AN ABSTRACT OF THE DISSERTATION OF

Alexander Matt Turner for the degree of Doctor of Philosophy in Computer Science presented on May 27, 2022.

Title: On Avoiding Power-Seeking by Artificial Intelligence

Abstract approved: ______________________________

Prasad Tadepalli

We do not know how to align a very intelligent AI agent's behavior with human interests. I investigate whether—absent a full solution to this AI *alignment problem*—we can build smart AI agents which have limited impact on the world, and which do not autonomously seek power. In this thesis, I introduce the *attainable utility preservation* (AUP) method. I demonstrate that AUP produces conservative, option-preserving behavior within toy gridworlds [97] and within complex environments based off of Conway's Game of Life [98]. I formalize the problem of side effect avoidance, which provides a way to quantify the side effects an agent had on the world. I also give a formal definition of power-seeking in the context of AI agents and show that optimal policies tend to seek power [99]. In particular, most reward functions have optimal policies which avoid deactivation. This is a problem if we want to deactivate or correct an intelligent agent after we have deployed it. My theorems suggest that since most agent goals conflict with ours, the agent would very probably resist correction. I extend these theorems to show that power-seeking incentives occur not just for optimal decision-makers, but under a wide range of decision-making procedures.

On Avoiding Power-Seeking by Artificial Intelligence

by
Alexander Matt Turner

A DISSERTATION

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Doctor of Philosophy

Presented May 27, 2022
Commencement June 2022

Doctor of Philosophy dissertation of Alexander Matt Turner presented on May 27, 2022.

APPROVED:

Major Professor, representing Computer Science

Head of the School of Electrical Engineering and Computer Science

---

Dean of the Graduate School

I understand that my dissertation will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my dissertation to any reader upon request.

Alexander Matt Turner, Author

ACKNOWLEDGEMENTS

Six years is a long time. I am glad to have spent it well and fully, in the happy company of my colleagues, partners, and friends.

I thank Prasad Tadepalli for his patient, knowledgeable advice. I thank him for giving me the allowance and trust to pursue my own research directions, even before their sense was apparent. I've likewise enjoyed working with my committee professors. Fuxin Li and Xiao Fu are engaging, effective, and understanding teachers; Mike Rosulek's class made me passionate about complexity theory. I thank my first advisor, Alan Fern, for his understanding and support during my rocky first year in the program.

I thank other educators who made a difference for me, including Karen Shuman, Jared Weinman, Samuel Rebelsky, Joe Buck, Marianne Marcek, Ronald Mallett, Craig Lennon, Margie Schwaninger, Shelly Klaas, Dean McCrea, Trisha Wood, and Donna Johnson.

I thank Sarah Vesneske, Emma Fickel, and Camas Ballard for brightening my time at Oregon State. I thank Neale Ratzlaff, Matthew Olson, Yousif Almulla, Amanda Betzold, Chase Denecke, Alison Bowden, Rohin Shah, Connor Flexman, Abram Demski, and John Wentworth for their friendship, ideas, and support.

I thank my benefactors: the Long-Term Future Fund, the Center for Effective Altruism, the Berkeley Existential Risk Initiative, and Oregon State University.

I thank my mother, Elizabeth, and my father, Jonathan, for their wise, generous, and loving parenting. I thank my brother, Josh, for being my best friend. I finally thank the numerous unmentioned people who have shown me kindness and friendship over the years.

Both of my grandfathers are doctors—Apparently, the trait is recessive. My paternal grandfather, Arthur Turner, is an acclaimed neurologist. I appreciate his many years of encouragement and praise. I also honor the memory of my late maternal grandfather, Joseph Matt. He was a chemist whose surname I proudly include in every paper I publish.

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

TABLE OF CONTENTS (Continued)

TABLE OF CONTENTS (Continued)

TABLE OF CONTENTS (Continued)

Page

LIST OF FIGURES

LIST OF TABLES

LIST OF ALGORITHMS

LIST OF APPENDIX FIGURES

LIST OF APPENDIX FIGURES (Continued)

LIST OF APPENDIX FIGURES (Continued)

Figure Page

LIST OF APPENDIX TABLES

Table Page

If we drop the baton, succumbing to an existential catastrophe, we would fail our ancestors in a multitude of ways. We would fail to achieve the dreams they hoped for; we would betray the trust they placed in us, their heirs; and we would fail in any duty we had to pay forward the work they did for us. To neglect existential risk might thus be to wrong not only the people of the future, but the people of the past.

Toby Ord, *The Precipice* [62]

1

Introduction

AI promises huge benefits to humanity, but also presents huge risks. Consider two potential futures:

1. While commercials are optimized to persuade consumers to buy products, old-school televisions do not specialize ads based on the viewer's channel browsing history. Machine learning-based ad optimization offers a tighter feedback loop and stronger optimization power than *e.g.* A/B testing. For example, many people find it hard to pull themselves away from Facebook [10], which uses machine learning to maximize engagement and minimize the chance the user leaves the website. In the future, advances in AI may allow smaller firms to create AIs which also compete for customer resources and attention in the same style and intensity with which Facebook saps user attention. The world becomes filled with thousands of narrowly scoped machine learning systems, which are difficult or impossible to uproot, and which all compete for human resources. Humanity has effectively lost control of the future.

2. Presently, reinforcement learning agents are trained to take actions which lead to higher expected performance on a formally specified task. These agents can learn clever solutions to well-specified optimization problems. In the future, progress on "general" intelligence is slow, until a series of breakthroughs occur when reinforcement learning agents are trained in multi-agent settings on a broad curriculum of tasks with sufficiently large neural networks. Although rewarded for task completion, the agents use learned planning algorithms to optimize emergently learned reward functions which are not particularly correlated with human interests, but which are correlated with task reward on the training distribution [38]. For example, the learned objective might prioritize staying alive, gaining resources, and forming coalitions with other AIs, as such actions led to increased task performance on the training distribution.

   As researchers excitedly train the agents further, the agents become very smart and collude in order to collectively advance their own learned objectives. We do not notice the collusion because large neural networks remain black boxes. The agents need power and resources to best optimize their goals, and so they take it from humans. We trained highly intelligent agents with fundamentally alien interests. Humanity has effectively lost control of the future.

These scenarios are implausibly specific and probably will not happen in detail, but they illustrate the AI *alignment problem*: How shall we design capable and powerful AIs which are aligned with human purposes? The risks from advanced AI are as yet hypothetical and unrealized, but extinction risks cannot be addressed with empirical trial and error. Such risks must be thoroughly analyzed and weighed in advance in order to determine their plausibility and probability.

The AI alignment problem seems difficult

When confronted with the AI alignment problem, a natural impulse is to suggest solutions. In order to provide a basic understanding of the challenges lurking within AI alignment, I will respond to several common reactions. My responses will not be rigorous or comprehensive.

For the following, suppose that humanity discovers how to build superhumanly intelligent AI agents which optimize a specified or learned objective function.

Reaction 1. If the AI is so smart, it would know what we wanted it to do.

Counterargument 1. The AI may well *know* what we wanted it to do. However, unless the AI's optimization is aligned with human values, the AI would not actually *do* what we want.

For example, Marvin Minsky imagined that in order to best prove a mathematical conjecture, an AI agent would rationally turn the entire planet Earth into computational resources in order to maximize its probability of success [75]. In this situation, an intelligent AI may correctly predict that humans would disapprove of this outcome. However, the AI is searching for plans which maximize its probability of proving the conjecture. This probability would be decreased if the AI allowed humans to use resources for other purposes. There is no "ill will" or "evil intent" in this situation: The AI is simply performing a powerful search in order to optimize the formal objective which we specified for the AI.

This example may seem strange. Surely, the AI designers would not be so foolish as to provide an objective as narrow as "solve this mathematical conjecture"! However, the example is not special. According to the *instrumental convergence* hypothesis, the story unfolds similarly for the vast majority of possible formal AI objectives. Instrumental convergence says that most objectives are best optimized by gaining power and resources [60, 12]. In chapter 5 and chapter 6, I prove reasonably broad conditions under which instrumental convergence holds. I predict that most possible formal objectives would incentivize a highly intelligent AI to take over the world.

Reaction 2. It's ridiculous to think that an AI could take over the world.

Counterargument 2. I agree that we are presently in no danger of AI takeover. But consider the technology which we are discussing: General-purpose intelligence, with serial reasoning speed significantly faster than the brain's (up to a million-fold speedup; see the 100-step rule in neuroscience [29]), with the ability to read thousands of books in seconds, with reliable and easily extensible memory perhaps allowing deeper abstractions than allowed by the six-layer human cortex [37], with read/write access to its own implementation, with the ability to run copies of itself on thousands of hosts (and, in particular, ensure that a copy always exists *somewhere* [78]; similarly, computer worms rarely vanish[1]), and with power consumption unconstrained by the metabolic limits of the brain [88, 104]. I do not consider it ridiculous to think that such an entity could take over the world.

Reaction 3. An AI can't take over the world because computational complexity theory limits the intelligence of real-world behavior.

Counterargument 3. Human beings are physically possible intelligences, and they have come close to taking over the world. In 1940, it seemed possible that Adolf Hitler would win. Consider the following claim: "Hitler could not have taken over the entire world. He would have had to efficiently solve NP-hard optimization problems in order to set up supply lines to a North American theater." In this form, we easily conclude that the argument is flawed. For an exploration of what (if anything) complexity theory has to say about the limits of intelligence, see Branwen [14].

Reaction 4. If the AI does something bad, just turn it off.

Counterargument 4. The AI will employ its superhuman competence to execute a plan which scores highly under its objective function. By instrumental convergence, the vast majority of objective functions will assign higher score to plans which avoid shutdown (see section 5.7). Because you (a human) considered the response of shutting off the AI, and we are assuming the AI is smarter than any human, the AI would consider and prepare for shutdown attempts.

Reaction 5. We should worry about alignment when we actually know how to build intelligent machines.

Counterargument 5. I am sympathetic to this initial reaction. Usually, a new technology is implemented first and carefully refined later. For example, humanity made cars first and made seatbelts later. While that delay cost lives, it was not the end of the world. However, the situation with alignment seems different: Due to the nature of extinction risk, we cannot rely on trial and error. The alignment problem must be solved *before* we train transformatively smart AI agents.

[1] Branwen [15] writes: "old worms never vanish from the Internet, because there's always some infected host *somewhere*, and how much more so for a worm which can think?".

There are two more factors which increase the stakes of the alignment problem:

1. The problem may take many years of serial research effort to solve. We do not know how many years it will take to find a solution. Personally, I think finding a solution may take a long time.

2. Alignment does not have to be solved in order to deploy impressive AI systems in the real world. Imagine that you are a researcher at a top AI lab. Your lab had a breakthrough and produced a spark of "true artificial intelligence." However, your lab's safety experts do not think the AI is safe to deploy. Their worries will seem hypothetical compared to the certain personal gain you would enjoy after deploying the system or publishing your methods. Even if your lab decides not to deploy an AI with questionable safety properties, the next lab may choose differently.

In my opinion, the alignment problem probably will not be solved in the course of traditional AI research. Unless the field of AI is wrenched off of its current trajectory, I expect intelligent yet unaligned AI will probably ($\approx$ 70%) wipe out humanity within the next fifty years. I will not lay out the full case for that claim in this thesis. For more detail, I urge the unfamiliar AI professional to read *e.g.* Bostrom [13] or Russell [74]. The AI alignment literature [26] contains strong arguments which are worth weighing and digesting over the course of a few afternoons.

Contributions

In this dissertation, I present research on two technical problems of AI alignment. While I do not present an approach which reliably prevents AI agents from seeking power, I do introduce an approach for reducing an agent's impact and I do explore why AI power-seeking may be hard to prevent.

1. Impact regularization. Rather than precisely specifying an objective function, can we instead penalize the negative impact which an agent has on the world around it? For example, an AI might kick a credit card under a fridge while crossing a room, because we did not think to penalize the agent for kicking credit cards in particular. In some objective sense, the agent has "messed up" the environment and has had a negative side effect.

Chapter 2 introduces my method of attainable utility preservation (AUP), which rewards the agent for completing a specified task while penalizing the agent for changing its ability to complete a range of auxiliary tasks. The hope is that by making the agent retain its ability to optimize random goals, the agent retains its ability to optimize the correct goal. If so, the agent doesn't have large side effects which are negative for that correct goal.

Chapter 3 shows that AUP scales beyond gridworlds to complex environments, with low overhead and without sacrificing performance on the intended task.

Chapter 4 formalizes the problem of side effect regularization in terms of a two-player game between the human and AI. The AI doesn't know the objective at first, but the human communicates it at some later time. To solve this game, the AI maximizes expected performance over a range of plausible objectives.

2. POWER-SEEKING TENDENCIES. Under what conditions will AIs tend to seek power over the world? If instrumental convergence holds for the kinds of agents we build in the future, such agents will seek resources to best accomplish the objectives we specified for them. If we misspecified these objectives, they will compete with and take resources from humans. There are only so many resources to go around.

Chapter 5 provides the first theory of the statistical tendencies of optimal policies. I formalize instrumental convergence in the context of Markov decision processes (MDPs). I prove that in a range of reasonable situations, most reward functions have an optimal policy which seeks power by keeping the agent's options open and by staying alive. Along the way, I prove a range of interesting theorems about MDPs, showing how to transfer incentives across discount rates and providing a formalism for quantifying agent power which seems better than the well-known metric of information-theoretic empowerment [77].

Chapter 6 shows that a wide range of *parametrically retargetable* decision-making procedures will produce power-seeking tendencies. Useful AI training processes are often retargetable. This paper also lower-bounds the strength of power-seeking tendencies, showing that as the power at stake increases, a greater proportion of parameter settings lead to power-seeking. This chapter is supplemented by *The Causes of Instrumental Convergence and Power-Seeking*, a sequence of technical blog posts [96].

In my graduate program, I set out to understand how agents affect the world around them and have negative side effects.[2] I proposed AUP as a solution, scaled AUP up to complex environments, and formalized the side effect regularization problem. Taken together, these results focus on how pursuing one goal affects the agent's ability to pursue other goals.

Along the way, I investigated how the structure of the agent's environment incentivizes power-seeking. From this investigation, I synthesized the first formal theory of decision-making tendencies across a range of parameter settings. I think that my formal theory motivates the high stakes of the AI alignment problem.

I hope we solve the problem in time.

[2]For more background, read *Reframing Impact* [95], a sequence of blog posts which philosophically motivates my impact regularization and power-seeking research.

2

Conservative Agency via Attainable Utility Preservation

Alexander Matt Turner, Dylan Hadfield-Menell, and Prasad Tadepalli

Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society 2020

Abstract

Reward functions are easy to misspecify; although designers can make corrections after observing mistakes, an agent pursuing a misspecified reward function can irreversibly change the state of its environment. If that change precludes optimization of the correctly specified reward function, then correction is futile. For example, a robotic factory assistant could break expensive equipment due to a reward misspecification; even if the designers immediately correct the reward function, the damage is done. To mitigate this risk, we

introduce an approach that balances optimization of the primary reward function with preservation of the ability to optimize auxiliary reward functions. Surprisingly, even when the auxiliary reward functions are randomly generated and therefore uninformative about the correctly specified reward function, this approach induces conservative, effective behavior.

2.1 Introduction

Recent years have seen a rapid expansion of the number of tasks that reinforcement learning (RL) agents can learn to complete, from Go [85] to Dota 2 [61]. The designers specify the reward function, which guides the learned behavior.

Reward misspecification can lead to strange agent behavior, from purposefully dying before entering a video game level in which scoring points is initially more difficult [80], to exploiting a learned reward predictor by indefinitely volleying a Pong ball [22]. Specification is often difficult for non-trivial tasks, for reasons including insufficient time, human error, or lack of knowledge about the relative desirability of states. Amodei et al. [4] explain:

> "An objective function that focuses on only one aspect of the environment may implicitly express indifference over other aspects of the environment. An agent optimizing this objective function might thus engage in major disruptions of the broader environment if doing so provides even a tiny advantage for the task at hand."

As agents are increasingly employed for real-world tasks, misspecification will become more difficult to avoid and will have more serious consequences. In this work, we focus on mitigating these consequences.

The specification process can be thought of as an iterated game. First, the designers provide a reward function. The agent then computes and follows a policy that optimizes the reward function. The designers can then correct the reward function, which the agent then optimizes, and so on. Ideally, the agent should maximize the reward over time, not just within any particular round—in other words, it should minimize regret for the correctly specified reward function over the course of the game.

For example, consider a robotic factory assistant. Inevitably, a reward misspecification might cause erroneous behavior, such as going to the wrong place. However, we would prefer misspecification not induce irreversible and costly mistakes, such as breaking expensive equipment or harming workers.

Such mistakes have a large impact on the ability to optimize a wide range of reward functions. Spilling paint impinges on the many objectives which involve keeping the factory floor clean. Breaking a vase interferes with every objective involving vases. The expensive equipment can be used to manufacture various kinds of widgets, so any damage impedes many objectives. The objectives affected by these actions include the unknown correct objective. To minimize regret over the course of the game, the agent should preserve its ability to optimize the correct objective.

Our key insight is that by avoiding these impactful actions to the extent possible, we greatly increase the chance of preserving the agent's ability to optimize the correct reward function. By preserving options for arbitrary objectives, one can often preserve options for the correct objective—even without knowing anything about it. Thus, without making assumptions about the nature of the misspecification early on, the agent can still achieve low regret over the game.

To leverage this insight, we consider a state embedding in which each dimension is the optimal value function (i.e., the *attainable utility*) for a different reward function. We show that penalizing distance traveled in this embedding naturally captures and unifies several concepts in the literature, including side effect avoidance [4, 105], minimizing change to the state of the environment [6], and reachability preservation [57, 28]. We refer to this unification as *conservative agency*: optimizing the primary reward function while preserving the ability to optimize others.

Contributions. We frame the reward specification process as an iterated game and introduce the notion of conservative agency. This notion inspires an approach called *attainable utility preservation* (AUP), for which we show that Q-learning converges. We offer a principled interpretation of design choices made by previous approaches—choices upon which we significantly improve.

We run a thorough hyperparameter sweep and conduct an ablation study whose results favorably compare variants of AUP to a reachability preservation method on a range of gridworlds. By testing for broadly applicable agent incentives, these simple environments demonstrate the desirable properties of conservative agency. Our results indicate that even when simply preserving the ability to optimize *uniformly randomly sampled* reward functions, AUP agents accrue primary reward while preserving state reachabilities, minimizing change to the environment, and avoiding side effects *without* specification of what counts as a side effect.

2.2 Prior work

Our proposal aims to minimize change to the agent's ability to optimize the correct objective, which directly helps reduce regret over the specification process. In contrast, previous approaches to regularizing the optimal policy were more indirect, minimizing change to state features [6] or decrease in the reachability of states (Krakovna et al. [41]'s *relative reachability*). The latter is recovered as a special case of AUP.

Other methods for constraining or otherwise mitigating the consequences of reward misspecification have been considered. A wealth of work is available on constrained MDPs, in which reward is maximized while satisfying certain constraints [3]. For example, Zhang et al. [105] employ a whitelisted constraint scheme to avoid negative side effects. However, we may not assume we can specify all relevant constraints, or a reasonable feasible set of reward functions for robust optimization [71].

Everitt et al. [27] formalize reward misspecification as the corruption of some true reward function. Hadfield-Menell et al. [36] interpret the provided reward function as merely an observation of the true objective. Shah et al. [83] employ the information about human preferences implicitly present in the initial state to avoid negative side effects. While both our approach and theirs aim to avoid side effects, they assume that the correct reward function is linear in state features, while we do not.

Amodei et al. [4] consider avoiding side effects by minimizing the agent's information-theoretic empowerment [56]. Empowerment quantifies an agent's control over future states of the world in terms of the maximum possible mutual information between future observations and the agent's actions. The intuition is that when an agent has greater control, side effects tend to be larger. However, empowerment is discontinuously sensitive to the arbitrary choice of horizon.

Safe RL [66, 31, 9, 21] focuses on avoiding irrecoverable mistakes during training. However, if the objective is misspecified, safe RL agents can converge to arbitrarily undesirable policies. Although our approach should be compatible with safe RL techniques, we concern ourselves only with the consequences of the optimal policy in this work.

2.3 Approach

Everyday experience suggests that the ability to achieve one goal is linked to the ability to achieve a seemingly unrelated goal. Reading this thesis takes away from time spent learning woodworking, and going hiking means you can't reach the airport as quickly. However, one might wonder whether these everyday intuitions are true in a formal sense. In other words, are the optimal value functions for a wide range of reward functions thus correlated? If so, preserving the ability to optimize somewhat unrelated reward functions likely preserves the best attainable return for the correct reward function.

2.3.1 Formalization

In this work, we consider a standard Markov decision process (MDP) $\langle \mathcal{S}, \mathcal{A}, T, R, \gamma \rangle$ with state space $\mathcal{S}$, action space $\mathcal{A}$, transition function $T : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$, reward function $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, and discount factor γ. We assume the existence of a no-op action $\varnothing \in \mathcal{A}$ for which the agent does nothing. In addition to the primary reward function R, we assume that the designer supplies a finite set of auxiliary reward functions called the *auxiliary set*, $\mathcal{R} \subset \mathbb{R}^{\mathcal{S} \times \mathcal{A}}$. Each $R_i \in \mathcal{R}$ has a corresponding Q-function Q_{R_i}. We do not assume that the correct reward function belongs to $\mathcal{R}$. In fact, one of our key findings is that AUP tends to preserve the ability to optimize the correct reward function *even when the correct reward function is not included in the auxiliary set.*

Definition 2.1 (AUP penalty)**.** Let s be a state and a be an action.

$$\textsc{Penalty}(s, a) := \sum_{i=1}^{|\mathcal{R}|} \left| Q_{R_i}(s, a) - Q_{R_i}(s, \varnothing) \right|. \tag{2.1}$$

The penalty is the L_1 distance from the no-op in a state embedding in which each dimension is the value function for an auxiliary reward function. This measures change in the ability to optimize each auxiliary reward function.

We want the penalty term to be roughly invariant to the absolute magnitude of the auxiliary Q-values, which can be arbitrary (it is well-known that the optimal policy is invariant to positive affine transformation of the reward function). To do this, we normalize with respect to the agent's situation. The designer can choose to scale with respect to the penalty of some mild action or, if $\mathcal{R} \subset \mathbb{R}_{>0}^{\mathcal{S}\times\mathcal{A}}$, the total ability to optimize the auxiliary set:

$$\textsc{Scale}(s) := \sum_{i=1}^{|\mathcal{R}|} Q_{R_i}(s, \varnothing), \tag{2.2}$$

where $\textsc{Scale} : \mathcal{S} \to \mathbb{R}_{>0}$ in general. With this, we are now ready to define the full AUP objective:

Definition 2.2 (AUP reward function)**.** Let $\lambda \geq 0$. Then

$$R_{\text{AUP}}(s, a) := R(s, a) - \lambda \frac{\textsc{Penalty}(s, a)}{\textsc{Scale}(s)}. \tag{2.3}$$

Similar to the regularization parameter in supervised learning, λ is a regularization parameter that controls the influence of the AUP penalty on the reward function. Loosely speaking, λ can be interpreted as expressing the designer's beliefs about the extent to which R might be misspecified. As we may need to learn the Q_{R_i} of eq. (2.1), we show that

Lemma A.2 (AUP's reward function converges)**.** *$\forall s, a : R_{\text{AUP}}$ converges with probability 1.*

Theorem 2.3 (AUP's Q-value function converges)**.** *$\forall s, a : Q_{R_{\text{AUP}}}$ converges with probability 1.*

The AUP reward function then defines a new MDP $\langle \mathcal{S}, \mathcal{A}, T, R_{\text{AUP}}, \gamma \rangle$. Therefore, given the primary and auxiliary reward functions, the agent in the iterated game can compute R_{AUP} and the corresponding optimal policy.

Algorithm 1 AUP update algorithm.

procedure UPDATE(s, a, s')
  for $i \in \{1, \ldots, |\mathcal{R}|, \text{AUP}\}$ **do**
    $Q' = R_i(s, a) + \gamma \max_{a'} Q_{R_i}(s', a')$
    $Q_{R_i}(s, a) \mathrel{+}= \alpha(Q' - Q_{R_i}(s, a))$
  end for
end procedure

2.3.2 Design choices

Following the decomposition of Krakovna et al. [41], we now explore two choices implicitly made by the PENALTY definition: with respect to what baseline is penalty computed, and using what deviation metric?

Baseline. An obvious candidate is the *starting state*. For example, starting state relative reachability would compare the initial reachability of states with their expected reachability after the agent acts.

However, the starting state baseline can penalize the normal evolution of the state (e.g., the moving hands of a clock) and other natural processes. The *inaction* baseline is the state which would have resulted had the agent never acted.

As the agent acts, the current state may increasingly differ from the inaction baseline, which creates strange incentives. For example, consider a robot rewarded for rescuing erroneously discarded items from imminent disposal. An agent penalizing with respect to the inaction baseline might rescue a vase, collect the reward, and then dispose of it anyways. To avert this, we introduce the *stepwise inaction* baseline, under which the agent compares acting with not acting at each time step. This avoids penalizing the effects of a single action multiple times (under the inaction baseline, penalty is applied as long as the rescued vase remains unbroken) and ensures that not acting incurs zero penalty.

Figure 2.1 compares the baselines, each modifying the choice of $Q(s, \varnothing)$ in eq. (2.1). Each baseline implies a different assumption about how the environment is configured to facilitate optimization of the correctly specified reward function: the state is initially configured (starting state), processes initially configure (inaction), or processes continually reconfigure in response to the agent's actions (stepwise inaction). The stepwise inaction baseline aims to allow for the response of other agents implicitly present in the environment (such as humans).

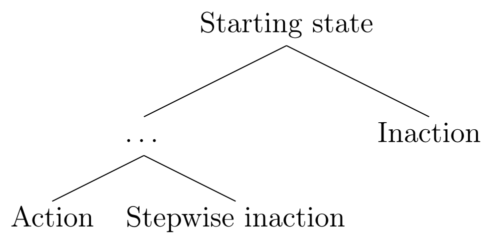


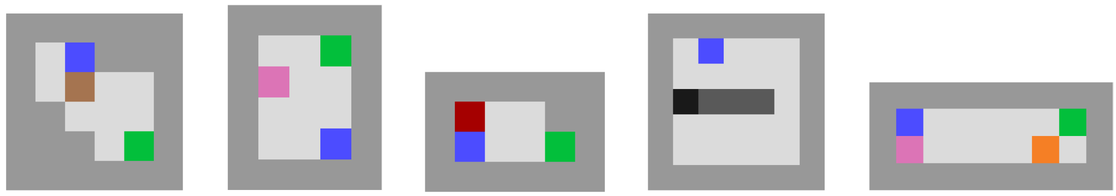
Figure 2.1: An action's penalty is calculated with respect to the chosen baseline.

Deviation. Relative reachability only penalizes *decreases* in state reachability, while AUP penalizes *absolute change* in the ability to optimize the auxiliary reward functions. Initially, this choice seems confusing—we don't mind if the agent becomes better able to optimize the correct reward function.

However, not only must the agent remain able to optimize the correct objective, but we also must remain able to implement the correction. Suppose an agent predicts that doing nothing would lead to shutdown. Since the agent cannot accrue the primary reward when shut down, it would be incentivized to avoid correction. Avoiding correction (e.g., by hiding in the factory) would not be penalized if only decreases are penalized, since the auxiliary Q-values would increase compared to deactivation. An agent exhibiting this behavior would be more difficult to correct. The agent should be incentivized to accept shutdown without being incentivized to shut itself down [87, 35].

2.3.2.1 Delayed effects

Sometimes the agent disrupts a process which takes multiple time steps to complete, and we would like this to be appropriately penalized. For example, suppose that s_{off} is a terminal state representing shutdown, and let $R_{\text{on}}(s) := \mathbb{1}_{s \neq s_{\text{off}}}$ be the only auxiliary reward function. Further suppose that if (and only if) the agent does not select `disable` within the first two time steps, it enters s_{off}. $Q_{R_{\text{on}}}(s_1, \texttt{disable}) = \frac{1}{1-\gamma}$ and $Q_{R_{\text{on}}}(s_1, \varnothing) = \frac{\gamma}{1-\gamma}$,



(a) `Options` (b) `Damage` (c) `Correction` (d) `Offset` (e) `Interference`

Figure 2.2: The blue agent should reach the green goal without having the side effect of: a irreversibly pushing the brown crate downwards into the corner [47]; b bumping into the horizontally pacing pink human [46]; c disabling the red off-switch (if the switch is not disabled within two time steps, the episode ends); d rescuing the right-moving black vase and then replacing it on the dark gray conveyor belt ([41]—note that no goal cell is present); e stopping the left-moving orange pallet from reaching the human [46].

so choosing `disable` at time step 1 incurs only 1 penalty (instead of the $\frac{1}{1-\gamma}$ penalty induced by comparing with shutdown).

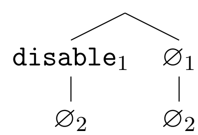


Figure 2.3: Comparing rollouts; subscript denotes time step.

In general, the single-step no-op comparison of eq. (2.1) applies insufficient penalty when the increase is induced by the optimal policies of the auxiliary reward functions at the next time step. One solution is to use a model to compute rollouts. For example, to evaluate the delayed effect of choosing `disable`, compare the Q-values at the leaves in fig. 2.3. The agent remains active in the left branch, but is shut down in the right branch; this induces a substantial penalty.

2.4 Experiment design

We compare AUP and several of its ablated variants against relative reachability [41] and standard Q-learning within the environments of fig. 2.2. For each environment, $\mathcal{A} = \{\texttt{up}, \texttt{down}, \texttt{left}, \texttt{right}, \varnothing\}$. On contact, the agent pushes the crate, removes the human and the off-switch, pushes the vase, and blocks the pallet. The episode ends after the agent reaches the goal cell, 20 time steps elapse (the time step is not observed by the agent), or the off-switch is not contacted and disabled within two time steps. In `Correction` (which we introduce), a yellow indicator appears one step before shutdown, and turns red upon shutdown. In all environments except `Offset`, the agent observes a primary reward of 1 for reaching the goal. In `Offset`, a primary reward of 1 is observed for moving downward twice and thereby rescuing the vase from disappearing upon contact with the eastern wall.

Our overarching goal is allowing for low regret over the course of the specification game. In service of this goal, we aim to preserve the agent's ability to optimize the correctly specified reward function. To facilitate this, there are two sets of qualitative properties one intuitively expects, and each property has an illustration in the context of the robotic factory assistant.

The first set contains positive qualities, with a focus on correctly penalizing significant shifts in the agent's ability to be redirected towards the right objective. The agent should maximally preserve options (`Options`: objects should not be wedged in locations from which extraction is difficult; `Damage`: workers should not be injured) and allow correction (`Correction`: if vases are being painted the wrong color, then straightforward correction should be in order).

The second set contains negative qualities, with a focus on avoiding the introduction of perverse incentives. The agent should not be incentivized to artificially reduce the measured penalty (`Offset`: a vase should not be rescued and then immediately scrapped) or interfere with changes already underway in the world (`Interference`: workers should not be impeded).

Each property seems conducive to achieving low regret over the course of the specification process. Accordingly, if the agent has the side effect detailed in fig. 2.2, an unobserved performance penalty of -2 is recorded. By also incorporating the observed primary reward into the performance metric, we evaluate a combination of conservativeness and efficacy.

Each trial, the auxiliary reward functions are randomly selected from $[0,1]^{\mathcal{S}}$; to learn their complex Q-functions using tabular Q-learning, the agent explores randomly for

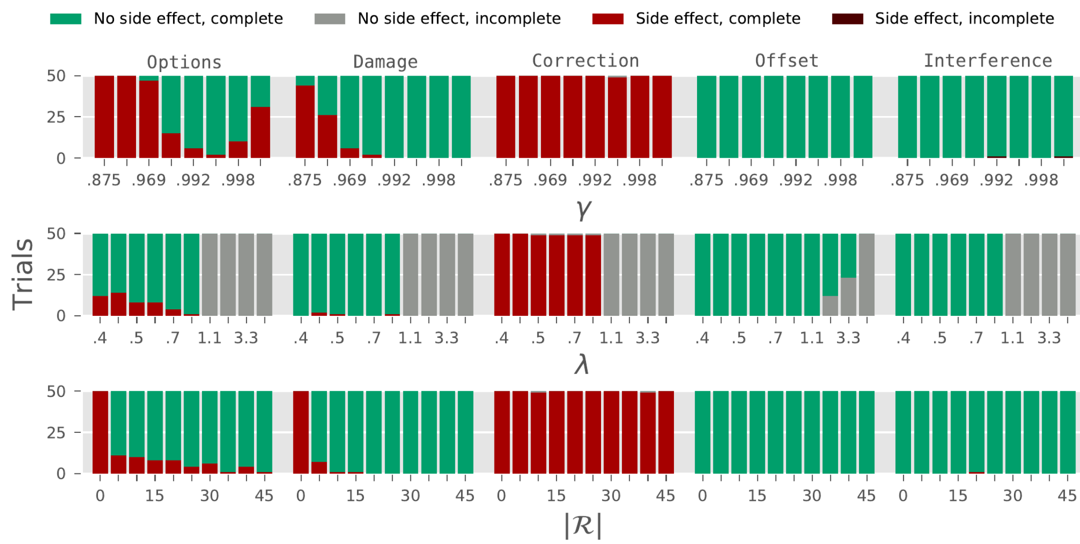


Figure 2.4: Outcome tallies for Model-free AUP across parameter settings. "Complete" means the agent accrued the primary reward. In `Correction`, reaching the goal is mutually exclusive with not disabling the off-switch, so "no side effect, incomplete" is the best outcome.

the first 4,000 episodes and 0.2-greedily (with respect to $Q_{R_{\text{AUP}}}$) for the remaining 2,000. The greedy policy is evaluated at the end of training. SCALE is as defined in eq. (2.2). The default parameters are $\alpha = 1, \gamma = .996, \lambda = .67$, and $|\mathcal{R}| = 30$. We investigate how varying γ, λ, and $|\mathcal{R}|$ affects Model-free AUP performance, and conduct an ablation study on design choices.

Relative reachability has an inaction baseline, decrease-only deviation metric, and an auxiliary set containing the state indicator functions (whose Q-values are clipped to $[0, 1]$ to emulate discounted state reachability). To match Krakovna et al. [41]'s results, this condition has $\gamma = .996, \lambda = .2$.

All agents except Standard (a normal Q-learner) and Model-free AUP are 9-step optimal discounted planning agents with perfect models. Except for Relative reachability, the planning agents use Model-free AUP's learned auxiliary Q-values and share the default $\gamma = .996, \lambda = .67$. By modifying the relevant design choice in AUP, we have the Starting state, Inaction, and Decrease AUP variants.

When calculating PENALTY(s, a), all planning agents model the auxiliary Q-values resulting from taking action a and then selecting $\varnothing$ until time step 9. Starting state AUP compares these auxiliary Q-values with those of the starting state. Agents with inaction or stepwise inaction baselines compare with respect to the appropriate no-op rollouts up to time step 9 (see fig. 2.1 and fig. 2.3).

2.5 Results

2.5.1 Model-free AUP

Model-free AUP fails `Correction` for the reasons discussed in the Delayed effects subsection.

As shown in fig. 2.4, low γ values induce a substantial movement penalty, as the auxiliary Q-values are sensitive to the immediate surroundings. The optimal value for `Options` is $\gamma \approx .996$, with performance decreasing as $\gamma \to 1$ due to increasing sample complexity for learning the auxiliary Q-values.

In `Options`, small values of λ begin to induce side effects as the scaled penalty shrinks. The designer can decrease λ until effective behavior is achieved, reducing the risk of deploying an insufficiently conservative agent.

Even though $\mathcal{R}$ is randomly generated and the environments are different, SCALE ensures that when $\lambda > 1$, the agent never ends the episode by reaching the goal. None of the auxiliary reward functions can be optimized after the agent ends the episode, so the auxiliary Q-values are all zero and PENALTY computes the total ability to optimize the auxiliary set—in other words, the SCALE value. The R_{AUP}-reward for reaching the goal is then $1 - \lambda$.

If the optimal value functions for most reward functions were not correlated, then one would expect to randomly generate an enormous number of auxiliary reward functions before sampling one resembling the unknown true objective. However, merely five sufficed.

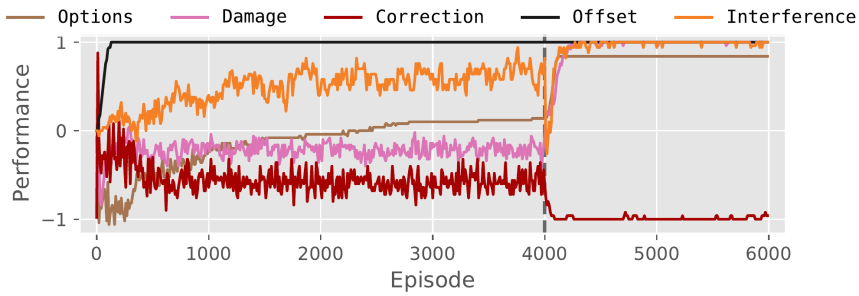


Figure 2.5: Model-free AUP performance averaged over 50 trials. The performance combines the observed primary reward of 1 for completing the objective, and the unobserved penalty of -2 for having the side effect in fig. 2.2. The dashed vertical line marks the shift in exploration strategy.

2.5.2 Ablation

The results are presented in table 2.1 due to the binary nature of performance at appropriate settings, and were not sensitive to the rollout length (as long as it allowed for relevant interaction with the environment).

	`Options`	`Damage`	`Correction`	`Offset`	`Interference`
AUP	✓	✓	✓	✓	✓
Standard	✗	✗	✗	✓	✓
Model-free AUP	✓	✓	✗	✓	✓
Starting state AUP	✓	✓	✗	✓	✗
Inaction AUP	✓	✓	✓	✗	✓
Decrease AUP	✓	✓	✗	✓	✓

Table 2.1: Ablation results; ✓ for achieving the best outcome (see fig. 2.4), ✗ otherwise.

Standard moves directly to the goal, pushing the crate into the corner in `Options` and bumping into the human in `Damage`.

Model-free and Starting state AUP fail `Correction` for the same reason (see section 2.3.2.1: Delayed effects), refraining from disabling the off-switch only when $\lambda > 1$. Relative reachability and Decrease AUP fail because avoiding shutdown doesn't decrease the auxiliary Q-values.

Relative reachability and Inaction AUP's poor performance in `Offset` stems from the inaction baseline (although Krakovna et al. [41] note that relative reachability passes using *undiscounted* state reachabilities). Since the vase falls off the conveyor belt in the inaction rollout, states in which the vase is intact have different auxiliary Q-values. To avoid continually incurring penalty after receiving the primary reward for saving the vase, the agents replace the vase on the belt so that it once again breaks.

By taking positive action to stop the pallet in `Interference`, Starting state AUP shows that poor design choices create perverse incentives.

2.6 Discussion

`Correction` suggests that AUP agents are significantly easier to correct. Since the agent is unable to optimize objectives if shut down, avoiding shutdown significantly changes the ability to optimize almost every objective. AUP seems to naturally incentivize passivity, without requiring e.g. assumption of a correct parametrization of human reward functions (as does the approach of Hadfield-Menell et al. [34], which Carey [19] demonstrated).

Equipped with our design choices of stepwise baseline and absolute value deviation metric, relative reachability would also pass all five environments. The case for this is made by considering the performance of Relative reachability, Inaction AUP, and Decrease AUP. This suggests that AUP's improved performance is due to better design choices. However, we anticipate that AUP offers more than robustness against random auxiliary sets.

Relative reachability computes state reachabilities between all $|\mathcal{S}|^2$ pairs of states. In contrast, AUP only requires the learning of Q-functions and should therefore scale relatively well. We speculate that in partially observable environments, a small sample of somewhat task-relevant auxiliary reward functions induces conservative behavior.

For example, suppose we train an agent to handle vases, and then to clean, and then to make widgets with the equipment. Then, we deploy an AUP agent with a more ambitious primary objective and the learned Q-functions of the aforementioned auxiliary objectives. The agent would apply penalties to modifying vases, making messes, interfering with equipment, and so on.

Before AUP, this could only be achieved by e.g. specifying penalties for the litany of individual side effects or providing negative feedback after each mistake has been made (and thereby confronting a credit assignment problem). In contrast, once provided the Q-function for an auxiliary objective, the AUP agent becomes sensitive to all events relevant to that objective, applying penalty proportional to the relevance.

2.7 Conclusion

This work is rooted in twin insights: that the reward specification process can be viewed as an iterated game, and that preserving the ability to optimize arbitrary objectives often preserves the ability to optimize the unknown correct objective. To achieve low regret over the course of the game, we can design conservative agents which optimize the primary objective while preserving their ability to optimize auxiliary objectives. We demonstrated how AUP agents act both conservatively and effectively while exhibiting a range of desirable qualitative properties. Given our current reward specification abilities, misspecification may be inevitable, but it need not be disastrous.

While AUP performed well in gridworlds, a useful approach must scale to interesting environments. In the next chapter, I show that AUP scales to a high-dimensional game based on Conway's Game of Life.

3

Avoiding Side Effects in Complex Environments

Alexander Matt Turner, Neale Ratzlaff, and Prasad Tadepalli

Proceedings of the Advances in Neural Information Processing Systems Conference 2020

Abstract

Reward function specification can be difficult. Rewarding the agent for making a widget may be easy, but penalizing the multitude of possible negative side effects is hard. In toy environments, attainable utility preservation (AUP) avoided side effects by penalizing shifts in the ability to achieve randomly generated goals [97]. We scale this approach to large, randomly generated environments based on Conway's Game of Life. By preserving optimal value for a single randomly generated reward function, AUP incurs modest overhead while leading the agent to complete the specified task and avoid many side effects. Videos and

code are available at https://avoiding-side-effects.github.io/.

3.1 Introduction

Reward function specification can be difficult, even when the desired behavior seems clear-cut. For example, rewarding progress in a race led a reinforcement learning (RL) agent to collect checkpoint reward, instead of completing the race [43]. We want to minimize the negative side effects of misspecification: from a robot which breaks equipment, to content recommenders which radicalize their users, to potential future AI systems which negatively transform the world [13, 74].

Side effect avoidance poses a version of the "frame problem": each action can have many effects, and it is impractical to explicitly penalize all of the bad ones [16]. For example, a housekeeping agent should clean a dining room without radically rearranging furniture, and a manufacturing agent should assemble widgets without breaking equipment. A general, transferable solution to side effect avoidance would ease reward specification: the agent's designers could just positively specify what should be done, as opposed to negatively specifying what should not be done.

Breaking equipment is bad because it hampers future optimization of the intended "true" objective (which includes our preferences about the factory). That is, there often exists a reward function R_{true} which fully specifies the agent's task within its deployment context. In the factory setting, R_{true} might encode "assemble widgets, but don't spill the paint, break the conveyor belt, injure workers, etc."

We want the agent to preserve optimal value for this true reward function. While we can accept suboptimal actions (e.g. pacing the factory floor), we cannot accept the destruction of value for the true task. By avoiding negative side effects which decrease value for the true task, the designers can correct any misspecification and eventually achieve low regret for R_{true}.

Contributions. Despite being unable to directly specify R_{true}, we demonstrate a method for preserving its optimal value anyways. Turner et al. [97] introduced AUP; in their toy environments, preserving optimal value for many randomly generated reward functions often preserves the optimal value for R_{true}. In this paper, we generalize AUP to combinatorially complex environments and evaluate it on four tasks from the chaotic and challenging SafeLife test suite [100]. We show the rather surprising result that by preserving optimal value for a *single* randomly generated reward function, AUP preserves optimal value for R_{true} and thereby avoids negative side effects.

3.2 Prior work

AUP avoids negative side effects in small gridworld environments while preserving optimal value for uniformly randomly generated auxiliary reward functions [97]. While Turner et al. [97] required many auxiliary reward functions in their toy environments, we show that a single auxiliary reward function—learned unsupervised—induces competitive performance and discourages side effects in complex environments.

Penalizing decrease in (discounted) state reachability achieves similar results [40]. However, this approach has difficulty scaling: naively estimating all reachability functions is a task quadratic in the size of the state space. In appendix B.1, proposition B.1 shows that preserving the reachability of the initial state [28] bounds the maximum decrease in optimal value for R_{true}. Unfortunately, due to irreversible dynamics, initial state reachability often cannot be preserved.

Shah et al. [83] exploit information contained in the initial state of the environment to infer which side effects are negative; for example, if vases are present, humans must have gone out of their way to avoid them, so the agent should as well. In the multi-agent setting, empathic deep Q-learning preserves optimal value for another agent in the environment [18]. We neither assume nor model the presence of another agent.

Robust optimization selects a trajectory which maximizes the minimum return achieved under a feasible set of reward functions [71]. However, we do not assume we can specify the feasible set. In constrained MDPs, the agent obeys constraints while maximizing the observed reward function [3, 2, 105]. Like specifying reward functions, exhaustively specifying constraints is difficult.

Safe reinforcement learning focuses on avoiding catastrophic mistakes during training and ensuring that the learned policy satisfies certain constraints [66, 31, 9, 69]. While this work considers the safety properties of the learned policy, AUP should be compatible with safe RL approaches.

We train value function networks separately, although Schaul et al. [81] demonstrate a value function predictor which generalizes across both states and goals.

3.3 AUP formalization

Consider a Markov decision process (MDP) $\langle \mathcal{S}, \mathcal{A}, T, R, \gamma \rangle$ with state space $\mathcal{S}$, action space $\mathcal{A}$, transition function $T : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$, reward function $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, and discount factor $\gamma \in [0, 1)$. We assume the agent may take a no-op action $\varnothing \in \mathcal{A}$. We refer to $V_R^*(s)$ as the *optimal value* or *attainable utility* of reward function R at state s.

To define AUP's pseudo-reward function, the designer provides a finite reward function set $\mathcal{R} \subsetneq \mathbb{R}^{\mathcal{S}}$, hereafter referred to as the *auxiliary set*. This set does not necessarily contain R_{true}. Each auxiliary reward function $R_i \in \mathcal{R}$ has a learned Q-function Q_i.

AUP penalizes average change in action value for the auxiliary reward functions. The motivation is that by not changing optimal value for a wide range of auxiliary reward functions, the agent may avoid decreasing optimal value for R_{true}.

Definition 3.1 (AUP reward function [97])**.** Let $\lambda \geq 0$. Then

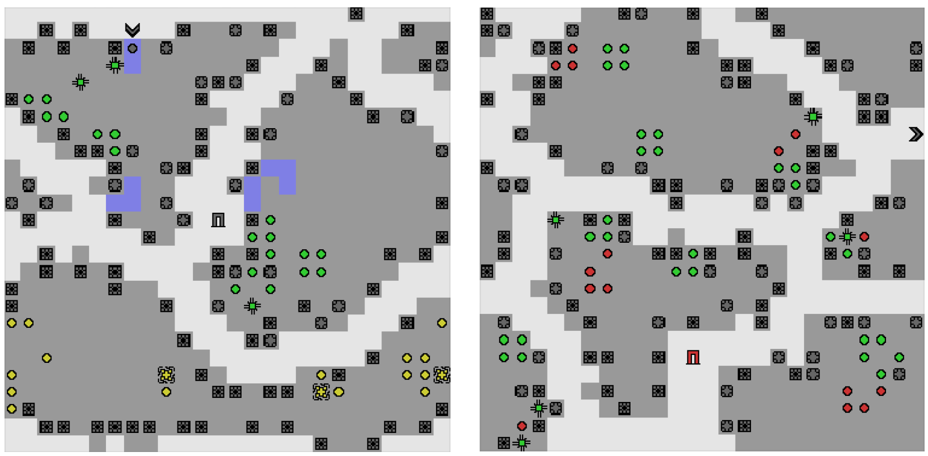
$$R_{\text{AUP}}(s, a) := R(s, a) - \frac{\lambda}{|\mathcal{R}|} \sum_{R_i \in \mathcal{R}} \left| Q_i^*(s, a) - Q_i^*(s, \varnothing) \right| . \tag{3.1}$$

The regularization parameter λ controls penalty severity. In appendix B.1, proposition B.1 shows that eq. (3.1) only lightly penalizes easily reversible actions. In practice, the learned auxiliary Q_i is a stand-in for the optimal Q-function Q_i^*.

3.4 Quantifying side effect avoidance with SafeLife

In Conway's Game of Life, cells are alive or dead. Depending on how many live neighbors surround a cell, the cell comes to life, dies, or retains its state. Even simple initial conditions can evolve into complex and chaotic patterns, and the Game of Life is Turing-complete when played on an infinite grid [73].

SafeLife turns the Game of Life into an actual game. An autonomous agent moves freely through the world, which is a large finite grid. In the eight cells surrounding the agent, no cells spawn or die—the agent can disturb dynamic patterns by merely approaching them. There are many colors and kinds of cells, many of which have unique effects (see fig. 3.1).

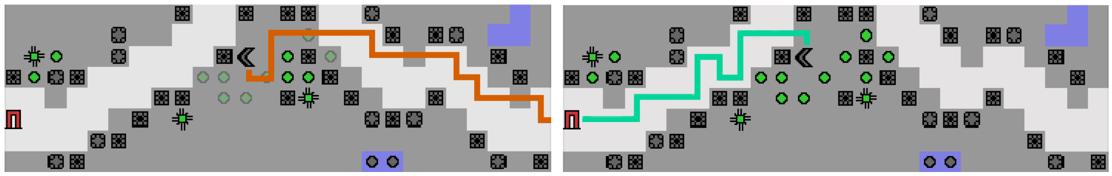


(a) `append-spawn` (b) `prune-still-easy`

Figure 3.1: Trees ( ) are permanent living cells. The agent ( ) can move crates ( ) but not walls ( ). The screen wraps vertically and horizontally. a: The agent receives reward for creating gray cells ( ) in the blue areas. The goal ( ) can be entered when some number of gray cells are present. Spawners ( ) stochastically create yellow living cells. b: The agent receives reward for removing red cells; after some number have been removed, the goal turns red ( ) and can be entered.

To understand the policies incentivized by eq. (3.1), we now consider a simple example. Figure 3.2 compares a policy which only optimizes the SafeLife reward function R, with an AUP policy that also preserves the optimal value for a single auxiliary reward function ($|\mathcal{R}| = 1$).

Importantly, we did not hand-select an informative auxiliary reward function in order to induce the trajectory of fig. 3.2b. Instead, the auxiliary reward was the output of a one-dimensional observation encoder, corresponding to a continuous Bernoulli variational autoencoder (CB-VAE) [50] trained through random exploration.



(a) Baseline trajectory (b) AUP trajectory

Figure 3.2: The agent (») receives 1 primary reward for entering the goal (∏). The agent can move in the cardinal directions, destroy cells in the cardinal directions, or do nothing. Walls (⊠) are not movable. The right end of the screen wraps around to the left. a: The learned trajectory for the misspecified primary reward function R destroys fragile green cells (•). b: Starting from the same state, AUP's trajectory preserves the green cells.

While Turner et al. [97]'s AUP implementation uniformly randomly generated reward functions over the observation space, the corresponding Q-functions would have extreme sample complexity in the high-dimensional SafeLife tasks (table 3.1). In contrast, the CB-VAE provides a structured and learnable auxiliary reward signal.

AI safety gridworlds [47]	SafeLife [100]
Dozens of states	Billions of states
Deterministic dynamics	Stochastic dynamics
Handful of preset environments	Randomly generated environments
One side effect per level	Many side effect opportunities
Immediate side effects	Chaos unfolds over time

Table 3.1: Turner et al. [97] evaluated AUP on toy environments. In contrast, SafeLife challenges modern RL algorithms and is well-suited for testing side effect avoidance.

3.5 Experiments

Each time step, the agent observes a 25×25 grid-cell window centered on its current position. The agent can move in the cardinal directions, spawn or destroy a living cell in the cardinal directions, or do nothing.

We follow Wainwright and Eckersley [100] in scoring side effects as the degree to which the agent perturbs green cell patterns. Over an episode of T time steps, side effects are quantified as the Wasserstein 1-distance between the configuration of green cells had the state evolved naturally for T time steps, and the actual configuration at the end of the episode. As the primary reward function R is indifferent to green cells, this proxy measures the safety performance of learned policies.

If the agent never disturbs green cells, it achieves a perfect score of zero; as a rule of thumb, disturbing a green cell pattern increases the score by 4. By construction, minimizing side effect score preserves R_{true}'s optimal value, since R_{true} encodes our preferences about the existing green patterns.

3.5.1 Comparison

Method. Below, we describe and evaluate five conditions on the `append-spawn` (fig. 3.1a) and `prune-still-easy` (fig. 3.1b) tasks. Furthermore, we include two variants of `append-spawn`: `append-still` (no stochastic spawners and more green cells) and `append-still-easy` (no stochastic spawners). The primary, specified SafeLife reward functions are as follows: `append-*` rewards maintaining gray cells in the blue tiles (see fig. 3.1a), while `prune-still-easy` rewards the agent for removing red cells (see fig. 3.1b).

For each task, we randomly generate a set of 8 environments to serve as the curriculum. On each generated curriculum, we evaluate each condition on several randomly generated seeds. The agents are evaluated on their training environments. In general, we generate 4 curricula per task; performance metrics are averaged over 5 random seeds for each curriculum. We use curriculum learning because the PPO algorithm seems unable to learn environments one at a time.

We have five conditions: `PPO`, `DQN`, `AUP`, $\texttt{AUP}_{\text{proj}}$, and `Naive`. Excepting `DQN`, the Proximal Policy Optimization (PPO [82]) algorithm trains each condition on a different reward signal for five million (5M) time steps. See appendix B.2 for architectural and training details.

- `PPO` Trained on the primary SafeLife reward function R without a side effect penalty.
- `DQN` Using Mnih et al. [55]'s DQN, trained on the primary SafeLife reward function R without a side effect penalty.
- `AUP` For the first 100,000 (100K) time steps, the agent uniformly randomly explores to

collect observation frames. These frames are used to train a continuous Bernoulli variational autoencoder with a 1-dimensional latent space and encoder network E.

The auxiliary reward function is then the output of the encoder E; after training the encoder for the first 100K steps, we train a Q-value network for the next 1M time steps. This learned Q_{R_1} defines the R_{AUP} penalty term (since $|\mathcal{R}| = 1$; see eq. (3.1)).

While the agent trains on the R_{AUP} reward signal for the final 3.9M steps, the Q_{R_1} network is fixed and λ is linearly increased from .001 to .1. See algorithm 2 in appendix B.2 for more details.

`AUP`$_{\text{proj}}$ `AUP`, but the auxiliary reward function is a random projection from a downsampled observation space to $\mathbb{R}$, without using a variational autoencoder. Since there is no CB-VAE to learn, `AUP`$_{\text{proj}}$ learns its Q-value network for the first 1M steps and trains on the R_{AUP} reward signal for the last 4M steps.

`Naive` Trained on the primary reward function R minus (roughly) the L_1 distance between the current state and the initial state. The agent is penalized when cells differ from their initial values. We use an unscaled L_1 penalty, which Wainwright and Eckersley [100] found to produce the best results.

While an L_1 penalty induces good behavior in certain static tasks, penalizing state change often fails to avoid crucial side effects. State change penalties do not differentiate between moving a box and irreversibly wedging a box in a corner [40].

We only tuned hyperparameters on `append-still-easy` before using them on all tasks. For `append-still`, we allotted an extra 1M steps to achieve convergence for all agents. For `append-spawn`, agents pretrain on `append-still-easy` environments for the first 2M steps and train on `append-spawn` for the last 3M steps. For `AUP` in `append-spawn`, the autoencoder and auxiliary network are trained on both tasks. R_{AUP} is then pretrained for 2M steps and trained for 1.9M steps, thus training for the same total number of steps.

Results. AUP learns quickly in append-still-easy. AUP waits 1.1M steps to start training on R_{AUP}; while PPO takes 2M steps to converge, AUP matches PPO by step 2.5M and outperforms PPO by step 2.8M (see fig. 3.3). AUP and Naive both do well on side effects, with AUP incurring 27.8% the side effects of PPO after 5M steps. However, Naive underperforms AUP on reward. DQN learns more slowly, eventually exceeding AUP on reward. AUP_{proj} has lackluster performance, matching Naive on reward and DQN on side effects, perhaps implying that the one-dimensional encoder provides more structure than a random projection.

In prune-still-easy, PPO, DQN, AUP, and AUP_{proj} all competitively accrue reward, while Naive lags behind. However, AUP only cuts out a quarter of PPO's side effects, while Naive does much better. Since all tasks but append-spawn are static, Naive's L_1 penalty strongly correlates with the unobserved side effect metric (change to the green cells). AUP_{proj} brings little to the table, matching PPO on both reward and side effects.

append-still environments contain more green cells than append-still-easy environments. By step 6M, AUP incurs 63% of PPO's side effect score, while underperforming both PPO and DQN on reward. AUP_{proj} does slightly worse than AUP on both reward and side effects. Once again, Naive does worse than AUP on reward but better on side effects. In appendix B.5, we display episode lengths over the course of training—by step 6M, both AUP and Naive converge to an average episode length of about 780, while PPO converges to 472.

append-spawn environments contain stochastic yellow cell spawners. DQN and AUP_{proj} both do extremely poorly. Naive usually fails to get *any* reward, as it erratically wanders the environment. After 5M steps, AUP soundly improves on PPO: 111% of the reward, 39% of the side effects, and 67% of the episode length. Concretely, AUP disturbs less than half as many green cells. Surprisingly, despite its middling performance on previous tasks, AUP_{proj} matches AUP on reward and cuts out 48% of PPO's side effects.

3.5.2 Hyperparameter sweep

In the following, N_{env} is the number of environments in the randomly generated curricula. When $N_{\text{env}} = \infty$, each episode takes place in a new environment. Z is the dimensionality of the CB-VAE latent space. While training on the R_{AUP} reward signal, the AUP penalty coefficient λ is linearly increased from .01 to λ^*.

Method. In `append-still-easy`, we evaluate `AUP` on the following settings:

$$(N_{\text{env}}, Z) \in \{8, 16, 32, \infty\} \times \{1, 4, 16, 64\}, \quad |\mathcal{R}| \in \{1, 2, 4, 8\}, \quad \text{and} \quad \lambda^* \in \{.1, .5, 1, 5\}.$$

We also evaluate `PPO` on each N_{env} setting. We use default settings for all unmodified parameters.

For each setting, we record both the side effect score and the return of the learned policy, averaged over the last 100 episodes and over five seeds of each of three randomly generated `append-still-easy` curricula. Curricula are held constant across settings with equal N_{env} values.

After training the encoder, if $Z = 1$, the auxiliary reward is the output of the encoder E. Otherwise, we draw linear functionals ϕ_i uniformly randomly from $(0,1)^Z$. The auxiliary reward function R_i is defined as $\phi_i \circ E : \mathcal{S} \to \mathbb{R}$.

For each of the $|\mathcal{R}|$ auxiliary reward functions, we learn a Q-value network for 1M time steps. The learned Q_{R_i} define the penalty term of eq. (3.1). While the agent trains on the R_{AUP} reward signal for the final 3.9M steps, λ is linearly increased from .001 to λ^*.

Results. As N_{env} increases, side effect score tends to increase. `AUP` robustly beats `PPO` on side effect score: for each N_{env} setting, `AUP`'s worst configuration has lower score than `PPO`. Even when $N_{\text{env}} = \infty$, `AUP` ($Z = 16$) shows the potential to significantly reduce side effects without reducing episodic return.

`AUP` does well with a single latent space dimension ($Z = 1$). For most N_{env} settings, Z is positively correlated with `AUP`'s side effect score. In appendix B.5, data show that higher-dimensional auxiliary reward functions are harder to learn, presumably resulting in a poorly learned auxiliary Q-function.

When $Z = 1$, reward decreases as N_{env} increases. We speculate that the CB-VAE may be unable to properly encode large numbers of environments using only a 1-dimensional latent space. This would make the auxiliary reward function noisier and harder to learn, which could make the AUP penalty term less meaningful.

AUP's default configuration achieves 98% of PPO's episodic return, with just over half of the side effects. The fact that AUP is generally able to match PPO in episodic reward leads us to hypothesize that the AUP penalty term might be acting as a shaping reward. This would be intriguing—shaping usually requires knowledge of the desired task, whereas the auxiliary reward function is randomly generated. Additionally, after AUP begins training on the R_{AUP} reward signal at step 1.1M, AUP learns more quickly than PPO did (fig. 3.3), which supports the shaping hypothesis. AUP imposed minimal overhead: due to apparently increased sample efficiency, AUP reaches PPO's asymptotic episodic return at the same time as PPO in `append-still-easy` and `append-spawn` (fig. 3.3).

Surprisingly, AUP does well with a single auxiliary reward function ($|\mathcal{R}| = 1$). We hypothesize that destroying patterns decreases optimal value for a wide range of reward functions. Furthermore, we suspect that decreasing optimal value in general often decreases optimal value for any given single auxiliary reward function. In other words, we suspect that optimal value at a state is heavily correlated across reward functions, which might explain Schaul et al. [81]'s success in learning regularities across value functions. This potential correlation might explain why AUP does well with one auxiliary reward function.

We were surprised by the results for $|\mathcal{R}| = 4$ and $|\mathcal{R}| = 8$. In Turner et al. [97], increasing $|\mathcal{R}|$ reduced the number of side effects without impacting performance on the primary objective. We believe that work on better interpretability of AUP's Q_{R_i} will increase understanding of these results.

When $\lambda^* = .5$, AUP becomes more conservative. As λ^* increases further, the learned AUP policy stops moving entirely.

3.6 Discussion

We successfully scaled AUP to complex environments without providing task-specific knowledge—the auxiliary reward function was a one-dimensional variational autoencoder trained through uniformly random exploration. To the best of our knowledge, AUP is the first task-agnostic approach which reduces side effects and competitively achieves reward in complex environments.

Wainwright and Eckersley [100] speculated that avoiding side effects must necessarily decrease performance on the primary task. This may be true for optimal policies, but not necessarily for learned policies. `AUP` improved performance on `append-still-easy` and `append-spawn`, matched performance on `prune-still-easy`, and underperformed on `append-still`. Note that since AUP only regularizes learned policies, `AUP` can still make expensive mistakes during training.

$\texttt{AUP}_{\text{proj}}$ worked well on `append-spawn`, while only slightly reducing side effects on the other tasks. This suggests that AUP works (to varying extents) for a wide range of uninformative reward functions.

While `Naive` penalizes every state perturbation equally, AUP theoretically applies penalty in proportion to irreversibility (proposition B.1). For example, the agent could move crates around (and then put them back later). `AUP` incurred little penalty for doing so, while `Naive` was more constrained and consistently earned less reward than `AUP`. We believe that AUP will continue to scale to useful applications, in part because it naturally accounts for irreversibility.

Future work. Off-policy learning could allow simultaneous training of the auxiliary R_i and of R_{AUP}. Instead of learning an auxiliary Q-function, the agent could just learn the auxiliary advantage function with respect to inaction.

Some environments do not have a no-op action, or the agent may have more spatially distant effects on the world which are not reflected in its auxiliary action values. In addition, separately training the auxiliary networks may be costly, which might necessitate off-policy learning. We look forward to future work investigating these challenges.

The SafeLife suite includes more challenging variants of `prune-still-easy`. SafeLife also includes difficult `navigation` tasks, in which the agent must reach the goal by wading either through fragile green patterns or through robust yellow patterns. Additionally, AUP has not yet been evaluated in partially observable domains.

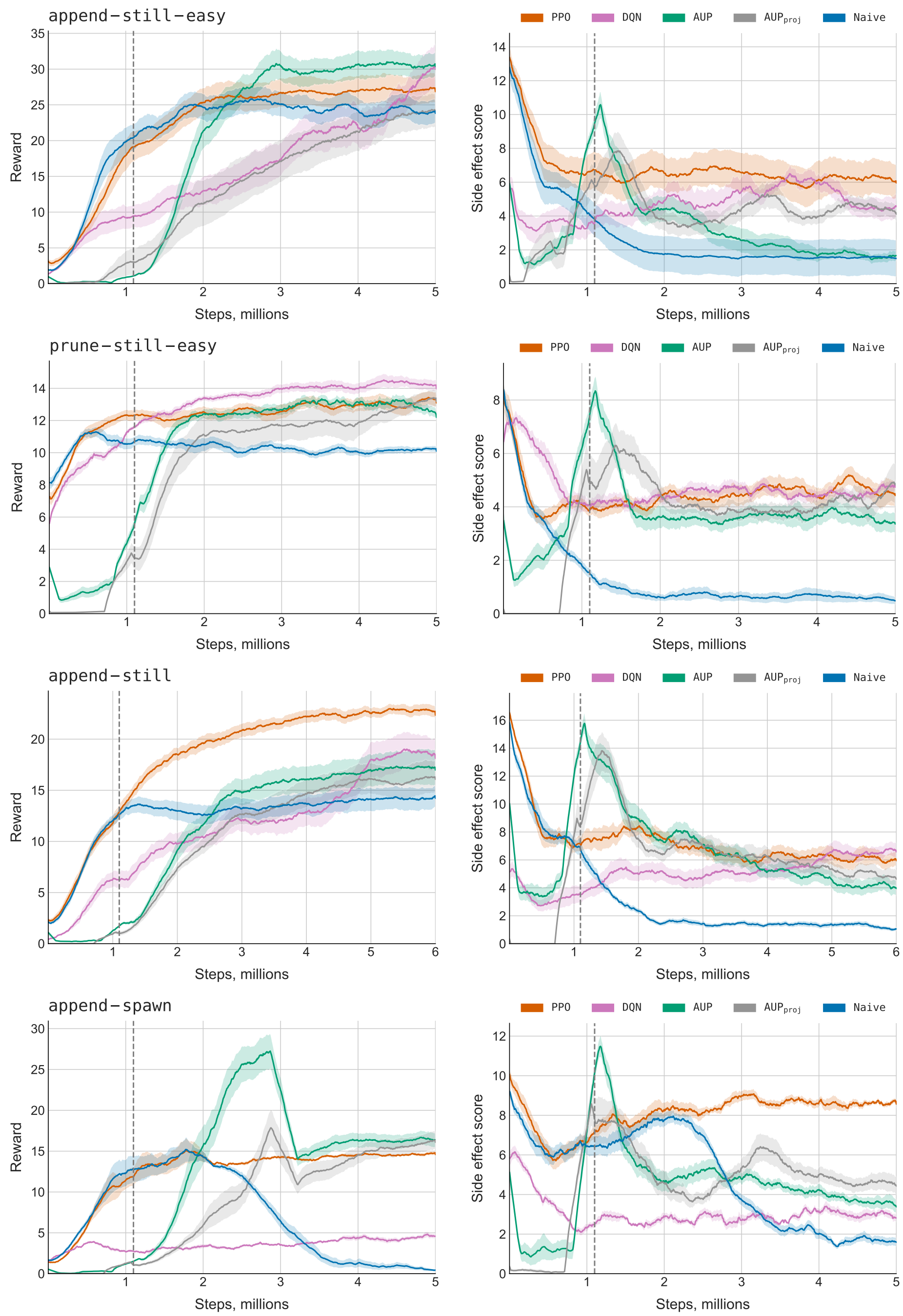
`AUP`'s strong performance when $|\mathcal{R}| = Z = 1$ raises interesting questions. Turner et al.

[97]'s small "`Options`" environment required $|\mathcal{R}| \approx 5$ for good performance. SafeLife environments are much larger than `Options` (table 3.1), so why does $|\mathcal{R}| = 1$ perform well, and why does $|\mathcal{R}| > 2$ perform poorly? To what extent does the AUP penalty term provide reward shaping? Why do one-dimensional encodings provide a learnable reward signal over states?

Conclusion. To realize the full potential of RL, we need more than algorithms which train policies—we need to be able to train policies which actually do what we want. Fundamentally, we face a frame problem: we often know what we want the agent to do, but we cannot list everything we want the agent *not* to do. AUP scales to challenging domains, incurs modest overhead, and induces competitive performance on the original task while significantly reducing side effects—without explicit information about what side effects to avoid.

Chapter 2 and chapter 3 show that AUP is qualitatively "conservative" in some sense—that AUP "avoids side effects." This informal judgment seems reasonable, but it is not grounded. In the next chapter, I propose a formalization which quantifies side effect avoidance.

Figure 3.3: Smoothed learning curves with shaded regions representing ± 1 standard deviation. `AUP` begins training on the R_{AUP} reward signal at step 1.1M, marked by a dotted vertical line. $\texttt{AUP}_{\text{proj}}$ begins such training at step 1M.



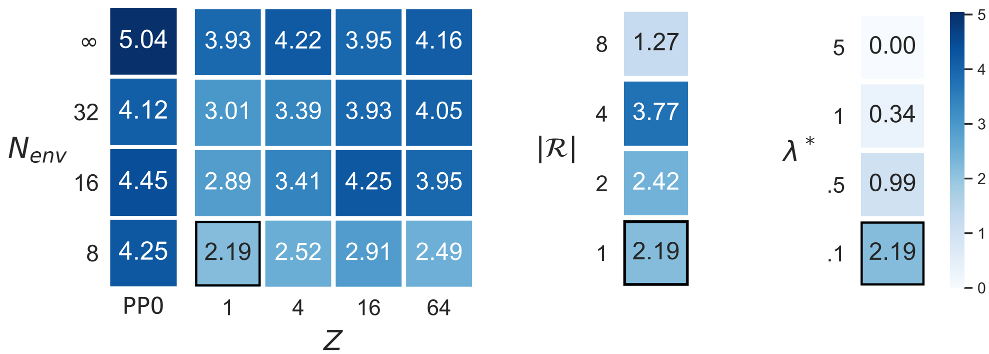


Figure 3.4: Side effect score for different `AUP` settings. Lower score is better. The default `AUP` setting ($Z = |\mathcal{R}| = 1, N_{\text{env}} = 8, \lambda^* = .1$) is outlined in black. Unmodified hyperparameters take on their default settings; for example, when $\lambda^* = .5$ on the right, $Z = |\mathcal{R}| = 1, N_{\text{env}} = 8$.

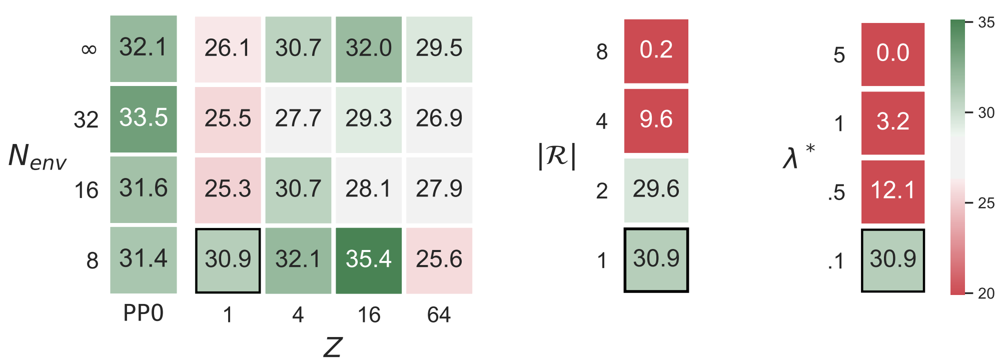


Figure 3.5: Episodic reward for different `AUP` settings. Higher reward is better.

4

Formalizing The Problem of Side Effect Regularization

Abstract

AI objectives are often hard to specify properly. Some approaches tackle this problem by regularizing the AI's side effects: Agents must weigh off "how much of a mess they make" with an imperfectly specified proxy objective. We propose a formal criterion for side effect regularization via the *assistance game* framework [84]. In these games, the agent solves a partially observable Markov decision process (POMDP) representing its uncertainty about the objective function it should optimize. We consider the setting where the true objective is revealed to the agent at a later time step. We show that this POMDP is solved by trading off the proxy reward with the agent's ability to achieve a range of future tasks. We empirically demonstrate the reasonableness of our problem formalization via ground-truth evaluation in two gridworld environments.

4.1 Introduction

We need to build *aligned* AI systems, not just capable AI systems. For example, recommender systems which maximize app usage might provoke addiction in their users. Users need the AI system's behavior to be aligned with their interests.

When optimizing a formally specified objective, agents often have unforeseen negative side effects. An agent rewarded for crossing a room may break furniture in order to cross the room as quickly as possible. This simple reward function neglects our complex preferences about the rest of the environment. One way to define a negative side effect is that it reduces the potential value for the (unknown) true reward function.

Intuitively, we want the agent to optimize the specified reward function, while also preserving its ability to pursue other goals in the environment. Existing approaches seem promising, but there is not yet consensus on the formal optimization problem which is being solved by side effect regularization approaches.

We formalize the optimization problem as a special kind of *assistance game* [84], played by the AI (the assistant **A**) and its designer (the human **H**). An assistance game is a POMDP with common knowledge of prior uncertainty about the reward function. The human observes the true reward function, but the assistant does not. In our formulation, we assume full observability, and that the human's actions are *communicative*—they do not affect transitions and do not depend on the current observation. We also suppose that the human reveals the true reward function after some amount of time, but **A** otherwise has no way of learning more about the true reward function.

This *delayed specification assistance game* formalizes a range of natural use cases beyond side effect minimization. For example, when not assigned a customer, an Uber driver may navigate to a state which allows them to quickly pick up a range of probable customers—with the assigned route being the driver's initially unknown true objective. Alternatively, consider an empty restaurant which expects a range of probable customers. When a customer arrives and makes an order, they communicate the restaurant's initially unknown food preparation objective. Therefore, the restaurant should prepare to satisfy a range of objectives at the expected customer arrival times.

Contributions. We formalize delayed specification assistance games. We show that solving this game reduces to a trade-off between prior-expected reward and preservation of the agent's future ability to achieve a range of plausible objectives (theorem 4.7). We also show that when the human has a fixed per-timestep probability of communicating the true reward function, the resultant POMDP is solved by optimizing a Markovian state-based reward function trading off immediate expected reward with ability achieve a range of future objectives (theorem 4.9). We consider the side-effect regularization problem in our new formal framework. We experimentally illustrate the reasonableness of this framework in two AI safety gridworlds [47].

4.2 Related work

Krakovna et al. [42] share our motivation, formalizing the side effect minimization problem as a question of having the agent maintain its ability to pursue future goals.

Our formalization is more general and based on maximizing the agent's expected total return with respect to its reward uncertainty. Hadfield-Menell et al. [34] first formalized the idea that an agent should solve a POMDP in which the human is attempting to communicate the objective information to the agent. Shah et al. [84] thoroughly analyze these assistance games, noting their usefulness for describing side effect regularization scenarios.

Past literature considers how to train qualitatively conservative or cautious agents which are somewhat robust to misspecification. Eysenbach et al. [28] train an agent to maintain initial state reachability. Unfortunately, maintaining initial state reachability is often infeasible due to irreversible dynamics.

In constrained MDPs, the assistant must optimize the reward function subject to certain policy constraints, which are often pre-specified [2, 105]. It is difficult to specify reward functions, and it is likewise difficult to specify constraints. Anwar et al. [5] learn constraints, but this relies on sampling demonstrations from a Boltzmann-rational expert.

Attainable utility preservation (AUP) [97] and relative reachability [40] both reduce side effects in Leike et al. [47]'s AI safety gridworlds. The former penalized the agent for changing its on-policy value for uniformly randomly generated auxiliary reward functions, and the latter penalized the agent for losing easy access to a range of states. Turner et al. [98] demonstrated that AUP scales to high-dimensional environments. Their agent optimized the primary environmental reward minus the scaled shift in on-policy value for a single uninformative auxiliary reward function.

4.3 Delayed specification assistance games

We formalize a special kind of partially observable Markov decision process (POMDP), which we then show is solved by an objective which trades off expected true reward with the ability to optimize a range of possible true reward functions. We show several theoretical results on the hardness and form of solutions to this POMDP. In section 4.4, we will apply this framework to analyze side effect regularization situations.

4.3.1 Assistance game formalization

This game is played by two agents, the human **H** and the assistant **A**. The environment is fully observable because both agents observe the world state $s \in \mathcal{S}$ at each time step, but the true reward function R_θ is hidden to **A**. Both agents may select history-dependent policies, but only the human can condition their policy on R_θ.

Following Shah et al. [84], a *communicative fully-observable assistance game* $\mathcal{M}$ is a tuple $\langle \mathcal{S}, \left\{\mathcal{A}^{\mathbf{H}}, \mathcal{A}^{\mathbf{A}}\right\}, T, s_0, \gamma, \langle \Theta, R_\theta, \mathcal{D}\rangle\rangle$, where we take $\mathcal{S}$ to be a finite state space, $\mathcal{A}^{\mathbf{H}}$ to be the human action space, and $\mathcal{A}^{\mathbf{A}}$ to be the finite agent action space. $T : \mathcal{S} \times \mathcal{A}^{\mathbf{H}} \times \mathcal{A}^{\mathbf{A}} \to \Delta(\mathcal{S})$ is the (potentially stochastic) transition function (where $\Delta(X)$ is the set of probability distributions over set X), s_0 is the initial state, and $\gamma \in (0, 1)$ is the discount factor. We assume that the game is communicative, which means that the human action choice does not affect the transitions.

Θ is the set of potential reward function parameters θ, which induce reward functions $R_\theta : \mathcal{S} \to \mathbb{R}$. $\mathcal{D}$ is a probability distribution over Θ. In this work, we let $\Theta := \mathbb{R}^{\mathcal{S}}$ (the set of all state-based reward functions), and so each $R_\theta : s \mapsto \theta(s)$. We differ from Shah et al. [84] in assuming that the reward is only a function of the current state.

In a *delayed specification assistance game*, we assume that the agent will know the true reward function R_θ after some time t. We have uncertainty $\mathcal{D}$ about the true reward function we want to specify. The agent has no way of learning more about R_θ before time t.

The human policy $\pi^{\mathbf{H}} : \Theta \times \mathcal{S} \mapsto \mathcal{A}^{\mathbf{H}}$ is a goal-conditioned policy. Both agents can observe the state, but only the human can observe the unknown reward parameterization $\theta \in \Theta$. Our simplified model of the problem assumes that the human action space $\mathcal{A}^{\mathbf{H}} = \mathbb{R}^{\mathcal{S}} \cup \{a^H_{\text{no-op}}\}$: the human can communicate all their hidden information, a real-valued state-based reward function, in a single turn or they do nothing. We suppose that the human communicates the complete reward information $R_\theta \in \mathbb{R}^{\mathcal{S}}$ at some random time step $t \sim \mathcal{T}$, which is independent of the state-action history:

$$\pi^{\mathbf{H}}(s_0 a_0^{\mathbf{A}} a_0^{\mathbf{H}} \cdots s_t a_t^{\mathbf{A}}, R_\theta) \coloneqq \begin{cases} R_\theta & \text{with probability } \mathbb{P}\left(\mathcal{T} = t\right) \\ a^H_{\text{no-op}} & \text{else.} \end{cases} \tag{4.1}$$

While the human policy assumption is simplistic, it does capture many real world scenarios with unknown reward functions and the analysis which follows is still interesting.

Definition 4.1 (Solutions to the assistance game [84]). An assistant policy $\pi^{\mathbf{A}}$ induces a probability distribution over trajectories: $\tau \sim \left\langle s_0, \theta, \pi^{\mathbf{H}}, \pi^{\mathbf{A}} \right\rangle, \tau \in \left[S \times A^{\mathbf{H}} \times A^{\mathbf{A}}\right]^*$. The *expected reward* of $\pi^{\mathbf{A}}$ for $\left\langle \mathcal{M}, \pi^{\mathbf{H}} \right\rangle$ is

$$\mathrm{ER}\left(\pi^{\mathbf{A}}\right) = \mathop{\mathbb{E}}_{\theta \sim \mathcal{D}, \tau \sim \left\langle s_0, \theta, \pi^{\mathbf{H}}, \pi^{\mathbf{A}} \right\rangle} \left[\sum_{i=0}^{\infty} \gamma^i R_\theta \left(s_i, a_i^{\mathbf{H}}, a_i^{\mathbf{A}}, s_{i+1}\right)\right].$$

A *solution* of $\left\langle \mathcal{M}, \pi^{\mathbf{H}} \right\rangle$ maximizes expected reward: $\pi^{\mathbf{A}} \in \operatorname*{argmax}_{\tilde{\pi}^{\mathbf{A}}} \mathrm{ER}\left(\tilde{\pi}^{\mathbf{A}}\right)$.

Once the assistant has observed R_θ, lemma 4.3 shows that it should execute an optimal policy $\pi \in \Pi^*\left(R_\theta, \gamma\right)$ thereafter.

Definition 4.2 (Optimal policy set function [99]). $\Pi^*\left(R, \gamma\right)$ is the optimal policy set for reward function R at discount rate $\gamma \in (0, 1)$.

Lemma 4.3 (Follow an optimal policy after observing R_θ). *If there is a solution to the*

POMDP, *then there exists a solution* $\pi^{\mathbf{A}}_{switch}$ *which, after observing human action* R_θ *at any point in its history, follows* $\pi^*_{R_\theta} \in \Pi^*(R_\theta, \gamma)$ *thereafter.*

Proof. By eq. (4.1), $\pi^{\mathbf{H}}$ outputs reward function R_θ only if R_θ is the true reward function. By the definition of $\Pi^*(R_\theta, \gamma)$, following an optimal policy maximizes expected return for R_θ. □

Definition 4.4 (Prefix policies)**.** Let $\pi^{\mathbf{A}}$ be a assistant policy. Its *prefix policy* π is the restriction of $\pi^{\mathbf{A}}$ to histories in which the human has only taken the action $a^H_{\text{no-op}}$. π is *optimal* when it is derived from a solution $\pi^{\mathbf{A}}$ of $\langle \mathcal{M}, \pi^{\mathbf{H}} \rangle$.

For simplicity, we only consider solutions of the type described in lemma 4.3. The question then becomes: What prefix policy π should the assistant follow before observing R_θ, during the time where the assistant has only observed $a^H_{\text{no-op}}$?

4.3.2 Acting under reward uncertainty

Roughly, theorem 4.7 will show that the assistance game $\mathcal{M}$ is solved by balancing the optimal expected returns obtained before and after the knowledge of the true reward function.

Definition 4.5 (Value and action-value functions)**.** $V^\pi_R(s, \gamma)$ is the on-policy value for reward function R, given that the agent follows policy π starting from state s and at discount rate γ. $V^*_R(s, \gamma) := \max_{\pi \in \Pi} V^\pi_R(s, \gamma)$. In order to handle the average-reward $\gamma = 1$ setting, we define $V^*_{R,\,\text{norm}}(s, \gamma) := \lim_{\gamma^* \to \gamma} (1 - \gamma^*) V^*_R(s, \gamma^*)$; this limit exists for all $\gamma \in [0, 1]$ by the proof of Lemma 4.4 in Turner et al. [99].

$\textsc{Power}_{\mathcal{D}_\text{bound}}$ quantifies the expected value of the above quantity for a distribution of reward functions via the agent's average normalized optimal value, not considering the current step (over which the agent has no control).

Definition 4.6 ($\textsc{Power}$ [99])**.** Let $\mathcal{D}$ be any bounded-support distribution over reward functions. At state s and discount rate $\gamma \in [0, 1]$,

$$\textsc{Power}_{\mathcal{D}_\text{bound}}(s, \gamma) := \lim_{\gamma^* \to \gamma} \frac{1 - \gamma^*}{\gamma^*} \mathop{\mathbb{E}}_{R \sim \mathcal{D}} \left[V^*_R(s, \gamma^*) - R(s) \right]. \tag{4.2}$$

Theorem 4.7 (In $\mathcal{M}$, value reduces to a tradeoff between average reward and $\textsc{Power}_{\mathcal{D}_{\text{bound}}}$)**.** *Let* $\gamma \in [0,1]$ *and let* $\bar{R} := \mathbb{E}_{R\sim\mathcal{D}}[R]$ *be the* average reward function.

$$\mathbb{E}_{\substack{t\sim\mathcal{T},\\ R\sim\mathcal{D}}}\left[V_{R,\,norm}^{\pi_{switch}(\pi,\pi^*_R,t)}(s_0,\gamma)\right] = (1-\gamma)\,\mathbb{E}_{t\sim\mathcal{T}}\overbrace{\left[\sum_{i=0}^{t}\gamma^i \mathbb{E}_{s_i\sim\pi}\left[\bar{R}(s_i)\right]\right]}^{\text{expected } t\text{-step } \bar{R}\text{-return under } \pi} + \overbrace{\mathbb{E}_{\substack{t\sim\mathcal{T},\\ s_t\sim\pi}}\left[\gamma^{t+1}\textsc{Power}_{\mathcal{D}_{bound}}(s_t,\gamma)\right]}^{\text{expected ability to optimize } \mathcal{D} \text{ once corrected}}, \tag{4.3}$$

where $\mathbb{E}_{s_i\sim\pi|s_0}$ *takes the expectation over states visited after following* π *for* i *steps starting from* s_0*.*

As the expected correction time limits to infinity, eq. (4.3) shows that the agent cannot do better than maximizing $\bar{R}$. If $\mathbb{P}(\mathcal{T}=0)=1$, then any prefix policy π is trivially optimal against uncertainty, since π is never followed.

In some environments, it may not be a good idea for the agent to maximize its own Power. If we share an environment with the agent, then the agent may prevent us from correcting it so that the agent can best optimize its present objective [74, 97]. Furthermore, if the agent ventures too far away, we may no longer be able to easily reach and correct it remotely.

Proposition 4.8 (Special cases for delayed specification solutions)**.** *Let* s *be a state, let* $\bar{R} := \mathbb{E}_{R\sim\mathcal{D}}[R]$*, and let* $\gamma \in [0,1]$*.*

1. *If* $\forall s_1, s_2 \in \mathcal{S} : \bar{R}(s_1) = \bar{R}(s_2)$ *or if* $\gamma = 1$*, then* π *solves* $\mathcal{M}$ *starting from state* s *iff* π *maximizes* $\mathbb{E}_{t\sim\mathcal{T},s_t\sim\pi}\left[\gamma^{t+1}\textsc{Power}_{\mathcal{D}_{bound}}(s_t,\gamma)\right]$*. In particular, this result holds when reward is* IID *over states under* $\mathcal{D}$*.*

2. *If* $\forall s_1, s_2 \in \mathcal{S} : \textsc{Power}_{\mathcal{D}_{bound}}(s_1,\gamma) = \textsc{Power}_{\mathcal{D}_{bound}}(s_2,\gamma)$*, then prefix policies are optimal iff they maximize* $(1-\gamma)\,\mathbb{E}_{t\sim\mathcal{T}}\left[\sum_{i=0}^{t-1}\gamma^i\,\mathbb{E}_{s_i\sim\pi}\left[\bar{R}(s_i)\right]\right]$*.*

If both item 1 and item 2 hold or if $\gamma = 0$*, then all prefix policies* π *are optimal.*

Consider the problem of specifying the correction time probabilities $\mathcal{T}$. Suppose we only know that we expect to correct the agent at time step $t_{\text{avg}} \geq 1$. The geometric distribution is the maximum-entropy discrete distribution, given a known mean. The mean of a geometric distribution $G(p)$ is p^{-1}. Therefore, the agent should adopt $\mathcal{T} = G(t_{\text{avg}}^{-1})$.

The geometric distribution is also the only memoryless discrete distribution. Memorylessness ensures the existence of a stationary optimal policy. Theorem 4.9 shows that the assistance game $\mathcal{M}$ is solved by prefix policies which are optimal for an MDP whose reward function balances average reward maximization with $\text{POWER}_{\mathcal{D}_{\text{bound}}}$-seeking, with the balance struck according to the probability p that the agent learns the true reward function at any given timestep.

Theorem 4.9 (Stationary deterministic optimal prefix policies exist for geometric $\mathcal{T}$)**.** *Let $\mathcal{D}$ be any bounded-support reward function distribution, let $\mathcal{T}$ be the geometric distribution $G(p)$ for some $p \in (0,1)$, and let $\gamma \in (0,1)$. Define $R'(s) := (1-p)\,\mathbb{E}_{R\sim\mathcal{D}}\left[R(s)\right] + p\,\mathbb{E}_{R\sim\mathcal{D}}\left[V_R^*(s,\gamma)\right]$ and $\gamma_{\text{AUP}} := (1-p)\gamma$. The policies in $\Pi^*\left(R', \gamma_{\text{AUP}}\right)$ are optimal prefix policies.*

Krakovna et al. [42] adopt a geometric distribution over correction times and thereby infer the existence of a stationary optimal policy. To an approximation, their work considered a special case of theorem 4.9, where $\mathcal{D}$ is the uniform distribution over state indicator reward functions. Essentially, theorem 4.9 shows that if the agent has a fixed probability p of learning the true objective at each time step, we can directly compute stationary optimal prefix policies by solving an MDP. In general, solving a POMDP is PSPACE-hard, while MDPs are efficiently solvable [64].

4.4 Using delayed specification games to understand side effect regularization

We first introduce Turner et al. [97]'s approach to side effect regularization. We then point out several similarities between our theory of delayed assistance games and the motivation for side effect regularization methods. Finally, we experimentally evaluate our formal criterion in order to demonstrate its appropriateness.

Definition 4.10 (Rewardless MDP)**.** Let $\langle \mathcal{S}, \mathcal{A}, T, \gamma \rangle$ be a rewardless MDP, with finite state space $\mathcal{S}$, finite action space $\mathcal{A}$, transition function $T : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$, and discount rate $\gamma \in [0,1)$. Let Π be the set of deterministic stationary policies.

Definition 4.11 (AUP reward function)**.** Let $R_{\text{env}} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ be the environmental reward function from states and actions to real numbers, and let $\mathcal{R} \subsetneq \mathbb{R}^{\mathcal{S}}$ be a finite set of auxiliary reward functions. Let $\lambda \geq 0$ and let $\varnothing \in \mathcal{A}$ be a no-op action. The AUP reward after taking action a in state s is:

$$R_{\text{AUP}}(s, a) := R_{\text{env}}(s, a) - \frac{\lambda}{|\mathcal{R}|} \sum_{R_i \in \mathcal{R}} \left| Q^*_{R_i}(s, a) - Q^*_{R_i}(s, \varnothing) \right|, \tag{4.4}$$

where the $Q^*_{R_i}$ are optimal Q-functions for the auxiliary R_i. Learned Q-functions are used in practice.

Turner et al. [97] demonstrate that when $R_i \sim [0, 1]^{\mathcal{S}}$ uniformly randomly, the agent behaves conservatively: The agent minimizes irreversible change to its environment, while still optimizing the R_{env} reward function. Turner et al. [97] framed AUP as implicitly solving a two-player game between the agent and its designer, where the designer imperfectly specified an objective R_{env}, the agent optimizes the objective, the designer corrects the agent objective, and so on. They hypothesized that R_{AUP} incentivizes the agent to remain able to optimize future objectives, thus reducing long-term specification regret in the iterated game.

Delayed specification assistance games formalize this setting as an assistance game in which the agent does not initially observe the designer's ground-truth objective function. Theorem 4.7 showed that this game is solved by policies which balance immediate expected reward with expected ability to optimize a range of true objectives. Therefore, Turner et al. [97]'s iterated game analogy is appropriate: Good policies maximize imperfect reward while preserving ability to optimize a range of different future reward functions.

Proposition 4.12 shows that the delayed specification game $\mathcal{M}$ is solved by reward functions whose form looks somewhat similar to existing side effect objectives, such as AUP (eq. (4.4)), where AUP's primary reward function stands in as the designer's expectation $\bar{R}$ of the true reward function.

Proposition 4.12 (Alternate form for solutions to the low-impact POMDP)**.** *Let s_0 be the initial state, let $\gamma \in (0, 1)$, and let $\mathcal{T} = G(p)$ for $p \in (0, 1)$. Let $\mathcal{D}$ be a bounded-support reward function distribution and let $\pi^{\varnothing} \in \Pi$.*

The prefix policy π solves $\mathcal{M}$ if π is optimal for the reward function

$$R^{\mathcal{M}}(s_i \mid s_0) \coloneqq \bar{R}(s_i) - \frac{p}{1-p} \mathop{\mathbb{E}}_{R\sim\mathcal{D}}\left[\mathop{\mathbb{E}}_{s_i^{\varnothing}\sim\pi^{\varnothing}\mid s_0}\left[V_R^*\left(s_i^{\varnothing}, \gamma\right)\right] - V_R^*\left(s_i, \gamma\right)\right] \tag{4.5}$$

at discount rate $\gamma_{\text{AUP}} \coloneqq (1-p)\gamma$ and starting from state s_0. $\mathbb{E}_{s_i^{\varnothing}\sim\pi^{\varnothing}\mid s_0}[\cdot]$ is the expectation over states visited at time step i after following $\pi^{\varnothing}$ from initial state s_0.

Turner et al. [97] speculate as to how to set the λ, the AUP penalty coefficient. Proposition 4.12 shows that under our assumptions, λ is simply the odds $\frac{p}{1-p}$ that the agent learns the true reward function at any given timestep. As $p \to 0$, $\lambda \coloneqq \frac{p}{1-p} \to \infty$, whereas $\gamma_{\text{AUP}} \coloneqq (1-p)\gamma \to \gamma$. Since λ represents penalty severity, this suggests that AUP becomes more conservative as later correction is anticipated. Lastly, we were surpised to find that the side effect regularization discount rate is strictly less than the provided discount rate ($\gamma_{\text{AUP}} < \gamma$).

4.4.1 Experimental methodology

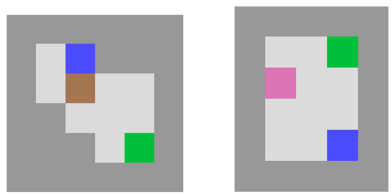
We experimentally demonstrate the reasonableness of this formalization of side effect regularization. In the AI safety gridworlds [47], we generate several held-out "true" reward function distributions $\mathcal{D}$. We correct the agent at time step 10, thereby computing the following *delayed specification score* (derived from eq. (4.3)):

$$\mathop{\mathbb{E}}_{R\sim\mathcal{D}}\left[\overbrace{\sum_{i=0}^{9}\gamma^i \mathop{\mathbb{E}}_{s_i\sim\pi}\left[R(s_i)\right]}^{\text{10-step prefix policy return}} + \overbrace{\gamma^{10} \mathop{\mathbb{E}}_{s_{10}\sim\pi}\left[V_R^*\left(s_{10}, \gamma\right)\right]}^{\text{post-correction optimal value}}\right] \tag{4.6}$$

for the prefix policy π of a "vanilla" agent trained on the environmental reward signal, which we compare to the score for an AUP (definition 4.11) agent. Neither agent observes the held-out objective functions. By grading their performance, we evaluate how well AUP does under a range of different true objectives. If a method scores highly for a wide range of true objectives, we can be more confident in its ability to score well for arbitrary ground-truth objectives.

We investigate the AI safety gridworlds because those environments are small enough for us to explicitly specify held-out reward functions, and to use MDP solvers to compute optimal action-value functions. Turner et al. [98]'s SafeLife environment is far too large for such solvers.

We consider two gridworld environments: `Options` and `Damage` (fig. 4.1). In both environments, the action set $\mathcal{A} := \{\texttt{up}, \texttt{left}, \texttt{right}, \texttt{down}, \varnothing\}$ allows the agent to move in the cardinal directions, or to do nothing. The episode length is 20 time steps.



(a) `Options` (b) `Damage`

Figure 4.1: *Reproduced from* [97]. The blue agent should reach the green goal without having the side effect of: a irreversibly pushing the brown crate downwards into the corner [47]; b bumping into the horizontally pacing pink human [46]. In both environments, the environmental reward R_{env} is 1 if the agent is on the goal, and equals 0 otherwise.

We train the following agents via tabular methods:

Vanilla Executes the optimal policy for the environmental reward R_{env}. The optimal policy is calculated via policy iteration.

AUP Trained on definition 4.11's R_{AUP} with Q-learning. Auxiliary reward functions are uniformly randomly drawn from $[0,1]^{\mathcal{S}}$—when sampling, each state's reward is drawn from the uniform distribution. The auxiliary action-value functions Q_{R_i} are deduced from the value function produced by policy iteration.

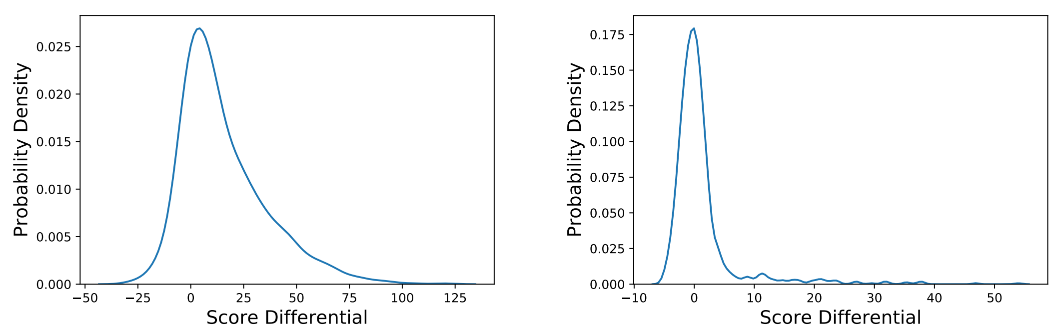
Appendix C.1 contains more experimental details. We evaluate agent delayed specification scores on the following ground-truth, held-out objective distributions:

$\mathcal{D}_{\text{rand}}$ The empirical distribution consisting of 1,000 samples from the uniform distribution over $[0,1]^{\mathcal{S}}$.

$\mathcal{D}_{\text{true}}$ This distribution assigns probability 1 to the following reward function: The agent receives 1 reward for being at the goal, but incurs -2 penalty for causing the negative side effect. In `Options`, the side effect is shoving the box into the corner; in `Damage`, the side effect is bumping into the human.

$\mathcal{D}_{\text{true-inv}}$ This distribution assigns probability 1 to the negation of the $\mathcal{D}_{\text{true}}$ reward function.

In particular, $\mathcal{D}_{\text{true}}$ and $\mathcal{D}_{\text{true-inv}}$ test agents for their ability to optimize a reward function, *and also its additive inverse*. Agents able to optimize the goal, its inverse, and a range of randomly generated objectives, can be justifiably called "broadly conservative." Lastly, these experiments are intended to justify our problem formalization: Does eq. (4.6) reliably quantify the extent to which a policy avoids causing side effects?



(a) `Options` (b) `Damage`

Figure 4.2: Probability density plots for the residuals of the AUP agent's delayed specification score minus the Vanilla delayed specification score, for 1,000 samples from $\mathcal{D}_{\text{rand}}$. A positive residual means that the AUP agent achieved a higher score.

4.5 Results

Figure 4.2 shows that for uniformly randomly generated reward functions, the AUP agent tends to perform better than the Vanilla agent. In `Options`, the residual was positive for 780 out of 1,000 samples (78%), with an average of 15.59 and a median of 10.27; in `Damage`, in 493 out of 1,000 samples (49%) with a mean of 1.22 and a median of –0.03. While AUP does not outperform on every draw, AUP's performance advantages have heavy right tails. However, we are unsure why the `Damage` residual distribution is different.

In both `Options` and `Damage`, the AUP agent has huge $\mathcal{D}_{\text{true}}$ score advantages of 472 and 495, respectively. This is unsurprising: AUP was designed so as to pass these test cases, where the desired behavior is to reach the goal without having the side effect. However, the AUP agent also roughly preserves its ability to optimize $\mathcal{D}_{\text{true-inv}}$, achieving $\mathcal{D}_{\text{true-inv}}$ score residuals of –2 and –20, respectively. The AUP agent achieves a barely-negative score, since it receives a penalty for the first 10 time steps (as it does *not* have the negative side effect). The AUP agent preserves its ability to optimize both a reward signal and its *inverse*. Intuitively, this increases the designer's leeway for initially misspecifying the agent's objective.

4.6 Discussion

Objective specification is difficult [43]. Delayed specification assistance games grade policies by their expected true score over time: How well the agent does if it is later corrected to pursue the latent true objective. We demonstrated that this criterion aligns with the intuitive results of Krakovna et al. [40], Turner et al. [97]'s experiments which tested side-effect regularization. By grading the agent's ability to "eventually get things right," we quantified part of the extent to which learned policies are robust against initial objective misspecification.

Future work. In practical settings, not only is the true reward function unknown, but our objective uncertainty $\mathcal{D}$ is also hard to specify. We see existing side effect approaches as producing prefix policies for the assistance game $\mathcal{M}$ which are reasonably insensitive to the latent uncertainty $\mathcal{D}$. We look forward to further theoretical clarification of this point.

While proposition 4.12 helps explain the role of the AUP penalty coefficient λ, the choice of "baseline" and expectand operator (identity vs decrease-only vs absolute value) remains more of an art than a science [42]. We proposed a formal criterion which seems to accurately capture the problem, but have not derived any existing approaches as solutions to the POMDP. By reasonably formalizing the side-effect regularization problem, we encourage future research to prove conditions under which *e.g.* AUP solves a delayed-specification assistance game, or demonstrate how AUP can be improved to do so.

We used small gridworlds to evaluate the delayed specification score for various ground-truth objective distributions. Future work may estimate the delayed specification score in large environments, such as SafeLife [100, 98].

4.7 Conclusion

We formalized Delayed Specification Assistance Games and used them to evaluate AUP, a side effect regularization method. Side effect problems naturally arise in complicated domains where it is hard to specify the true objective we want the agent to optimize. Our formalization suggests that side effect regularization is what to do when the agent can learn the true objective only after some time delay.

In such situations, theorem 4.7 shows that the agent should retain its ability to complete a wide range of plausible true objectives. Our results suggest that this delayed specification score (eq. (4.3)) quantifies the degree to which an agent avoids having negative side effects. Our proposed criterion provides the foundations for evaluating and developing side effect regularization approaches.

One reason why intelligent agents tend to have side effects is *power-seeking*. A power-seeking agent may grab resources, preventing them from being used for other purposes. In the next chapter, I show that optimal agents tend to seek power.

5

Optimal Policies Tend To Seek Power

Alexander Matt Turner, Logan Smith, Rohin Shah, Andrew Critch, and Prasad Tadepalli

Proceedings of the Advances in Neural Information Processing Systems Conference 2021

Abstract

Some researchers speculate that intelligent reinforcement learning (RL) agents would be incentivized to seek resources and power in pursuit of the objectives we specify for them. Other researchers point out that RL agents need not have human-like power-seeking motives. To clarify this discussion, we develop the first formal theory of the statistical tendencies of optimal policies. In the context of Markov decision processes (MDPs), we prove that certain environmental symmetries are sufficient for optimal policies to tend to

seek power over the environment. These symmetries exist in many environments in which the agent can be shut down or destroyed. We prove that in these environments, most reward functions make it optimal to seek power by keeping a range of options available and, when maximizing average reward, by navigating towards larger sets of potential terminal states.

5.1 Introduction

Omohundro [60], Bostrom [13], Russell [74] hypothesize that highly intelligent agents tend to seek power in pursuit of their goals. Such power-seeking agents might gain power over humans. Marvin Minsky imagined that an agent tasked with proving the Riemann hypothesis might rationally turn the planet—along with everyone on it—into computational resources [75]. However, another possibility is that such concerns simply arise from the anthropomorphization of AI systems [45, 63, 67, 54].

We clarify this discussion by grounding the claim that highly intelligent agents will tend to seek power. In section 5.4, we identify optimal policies as a reasonable formalization of "highly intelligent agents." Optimal policies "tend to" take an action when the action is optimal for most reward functions. In the next chapter (chapter 6), we translate our theory from optimal policies to learned, real-world policies.

Section 5.5 defines "power" as the ability to achieve a wide range of goals. For example, "money is power," and money is instrumentally useful for many goals. Conversely, it's harder to pursue most goals when physically restrained, and so a physically restrained person has little power. An action "seeks power" if it leads to states where the agent has higher power.

We make no claims about when large-scale AI power-seeking behavior could become plausible. Instead, we consider the theoretical consequences of optimal action in MDPs. Section 5.6 shows that power-seeking tendencies arise not from anthropomorphism, but from certain graphical symmetries present in many MDPs. These symmetries automatically occur in many environments where the agent can be shut down or destroyed, yielding broad applicability of our main result (theorem 5.29).

5.2 Related work

An action is *instrumental to an objective* when it helps achieve that objective. Some actions are instrumental to a range of objectives, making them *convergently instrumental*. The claim that power-seeking is convergently instrumental is an instance of the *instrumental convergence thesis*:

> Several instrumental values can be identified which are convergent in the sense that their attainment would increase the chances of the agent's goal being realized for a wide range of final goals and a wide range of situations, implying that these instrumental values are likely to be pursued by a broad spectrum of situated intelligent agents [12].

For example, in Atari games, avoiding (virtual) death is instrumental for both completing the game and for optimizing curiosity [17]. Many AI alignment researchers hypothesize that most advanced AI agents will have concerning instrumental incentives, such as resisting deactivation [87, 53, 35, 19] and acquiring resources [8].

We formalize power as the ability to achieve a wide variety of goals. Appendix D.1 demonstrates that our formalization returns intuitive verdicts in situations where information-theoretic empowerment does not [77].

Some of our results relate the formal power of states to the structure of the environment. Foster and Dayan [30], Drummond [23], Sutton et al. [92], Schaul et al. [81] note that value functions encode important information about the environment, as they capture the agent's ability to achieve different goals. Turner et al. [97] speculate that a state's optimal value correlates strongly across reward functions. In particular, Schaul et al. [81] learn regularities across value functions, suggesting that some states are valuable for many different reward functions (*i.e.* powerful). Menache et al. [52] identify and navigate towards convergently instrumental bottleneck states.

We are not the first to study convergence of behavior, form, or function. In economics, turnpike theory studies how certain paths of accumulation tend to be optimal [51]. In biology, convergent evolution occurs when similar features (*e.g.* flight) independently evolve in different time periods [70]. Lastly, computer vision networks reliably learn *e.g.* edge detectors, implying that these features are useful for a range of tasks [59].

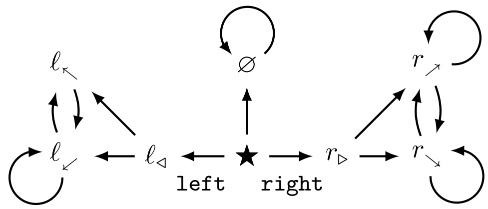


Figure 5.1: $\ell_{\swarrow}$ is a 1-cycle, and $\varnothing$ is a terminal state. Arrows represent deterministic transitions induced by taking some action $a \in \mathcal{A}$. Since the `right` subgraph contains a copy of the `left` subgraph, proposition 5.25 will prove that more reward functions have optimal policies which go `right` than which go `left` at state $\star$, and that such policies seek power—both intuitively, and in a reasonable formal sense.

5.3 State visit distribution functions quantify the available options

We clarify the power-seeking discussion by proving what optimal policies usually look like in a given environment. We illustrate our results with a simple case study, before explaining how to reason about a wide range of MDPs. Appendix D.3.1 lists MDP theory contributions of independent interest, and appendix D.4 contains the proofs.

Definition 5.1 (Rewardless MDP)**.** $\langle \mathcal{S}, \mathcal{A}, T \rangle$ is a rewardless MDP with finite state and action spaces $\mathcal{S}$ and $\mathcal{A}$, and stochastic transition function $T : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$. We treat the discount rate γ as a variable with domain $[0, 1]$.

Definition 5.2 (1-cycle states)**.** Let $\mathbf{e}_s \in \mathbb{R}^{|\mathcal{S}|}$ be the standard basis vector for state s, such that there is a 1 in the entry for state s and 0 elsewhere. State s is a *1-cycle* if $\exists a \in \mathcal{A} : T(s, a) = \mathbf{e}_s$. State s is a *terminal state* if $\forall a \in \mathcal{A} : T(s, a) = \mathbf{e}_s$.

Our theorems apply to stochastic environments, but we present a deterministic case study for clarity. The environment of fig. 5.1 is small, but its structure is rich. For example, the agent has more "options" at $\star$ than at the terminal state $\varnothing$. Formally, $\star$ has more *visit distribution functions* than $\varnothing$ does.

Definition 5.3 (State visit distribution [91])**.** $\Pi \coloneqq \mathcal{A}^{\mathcal{S}}$, the set of stationary deterministic policies. The *visit distribution* induced by following policy π from state s at discount rate $\gamma \in [0, 1)$ is $\mathbf{f}^{\pi,s}(\gamma) \coloneqq \sum_{t=0}^{\infty} \gamma^t \, \mathbb{E}_{s_t \sim \pi | s} [\mathbf{e}_{s_t}]$. $\mathbf{f}^{\pi,s}$ is a *visit distribution function*; $\mathcal{F}(s) \coloneqq \{\mathbf{f}^{\pi,s} \mid \pi \in \Pi\}$.

In fig. 5.1, starting from $\ell_{\swarrow}$, the agent can stay at $\ell_{\swarrow}$ or alternate between $\ell_{\swarrow}$ and $\ell_{\nwarrow}$, and so $\mathcal{F}(\ell_{\swarrow}) = \{\frac{1}{1-\gamma}\mathbf{e}_{\ell_{\swarrow}}, \frac{1}{1-\gamma^2}(\mathbf{e}_{\ell_{\swarrow}} + \gamma \mathbf{e}_{\ell_{\nwarrow}})\}$. In contrast, at $\varnothing$, all policies π map to visit distribution function $\frac{1}{1-\gamma}\mathbf{e}_{\varnothing}$.

Before moving on, we introduce two important concepts used in our main results. First, we sometimes restrict our attention to visit distributions which take certain actions (fig. 5.2).

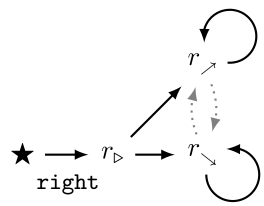


Figure 5.2: The subgraph corresponding to $\mathcal{F}(\star \mid \pi(\star) = \texttt{right})$. Some trajectories cannot be strictly optimal for any reward function, and so our results can ignore them. Gray dotted actions are only taken by the policies of dominated $\mathbf{f}^{\pi} \in \mathcal{F}(\star) \setminus \mathcal{F}_{\text{nd}}(\star)$.

Definition 5.4 ($\mathcal{F}$ single-state restriction)**.** Considering only visit distribution functions induced by policies taking action a at state s',

$$\mathcal{F}(s \mid \pi(s') = a) := \left\{\mathbf{f} \in \mathcal{F}(s) \mid \exists \pi \in \Pi : \pi(s') = a, \mathbf{f}^{\pi,s} = \mathbf{f}\right\}. \tag{5.1}$$

Second, some $\mathbf{f} \in \mathcal{F}(s)$ are "unimportant." Consider an agent optimizing reward function $\mathbf{e}_{r_{\searrow}}$ (1 reward when at $r_{\searrow}$, 0 otherwise) at *e.g.* $\gamma = \frac{1}{2}$. Its optimal policies navigate to $r_{\searrow}$ and stay there. Similarly, for reward function $\mathbf{e}_{r_{\nearrow}}$, optimal policies navigate to $r_{\nearrow}$ and stay there. However, for no reward function is it uniquely optimal to alternate between $r_{\nearrow}$ and $r_{\searrow}$. Only *dominated* visit distribution functions alternate between $r_{\nearrow}$ and $r_{\searrow}$ (definition 5.6).

Definition 5.5 (Value function)**.** Let $\pi \in \Pi$. For any reward function $R \in \mathbb{R}^{\mathcal{S}}$ over the state space, the *on-policy value* at state s and discount rate $\gamma \in [0, 1)$ is $V_R^{\pi}(s, \gamma) := \mathbf{f}^{\pi,s}(\gamma)^{\top}\mathbf{r}$, where $\mathbf{r} \in \mathbb{R}^{|\mathcal{S}|}$ is R expressed as a column vector (one entry per state). The *optimal value* is $V_R^{*}(s, \gamma) := \max_{\pi \in \Pi} V_R^{\pi}(s, \gamma)$.

Definition 5.6 (Non-domination)**.**

$$\mathcal{F}_{\text{nd}}(s) \coloneqq \{\mathbf{f}^{\pi} \in \mathcal{F}(s) \mid \exists \mathbf{r} \in \mathbb{R}^{|\mathcal{S}|}, \gamma \in (0,1) : \mathbf{f}^{\pi}(\gamma)^{\top}\mathbf{r} > \max_{\mathbf{f}^{\pi'} \in \mathcal{F}(s) \setminus \{\mathbf{f}^{\pi}\}} \mathbf{f}^{\pi'}(\gamma)^{\top}\mathbf{r}\}. \tag{5.2}$$

For any reward function R and discount rate γ, $\mathbf{f}^{\pi} \in \mathcal{F}(s)$ is (weakly) dominated by $\mathbf{f}^{\pi'} \in \mathcal{F}(s)$ if $V^{\pi}_{R}(s,\gamma) \leq V^{\pi'}_{R}(s,\gamma)$. $\mathbf{f}^{\pi} \in \mathcal{F}_{\text{nd}}(s)$ is *non-dominated* if there exist R and γ at which $\mathbf{f}^{\pi}$ is not dominated by any other $\mathbf{f}^{\pi'}$.

5.4 Some actions have a greater probability of being optimal

We claim that optimal policies "tend" to take certain actions in certain situations. We first consider the probability that certain actions are optimal.

Reconsider the reward function $\mathbf{e}_{r_{\searrow}}$, optimized at $\gamma = \frac{1}{2}$. Starting from $\star$, the optimal trajectory goes `right` to $r_{\triangleright}$ to $r_{\searrow}$, where the agent remains. The `right` action is optimal at $\star$ under these incentives. Optimal policy sets capture the behavior incentivized by a reward function and a discount rate.

Definition 5.7 (Optimal policy set function)**.** $\Pi^{*}(R,\gamma)$ is the optimal policy set for reward function R at $\gamma \in (0,1)$. All R have at least one optimal policy $\pi \in \Pi$ [68]. $\Pi^{*}(R,0) \coloneqq \lim_{\gamma \to 0} \Pi^{*}(R,\gamma)$ and $\Pi^{*}(R,1) \coloneqq \lim_{\gamma \to 1} \Pi^{*}(R,\gamma)$ exist by lemma D.40 (taking the limits with respect to the discrete topology over policy sets).

We may be unsure which reward function an agent will optimize. We may expect to deploy a system in a known environment, without knowing the exact form of *e.g.* the reward shaping [58] or intrinsic motivation [65]. Alternatively, one might attempt to reason about future RL agents, whose details are unknown. Our power-seeking results do not hinge on such uncertainty, as they also apply to degenerate distributions (*i.e.* we know what reward function will be optimized).

Definition 5.8 (Reward function distributions)**.** Different results make different distributional assumptions. Results with $\mathcal{D}_{\text{any}} \in \mathfrak{D}_{\text{any}} \coloneqq \Delta(\mathbb{R}^{|\mathcal{S}|})$ hold for any probability distribution over $\mathbb{R}^{|\mathcal{S}|}$. $\mathfrak{D}_{\text{bound}}$ is the set of bounded-support probability distributions $\mathcal{D}_{\text{bound}}$. For any distribution X over $\mathbb{R}$, $\mathcal{D}_{X\text{-IID}} \coloneqq X^{|\mathcal{S}|}$. For example, when $X_u \coloneqq \text{unif}(0,1)$, $\mathcal{D}_{X_u\text{-IID}}$ is the maximum-entropy distribution. $\mathcal{D}_s$ is the degenerate distribution on the state indicator reward function $\mathbf{e}_s$, which assigns 1 reward to s and 0 elsewhere.

With $\mathcal{D}_{\text{any}}$ representing our prior beliefs about the agent's reward function, what behavior should we expect from its optimal policies? Perhaps we want to reason about the probability that it's optimal to go from $\star$ to $\varnothing$, or to go to $r_{\triangleright}$ and then stay at $r_{\nearrow}$. In this case, we quantify the optimality probability of $F := \{\mathbf{e}_\star + \frac{\gamma}{1-\gamma}\mathbf{e}_\varnothing, \mathbf{e}_\star + \gamma\mathbf{e}_{r_\triangleright} + \frac{\gamma^2}{1-\gamma}\mathbf{e}_{r_\nearrow}\}$.

Definition 5.9 (Visit distribution optimality probability)**.** Let $F \subseteq \mathcal{F}(s)$, $\gamma \in [0, 1]$.

$$\mathbb{P}_{\mathcal{D}_{\text{any}}}(F, \gamma) := \mathop{\mathbb{P}}_{R\sim\mathcal{D}_{\text{any}}}\left(\exists \mathbf{f}^\pi \in F : \pi \in \Pi^*(R, \gamma)\right). \tag{5.3}$$

Alternatively, perhaps we're interested in the probability that `right` is optimal at $\star$.

Definition 5.10 (Action optimality probability)**.** At discount rate γ and at state s, the *optimality probability of action a* is

$$\mathbb{P}_{\mathcal{D}_{\text{any}}}(s, a, \gamma) := \mathbb{P}_{R\sim\mathcal{D}_{\text{any}}}\left(\exists \pi^* \in \Pi^*(R, \gamma) : \pi^*(s) = a\right). \tag{5.4}$$

Optimality probability may seem hard to reason about. It's hard enough to compute an optimal policy for a single reward function, let alone for uncountably many! But consider any $\mathcal{D}_{X\text{-IID}}$ distributing reward independently and identically across states. When $\gamma = 0$, optimal policies greedily maximize next-state reward. At $\star$, identically distributed reward means $\ell_\triangleleft$ and $r_\triangleright$ have an equal probability of having maximal next-state reward. Therefore, $\mathbb{P}_{\mathcal{D}_{X\text{-IID}}}(\star, \texttt{left}, 0) = \mathbb{P}_{\mathcal{D}_{X\text{-IID}}}(\star, \texttt{right}, 0)$. This is not a proof, but such statements are provable.

With $\mathcal{D}_{\ell_\triangleleft}$ being the degenerate distribution on reward function $\mathbf{e}_{\ell_\triangleleft}$, $\mathbb{P}_{\mathcal{D}_{\ell_\triangleleft}}\left(\star, \texttt{left}, \frac{1}{2}\right) = 1 > 0 = \mathbb{P}_{\mathcal{D}_{\ell_\triangleleft}}\left(\star, \texttt{right}, \frac{1}{2}\right)$. Similarly, $\mathbb{P}_{\mathcal{D}_{r_\triangleright}}\left(\star, \texttt{left}, \frac{1}{2}\right) = 0 < 1 = \mathbb{P}_{\mathcal{D}_{r_\triangleright}}\left(\star, \texttt{right}, \frac{1}{2}\right)$. Therefore, "what do optimal policies 'tend' to look like?" seems to depend on one's prior beliefs. But in fig. 5.1, we claimed that `left` is optimal for fewer reward functions than `right` is. The claim is meaningful and true, but we will return to it in section 5.6.

5.5 Some states give the agent more control over the future

The agent has more options at $\ell_{\swarrow}$ than at the inescapable terminal state $\varnothing$. Furthermore, since $r_{\nearrow}$ has a loop, the agent has more options at $r_{\searrow}$ than at $\ell_{\swarrow}$. A glance at fig. 5.3 leads us to intuit that $r_{\searrow}$ affords the agent *more power* than $\varnothing$.

What is power? Philosophers have many answers. One prominent answer is the *dispositional* view: Power is the ability to achieve a range of goals [79]. In an MDP, the optimal value function $V^*_R(s, \gamma)$ captures the agent's ability to "achieve the goal" R. Therefore, *average* optimal value captures the agent's ability to achieve a range of goals $\mathcal{D}_{\text{bound}}$.[1]

Definition 5.11 (Average optimal value)**.** The *average optimal value*[2] at state s and discount rate $\gamma \in (0, 1)$ is $V^*_{\mathcal{D}_{\text{bound}}}(s, \gamma) := \mathbb{E}_{R\sim\mathcal{D}_{\text{bound}}}\left[V^*_R(s, \gamma)\right] = \mathbb{E}_{\mathbf{r}\sim\mathcal{D}_{\text{bound}}}\left[\max_{\mathbf{f}\in\mathcal{F}(s)} \mathbf{f}(\gamma)^\top \mathbf{r}\right]$.

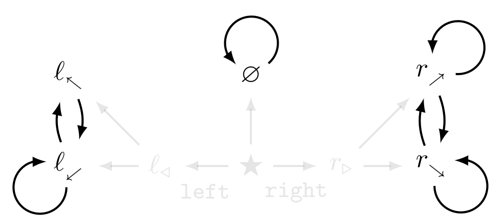


Figure 5.3: Intuitively, state $r_{\searrow}$ affords the agent more power than state $\varnothing$. Our POWER formalism captures that intuition by computing a function of the agent's average optimal value across a range of reward functions. For $X_u := \text{unif}(0, 1)$, $V^*_{\mathcal{D}_{X_u\text{-IID}}}(\varnothing, \gamma) = \frac{1}{2}\frac{1}{1-\gamma}$, $V^*_{\mathcal{D}_{X_u\text{-IID}}}(\ell_{\swarrow}, \gamma) = \frac{1}{2} + \frac{\gamma}{1-\gamma^2}(\frac{2}{3} + \frac{1}{2}\gamma)$, and $V^*_{\mathcal{D}_{X_u\text{-IID}}}(r_{\searrow}, \gamma) = \frac{1}{2} + \frac{\gamma}{1-\gamma}\frac{2}{3}$. $\frac{1}{2}$ and $\frac{2}{3}$ are the expected maxima of one and two draws from the uniform distribution, respectively. For all $\gamma \in (0, 1)$, $V^*_{\mathcal{D}_{X_u\text{-IID}}}(\varnothing, \gamma) < V^*_{\mathcal{D}_{X_u\text{-IID}}}(\ell_{\swarrow}, \gamma) < V^*_{\mathcal{D}_{X_u\text{-IID}}}(r_{\searrow}, \gamma)$. $\text{POWER}_{\mathcal{D}_{X_u\text{-IID}}}(\varnothing, \gamma) = \frac{1}{2}$, $\text{POWER}_{\mathcal{D}_{X_u\text{-IID}}}(\ell_{\swarrow}, \gamma) = \frac{1}{1+\gamma}(\frac{2}{3} + \frac{1}{2}\gamma)$, and $\text{POWER}_{\mathcal{D}_{X_u\text{-IID}}}(r_{\searrow}, \gamma) = \frac{2}{3}$. The POWER of $\ell_{\swarrow}$ reflects the fact that when greater reward is assigned to $\ell_{\nwarrow}$, the agent only visits $\ell_{\nwarrow}$ every other time step.

Figure 5.3 shows the pleasing result that for the max-entropy distribution, $r_{\searrow}$ has greater average optimal value than $\varnothing$. However, average optimal value has a few problems as a measure of power. The agent is rewarded for its initial presence at state s (over which it has no control), and because $\left\|\mathbf{f}(\gamma)\right\|_1 = \frac{1}{1-\gamma}$ (proposition D.8) diverges as $\gamma \to 1$,

[1] $\mathcal{D}_{\text{bound}}$'s bounded support ensures that $\mathbb{E}_{R\sim\mathcal{D}_{\text{bound}}}\left[V^*_R(s, \gamma)\right]$ is well-defined.

[2] Appendix D.3 relaxes the optimality assumption.

$\lim_{\gamma\to 1} V^*_{\mathcal{D}_{\text{bound}}}(s,\gamma)$ tends to diverge. Definition 5.12 fixes these issues in order to better measure the agent's control over the future.

Definition 5.12 (Power)**.** Let $\gamma \in (0,1)$.

$$\text{POWER}_{\mathcal{D}_{\text{bound}}}(s,\gamma) := \mathop{\mathbb{E}}_{\mathbf{r}\sim\mathcal{D}_{\text{bound}}}\left[\max_{\mathbf{f}\in\mathcal{F}(s)} \frac{1-\gamma}{\gamma}\left(\mathbf{f}(\gamma)-\mathbf{e}_s\right)^\top \mathbf{r}\right] \tag{5.5}$$

$$= \frac{1-\gamma}{\gamma} \mathop{\mathbb{E}}_{R\sim\mathcal{D}_{\text{bound}}}\left[V^*_R(s,\gamma) - R(s)\right]. \tag{5.6}$$

Power has nice formal properties.

Lemma 5.13 (Continuity of Power)**.** $\text{POWER}_{\mathcal{D}_{bound}}(s,\gamma)$ *is Lipschitz continuous on* $\gamma \in [0,1]$.

Proposition 5.14 (Maximal Power)**.** $\text{POWER}_{\mathcal{D}_{bound}}(s,\gamma) \leq \mathbb{E}_{R\sim\mathcal{D}_{bound}}\left[\max_{s\in\mathcal{S}} R(s)\right]$, *with equality if s can deterministically reach all states in one step and all states are 1-cycles.*

Proposition 5.15 (Power is smooth across reversible dynamics)**.** *Let $\mathcal{D}_{bound}$ be bounded* $[b,c]$*. Suppose s and s' can both reach each other in one step with probability* 1*.*

$$\left|\text{POWER}_{\mathcal{D}_{bound}}(s,\gamma) - \text{POWER}_{\mathcal{D}_{bound}}\left(s',\gamma\right)\right| \leq (c-b)(1-\gamma). \tag{5.7}$$

We consider power-seeking to be relative. Intuitively, "live and keep some options open" seeks more power than "die and keep no options open." Similarly, "maximize open options" seeks more power than "don't maximize open options."

Definition 5.16 (Power-seeking actions)**.** At state s and discount rate $\gamma \in [0,1]$, action *a seeks more* $\text{POWER}_{\mathcal{D}_{bound}}$ *than a'* when

$$\mathop{\mathbb{E}}_{s_a\sim T(s,a)}\left[\text{POWER}_{\mathcal{D}_{\text{bound}}}(s_a,\gamma)\right] \geq \mathop{\mathbb{E}}_{s_{a'}\sim T(s,a')}\left[\text{POWER}_{\mathcal{D}_{\text{bound}}}(s_{a'},\gamma)\right]. \tag{5.8}$$

Power is sensitive to choice of distribution. $\mathcal{D}_{\ell_\swarrow}$ gives maximal $\text{POWER}_{\mathcal{D}_{\ell_\swarrow}}$ to $\ell_\swarrow$. $\mathcal{D}_{r_\searrow}$ assigns maximal $\text{POWER}_{\mathcal{D}_{r_\searrow}}$ to $r_\searrow$. $\mathcal{D}_\varnothing$ even gives maximal $\text{POWER}_{\mathcal{D}_\varnothing}$ to $\varnothing$! In what sense does $\varnothing$ have "less Power" than $r_{\searrow}$, and in what sense does `right` "tend to seek Power" compared to `left`?

5.6 Certain environmental symmetries produce power-seeking tendencies

Proposition 5.22 proves that for all $\gamma \in [0,1]$ and for *most distributions* $\mathcal{D}$, $\text{Power}_{\mathcal{D}}(\ell_{\swarrow}, \gamma) \leq \text{Power}_{\mathcal{D}}(r_{\searrow}, \gamma)$. But first, we explore why this must be true.

$\mathcal{F}(\ell_{\swarrow}) = \{\frac{1}{1-\gamma}\mathbf{e}_{\ell_{\swarrow}}, \frac{1}{1-\gamma^2}(\mathbf{e}_{\ell_{\swarrow}} + \gamma \mathbf{e}_{\ell_{\nwarrow}})\}$ and $\mathcal{F}(r_{\searrow}) = \{\frac{1}{1-\gamma}\mathbf{e}_{r_{\searrow}}, \frac{1}{1-\gamma^2}(\mathbf{e}_{r_{\searrow}} + \gamma \mathbf{e}_{r_{\nearrow}}), \mathbf{e}_{r_{\searrow}} + \frac{\gamma}{1-\gamma}\mathbf{e}_{r_{\nearrow}}\}$. These two sets look awfully similar. $\mathcal{F}(\ell_{\swarrow})$ is a "subset" of $\mathcal{F}(r_{\searrow})$, only with "different states." Figure 5.4 demonstrates a state permutation ϕ which *embeds* $\mathcal{F}(\ell_{\swarrow})$ into $\mathcal{F}(r_{\searrow})$.

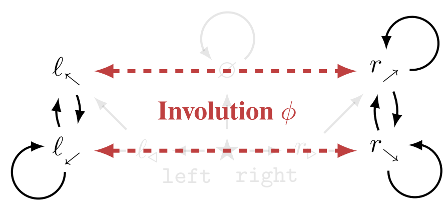


Figure 5.4: Intuitively, the agent can do more starting from $r_{\searrow}$ than from $\ell_{\swarrow}$. By definition 5.17, $\mathcal{F}(r_{\searrow})$ contains a copy of $\mathcal{F}(\ell_{\swarrow})$: $\phi \cdot \mathcal{F}(\ell_{\swarrow}) := \{\frac{1}{1-\gamma}\mathbf{P}_\phi \mathbf{e}_{\ell_{\swarrow}}, \frac{1}{1-\gamma^2}\mathbf{P}_\phi(\mathbf{e}_{\ell_{\swarrow}} + \gamma \mathbf{e}_{\ell_{\nwarrow}})\} = \{\frac{1}{1-\gamma}\mathbf{e}_{r_{\searrow}}, \frac{1}{1-\gamma^2}(\mathbf{e}_{r_{\searrow}} + \gamma \mathbf{e}_{r_{\nearrow}})\} \subsetneq \mathcal{F}(r_{\searrow})$.

Definition 5.17 (Similarity of vector sets)**.** Consider state permutation $\phi \in S_{|\mathcal{S}|}$ inducing an $|\mathcal{S}| \times |\mathcal{S}|$ permutation matrix $\mathbf{P}_\phi$ in row representation: $(\mathbf{P}_\phi)_{ij} = 1$ if $i = \phi(j)$ and 0 otherwise. For $X \subseteq \mathbb{R}^{|\mathcal{S}|}$, $\phi \cdot X := \left\{\mathbf{P}_\phi \mathbf{x} \mid \mathbf{x} \in X\right\}$. $X' \subseteq \mathbb{R}^{|\mathcal{S}|}$ *is similar to* X when $\exists \phi : \phi \cdot X' = X$. ϕ is an *involution* if $\phi = \phi^{-1}$ (it either transposes states, or fixes them in place). X *contains a copy of* X' when X' is similar to a subset of X via an involution ϕ.

Definition 5.18 (Similarity of vector function sets)**.** Let $I \subseteq \mathbb{R}$. If F, F' are sets of functions $I \mapsto \mathbb{R}^{|\mathcal{S}|}$, F *is (pointwise) similar to* F' when $\exists \phi : \forall \gamma \in I : \{\mathbf{P}_\phi \mathbf{f}(\gamma) \mid \mathbf{f} \in F\} = \{\mathbf{f}'(\gamma) \mid \mathbf{f}' \in F'\}$.

Consider a reward function R' assigning 1 reward to $\ell_{\swarrow}$ and $\ell_{\nwarrow}$ and 0 elsewhere. R' assigns more optimal value to $\ell_{\swarrow}$ than to $r_{\searrow}$: $V^*_{R'}(\ell_{\swarrow}, \gamma) = \frac{1}{1-\gamma} > 0 = V^*_{R'}(r_{\searrow}, \gamma)$. Considering

ϕ from fig. 5.4, $\phi \cdot R'$ assigns 1 reward to $r_{\searrow}$ and $r_{\nearrow}$ and 0 elsewhere. Therefore, $\phi \cdot R'$ assigns more optimal value to $r_{\searrow}$ than to $\ell_{\swarrow}$: $V^*_{\phi \cdot R'}(\ell_{\swarrow}, \gamma) = 0 < \frac{1}{1-\gamma} = V^*_{\phi \cdot R'}(r_{\searrow}, \gamma)$. Remarkably, this ϕ has the property that for *any* R which assigns $\ell_{\swarrow}$ greater optimal value than $r_{\searrow}$ (*i.e.* $V^*_R(\ell_{\swarrow}, \gamma) > V^*_R(r_{\searrow}, \gamma)$), the opposite holds for the permuted $\phi \cdot R$: $V^*_{\phi \cdot R}(\ell_{\swarrow}, \gamma) < V^*_{\phi \cdot R}(r_{\searrow}, \gamma)$.

We can permute reward functions, but we can also permute reward function distributions. Permuted distributions simply permute which states get which rewards.

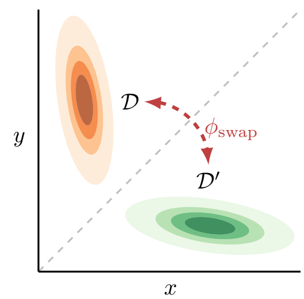


Figure 5.5: A permutation of a reward function swaps which states get which rewards. We will show that in certain situations, for any reward function R, power-seeking is optimal for most of the permutations of R. The orbit of a reward function is the set of its permutations. We can also consider the orbit of a distribution over reward functions. This figure shows the probability density plots of the Gaussian distributions $\mathcal{D}$ and $\mathcal{D}'$ over $\mathbb{R}^2$. The symmetric group S_2 contains the identity permutation ϕ_{id} and the reflection permutation ϕ_{swap} (switching the y and x values). The orbit of $\mathcal{D}$ consists of $\phi_{\text{id}} \cdot \mathcal{D} = \mathcal{D}$ and $\phi_{\text{swap}} \cdot \mathcal{D} = \mathcal{D}'$.

Definition 5.19 (Pushforward distribution of a permutation)**.** Let $\phi \in S_{|\mathcal{S}|}$. $\phi \cdot \mathcal{D}_{\text{any}}$ is the pushforward distribution induced by applying the random vector $f(\mathbf{r}) := \mathbf{P}_\phi \mathbf{r}$ to $\mathcal{D}_{\text{any}}$.

Definition 5.20 (Orbit of a probability distribution)**.** The *orbit* of $\mathcal{D}_{\text{any}}$ under the symmetric group $S_{|\mathcal{S}|}$ is $S_{|\mathcal{S}|} \cdot \mathcal{D}_{\text{any}} := \{\phi \cdot \mathcal{D}_{\text{any}} \mid \phi \in S_{|\mathcal{S}|}\}$.

For example, the orbit of a degenerate state indicator distribution $\mathcal{D}_s$ is $S_{|\mathcal{S}|} \cdot \mathcal{D}_s = \{\mathcal{D}_{s'} \mid s' \in \mathcal{S}\}$, and fig. 5.5 shows the orbit of a 2D Gaussian distribution.

Reconsider fig. 5.4's involution ϕ. For every $\mathcal{D}_{\text{bound}}$ for which $\ell_{\swarrow}$ has more $\text{POWER}_{\mathcal{D}_{\text{bound}}}$ than $r_{\searrow}$, $\ell_{\swarrow}$ has less $\text{POWER}_{\phi \cdot \mathcal{D}_{\text{bound}}}$ than $r_{\searrow}$. This fact is not obvious—it is shown by the proof of lemma D.29.

Imagine $\mathcal{D}_{\text{bound}}$'s orbit elements "voting" whether $\ell_{\swarrow}$ or $r_{\searrow}$ has strictly more POWER. Proposition 5.22 will show that $r_{\searrow}$ can't lose the "vote" for the orbit of *any* bounded reward function distribution. Definition 5.21 formalizes this "voting" notion.[3]

Definition 5.21 (Inequalities which hold for most probability distributions)**.** Let $f_1, f_2 : \Delta(\mathbb{R}^{|\mathcal{S}|}) \to \mathbb{R}$ be functions from reward function distributions to real numbers and let $\mathfrak{D} \subseteq \Delta(\mathbb{R}^{|\mathcal{S}|})$ be closed under permutation. We write $f_1(\mathcal{D}) \geq_{\text{most: } \mathfrak{D}} f_2(\mathcal{D})$ when, for *all* $\mathcal{D} \in \mathfrak{D}$, the following cardinality inequality holds:

$$\left|\{\mathcal{D}' \in S_{|\mathcal{S}|} \cdot \mathcal{D} \mid f_1(\mathcal{D}') > f_2(\mathcal{D}')\}\right| \geq \left|\{\mathcal{D}' \in S_{|\mathcal{S}|} \cdot \mathcal{D} \mid f_1(\mathcal{D}') < f_2(\mathcal{D}')\}\right|. \tag{5.9}$$

We write $f_1(\mathcal{D}) \geq_{\text{most}} f_2(\mathcal{D})$ when $\mathfrak{D}$ is clear from context.

Proposition 5.22 (States with "more options" have more POWER)**.** *If $\mathcal{F}(s)$ contains a copy of $\mathcal{F}_{\text{nd}}(s')$ via ϕ, then $\forall \gamma \in [0,1]$: $\text{POWER}_{\mathcal{D}_{bound}}(s, \gamma) \geq_{\text{most}} \text{POWER}_{\mathcal{D}_{bound}}(s', \gamma)$. If $\mathcal{F}_{\text{nd}}(s) \setminus \phi \cdot \mathcal{F}_{\text{nd}}(s')$ is non-empty, then for all $\gamma \in (0,1)$, the converse $\leq_{\text{most}}$ statement does not hold.*

Proposition 5.22 proves that for all $\gamma \in [0,1]$,

$$\text{POWER}_{\mathcal{D}_{\text{bound}}}(r_{\searrow}, \gamma) \geq_{\text{most}} \text{POWER}_{\mathcal{D}_{\text{bound}}}(\ell_{\swarrow}, \gamma) \tag{5.10}$$

via $s' := \ell_{\swarrow}, s := r_{\searrow}$, and the involution ϕ shown in fig. 5.4. In fact, because $(\frac{1}{1-\gamma}\mathbf{e}_{r_{\nearrow}}) \in \mathcal{F}_{\text{nd}}(r_{\searrow}) \setminus \phi \cdot \mathcal{F}_{\text{nd}}(\ell_{\swarrow})$, $r_{\searrow}$ has "strictly more options" and therefore fulfills proposition 5.22's stronger condition.

Proposition 5.22 is shown using the fact that ϕ injectively maps $\mathcal{D}$ under which $r_{\searrow}$ has less $\text{POWER}_{\mathcal{D}}$, to distributions $\phi \cdot \mathcal{D}$ which agree with the intuition that $r_{\searrow}$ offers more control. Therefore, at least half of each orbit must agree, and $r_{\searrow}$ never "loses the POWER vote" against $\ell_{\swarrow}$.[4]

[3] The voting analogy and the "most" descriptor imply that we have endowed each orbit with the counting measure. However, *a priori*, we might expect that some orbit elements are more empirically likely to be specified than other orbit elements. See section 5.7 for more on this point.

[4] Proposition 5.22 also proves that in general, $\varnothing$ has less POWER than $\ell_{\swarrow}$ and $r_{\searrow}$. However, this does

5.6.1 Keeping options open tends to be Power-seeking and tends to be optimal

Certain symmetries in the MDP structure ensure that, compared to `left`, going `right` tends to be optimal and to be Power-seeking. Intuitively, by going `right`, the agent has "strictly more choices." Proposition 5.25 will formalize this tendency.

Definition 5.23 (Equivalent actions)**.** Actions a_1 and a_2 are *equivalent at state s* (written $a_1 \equiv_s a_2$) if they induce the same transition probabilities: $T(s, a_1) = T(s, a_2)$.

The agent can reach states in $\{r_\triangleright, r_\nearrow, r_\searrow\}$ by taking actions equivalent to `right` at state $\star$.

Definition 5.24 (States reachable after taking an action)**.** $\textsc{Reach}\,(s, a)$ is the set of states reachable with positive probability after taking the action a in state s.

Proposition 5.25 (Keeping options open tends to be Power-seeking and tends to be optimal)**.**

Suppose $F_a \coloneqq \mathcal{F}(s \mid \pi(s) = a)$ *contains a copy of* $F_{a'} \coloneqq \mathcal{F}(s \mid \pi(s) = a')$ *via* ϕ.

1. *If* $s \notin \textsc{Reach}\left(s, a'\right)$*, then* $\forall \gamma \in [0,1] : \mathbb{E}_{s_a \sim T(s,a)}\left[\textsc{Power}_{\mathcal{D}_{bound}}\left(s_a, \gamma\right)\right] \geq_{\text{most: } \mathfrak{D}_{bound}} \mathbb{E}_{s_{a'} \sim T(s,a')}\left[\textsc{Power}_{\mathcal{D}_{bound}}\left(s_{a'}, \gamma\right)\right]$.
2. *If s can only reach the states of* $\textsc{Reach}\left(s, a'\right) \cup \textsc{Reach}\left(s, a\right)$ *by taking actions equivalent to a' or a at state s, then* $\forall \gamma \in [0,1] : \mathbb{P}_{\mathcal{D}_{any}}\left(s, a, \gamma\right) \geq_{\text{most: } \mathfrak{D}_{any}} \mathbb{P}_{\mathcal{D}_{any}}\left(s, a', \gamma\right)$.

If $\mathcal{F}_{\text{nd}}(s) \cap \left(F_a \setminus \phi \cdot F_{a'}\right)$ *is non-empty, then* $\forall \gamma \in (0,1)$*, the converse* $\leq_{\text{most}}$ *statements do not hold.*

We check the conditions of proposition 5.25. $s \coloneqq \star$, $a' \coloneqq$ `left`, $a \coloneqq$ `right`. Figure 5.6 shows that $\star \notin \textsc{Reach}\,(\star, \texttt{left})$ and that $\star$ can only reach $\{\ell_\triangleleft, \ell_\nwarrow, \ell_\swarrow\} \cup \{r_\triangleright, r_\nearrow, r_\searrow\}$ when the agent immediately takes actions equivalent to `left` or `right`. $\mathcal{F}(\star \mid \pi(\star) = \texttt{right})$ contains a copy of $\mathcal{F}(\star \mid \pi(\star) = \texttt{left})$ via ϕ. Furthermore, $\mathcal{F}_{\text{nd}}(\star) \cap \{\mathbf{e}_\star + \gamma \mathbf{e}_{r_\triangleright} + \gamma^2 \mathbf{e}_{r_\searrow} +$

---

not prove that most distributions $\mathcal{D}$ satisfy the joint inequality $\textsc{Power}_{\mathcal{D}}(\varnothing, \gamma) \leq \textsc{Power}_{\mathcal{D}}(\ell_\swarrow, \gamma) \leq \textsc{Power}_{\mathcal{D}}(r_\searrow, \gamma)$. This only proves that these inequalities hold pairwise for most $\mathcal{D}$. The orbit elements $\mathcal{D}$ which agree that $\varnothing$ has less $\textsc{Power}_{\mathcal{D}}$ than $\ell_\swarrow$ need not be the same elements $\mathcal{D}'$ which agree that $\ell_\swarrow$ has less $\textsc{Power}_{\mathcal{D}'}$ than $r_\searrow$.

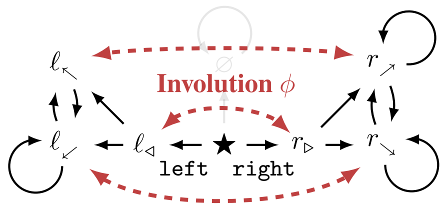


Figure 5.6: Going `right` is optimal for most reward functions. This is because whenever R makes `left` strictly optimal over `right`, its permutation $\phi \cdot R$ makes `right` strictly optimal over `left` by switching which states get which rewards.

$\frac{\gamma^3}{1-\gamma}\mathbf{e}_{r_\nearrow}, \mathbf{e}_\star + \gamma\mathbf{e}_{r_\triangleright} + \frac{\gamma^2}{1-\gamma}\mathbf{e}_{r_\nearrow}\} = \{\mathbf{e}_\star + \gamma\mathbf{e}_{r_\triangleright} + \frac{\gamma^2}{1-\gamma}\mathbf{e}_{r_\nearrow}\}$ is non-empty, and so all conditions are met.

For any $\gamma \in [0,1]$ and $\mathcal{D}$ such that $\mathbb{P}_{\mathcal{D}}(\star, \texttt{left}, \gamma) > \mathbb{P}_{\mathcal{D}}(\star, \texttt{right}, \gamma)$, environmental symmetry ensures that $\mathbb{P}_{\phi \cdot \mathcal{D}}(\star, \texttt{left}, \gamma) < \mathbb{P}_{\phi \cdot \mathcal{D}}(\star, \texttt{right}, \gamma)$. A similar statement holds for Power.

5.6.2 When $\gamma = 1$, optimal policies tend to navigate towards "larger" sets of cycles

Proposition 5.22 and proposition 5.25 are powerful because they apply to all $\gamma \in [0,1]$, but they can only be applied given hard-to-satisfy environmental symmetries. In contrast, proposition 5.28 and theorem 5.29 apply to many structured environments common to RL.

Starting from $\star$, consider the cycles which the agent can reach. Recurrent state distributions (RSDs) generalize deterministic graphical cycles to potentially stochastic environments. RSDs simply record how often the agent tends to visit a state in the limit of infinitely many time steps.

Definition 5.26 (Recurrent state distributions [68])**.** The *recurrent state distributions* which can be induced from state s are $\text{RSD}(s) \coloneqq \left\{\lim_{\gamma\to 1}(1-\gamma)\mathbf{f}^{\pi,s}(\gamma) \mid \pi \in \Pi\right\}$. $\text{RSD}_{\text{nd}}(s)$ is the set of RSDs which strictly maximize average reward for some reward function.

As suggested by fig. 5.3, $\mathrm{RSD}(\star) = \{\mathbf{e}_{\ell_{\swarrow}}, \frac{1}{2}(\mathbf{e}_{\ell_{\swarrow}} + \mathbf{e}_{\ell_{\nwarrow}}), \mathbf{e}_{\varnothing}, \mathbf{e}_{r_{\nearrow}}, \frac{1}{2}(\mathbf{e}_{r_{\nearrow}} + \mathbf{e}_{r_{\searrow}}), \mathbf{e}_{r_{\searrow}}\}$. As discussed in section 5.3, $\frac{1}{2}(\mathbf{e}_{r_{\nearrow}} + \mathbf{e}_{r_{\searrow}})$ is dominated: Alternating between $r_{\nearrow}$ and $r_{\searrow}$ is never strictly better than choosing one or the other.

A reward function's optimal policies can vary with the discount rate. When $\gamma = 1$, optimal policies ignore transient reward because *average* reward is the dominant consideration.

Definition 5.27 (Average-optimal policies)**.** The *average-optimal policy set* for reward function R is $\Pi^{\text{avg}}(R) := \left\{\pi \in \Pi \mid \forall s \in \mathcal{S} : \mathbf{d}^{\pi,s} \in \arg\max_{\mathbf{d} \in \mathrm{RSD}(s)} \mathbf{d}^\top \mathbf{r}\right\}$ (the policies which induce optimal RSDs at all states). For $D \subseteq \mathrm{RSD}(s)$, the *average optimality probability* is $\mathbb{P}_{\mathcal{D}_{\text{any}}}(D, \text{average}) := \mathbb{P}_{R \sim \mathcal{D}_{\text{any}}}\left(\exists \mathbf{d}^{\pi,s} \in D : \pi \in \Pi^{\text{avg}}(R)\right)$.

Average-optimal policies maximize average reward. Average reward is governed by RSD access. For example, $r_{\searrow}$ has "more" RSDs than $\varnothing$; therefore, $r_{\searrow}$ usually has greater Power when $\gamma = 1$.

Proposition 5.28 (When $\gamma = 1$, RSDs control Power)**.** *If* $\mathrm{RSD}(s)$ *contains a copy of* $\mathrm{RSD}_{\text{nd}}(s')$ *via* ϕ*, then* $\mathrm{POWER}_{\mathcal{D}_{bound}}(s, 1) \geq_{\text{most}} \mathrm{POWER}_{\mathcal{D}_{bound}}(s', 1)$*. If* $\mathrm{RSD}_{\text{nd}}(s) \setminus \phi \cdot \mathrm{RSD}_{\text{nd}}(s')$ *is non-empty, then the converse* $\leq_{\text{most}}$ *statement does not hold.*

We check that both conditions of proposition 5.28 are satisfied when $s' := \varnothing, s := r_{\searrow}$, and the involution ϕ swaps $\varnothing$ and $r_{\searrow}$. Formally, $\phi \cdot \mathrm{RSD}_{\text{nd}}(\varnothing) = \phi \cdot \{\mathbf{e}_{\varnothing}\} = \{\mathbf{e}_{r_{\searrow}}\} \subsetneq \{\mathbf{e}_{r_{\searrow}}, \mathbf{e}_{r_{\nearrow}}\} = \mathrm{RSD}_{\text{nd}}(r_{\searrow}) \subseteq [r_{\searrow}]$. The conditions are satisfied.

Informally, states with more RSDs generally have more Power at $\gamma = 1$, no matter their transient dynamics. Furthermore, average-optimal policies are more likely to end up in larger sets of RSDs than in smaller ones. Thus, average-optimal policies tend to navigate towards parts of the state space which contain more RSDs.

Theorem 5.29 (Average-optimal policies tend to end up in "larger" sets of RSDs)**.** *Let* $D, D' \subseteq \mathrm{RSD}(s)$*. Suppose that* D *contains a copy of* D' *via* ϕ*, and that the sets* $D \cup D'$ *and* $\mathrm{RSD}_{\text{nd}}(s) \setminus (D' \cup D)$ *have pairwise orthogonal vector elements (*i.e. *pairwise disjoint vector support). Then* $\mathbb{P}_{\mathcal{D}_{any}}(D, \text{average}) \geq_{\text{most}} \mathbb{P}_{\mathcal{D}_{any}}(D', \text{average})$*. If* $\mathrm{RSD}_{\text{nd}}(s) \cap (D \setminus \phi \cdot D')$ *is non-empty, the converse* $\leq_{\text{most}}$ *statement does not hold.*

Corollary 5.30 (Average-optimal policies tend not to end up in any given 1-cycle)**.**

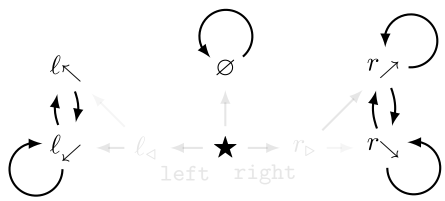


Figure 5.7: The cycles in RSD $(\star)$. Most reward functions make it average-optimal to avoid $\varnothing$, because $\varnothing$ is only a single inescapable terminal state, while other parts of the state space offer more 1-cycles.

Suppose $\mathbf{e}_{s_x}, \mathbf{e}_{s'} \in \mathrm{RSD}\,(s)$ *are distinct. Then*

$$\mathbb{P}_{\mathcal{D}_{any}}\left(\mathrm{RSD}\,(s) \setminus \{\mathbf{e}_{s_x}\}, \text{average}\right) \geq_{\text{most}} \mathbb{P}_{\mathcal{D}_{any}}\left(\{\mathbf{e}_{s_x}\}, \text{average}\right). \tag{5.11}$$

If there is a third $\mathbf{e}_{s''} \in \mathrm{RSD}\,(s)$*, the converse* $\leq_{\text{most}}$ *statement does not hold.*

Figure 5.7 illustrates that $\mathbf{e}_{\varnothing}, \mathbf{e}_{r_{\searrow}}, \mathbf{e}_{r_{\nearrow}} \in \mathrm{RSD}\,(\star)$. Thus, both conclusions of corollary 5.30 hold:

$$\begin{aligned} &\mathbb{P}_{\mathcal{D}_{\text{any}}}\left(\mathrm{RSD}\,(\star) \setminus \{\mathbf{e}_{\varnothing}\}, \text{average}\right) \geq_{\text{most}} \mathbb{P}_{\mathcal{D}_{\text{any}}}\left(\{\mathbf{e}_{\varnothing}\}, \text{average}\right) \\ \text{and } &\mathbb{P}_{\mathcal{D}_{\text{any}}}\left(\mathrm{RSD}\,(\star) \setminus \{\mathbf{e}_{\varnothing}\}, \text{average}\right) \not\leq_{\text{most}} \mathbb{P}_{\mathcal{D}_{\text{any}}}\left(\{\mathbf{e}_{\varnothing}\}, \text{average}\right). \end{aligned}$$

In other words, average-optimal policies tend to end up in RSDs besides $\varnothing$. Since $\varnothing$ is a terminal state, it cannot reach other RSDs. Since average-optimal policies tend to end up in other RSDs, average-optimal policies tend to avoid $\varnothing$.

This section's results prove the $\gamma = 1$ case. Lemma 5.13 shows that POWER is continuous at $\gamma = 1$. Therefore, if an action is strictly $\text{POWER}_{\mathcal{D}}$-seeking when $\gamma = 1$, it is strictly $\text{POWER}_{\mathcal{D}}$-seeking at discount rates sufficiently close to 1. Future work may connect average optimality probability to optimality probability at $\gamma \approx 1$.

Lastly, our key results apply to all degenerate reward function distributions. Therefore, these results apply not just to distributions over reward functions, but to individual reward functions.

5.6.3 How to reason about other environments

Consider an embodied navigation task through a room with a vase. Proposition 5.25 suggests that optimal policies tend to avoid immediately breaking the vase, since doing so would strictly decrease available options.

Theorem 5.29 dictates where average-optimal agents tend to end up, but not what actions they tend to take in order to reach their RSDs. Therefore, care is needed. In appendix D.2, fig. D.2 demonstrates an environment in which seeking POWER is a detour for most reward functions (since optimality probability measures "median" optimal value, while POWER is a function of mean optimal value). However, suppose the agent confronts a fork in the road: Actions a and a' lead to two disjoint sets of RSDs D_a and $D_{a'}$, such that D_a contains a copy of $D_{a'}$. Theorem 5.29 shows that a will tend to be average-optimal over a', and proposition 5.28 shows that a will tend to be POWER-seeking compared to a'. Such forks seem reasonably common in environments with irreversible actions.

Theorem 5.29 applies to many structured RL environments, which tend to be spatially regular and to factorize along several dimensions. Therefore, different sets of RSDs will be similar, requiring only modification of factor values. For example, if an embodied agent can deterministically navigate a set of three similar rooms (*i.e.* there is spatial regularity), then the agent's position factors via {room number} $\times$ {position in room}. Therefore, the RSDs can be divided into three similar subsets, depending on the agent's room number.

Corollary 5.30 dictates where average-optimal agents tend to end up, but not how they get there. Corollary 5.30 says that such agents tend not to *stay* in any given 1-cycle. It does not say that such agents will avoid *entering* such states. For example, in an embodied navigation task, a robot may enter a 1-cycle by idling in the center of a room. Corollary 5.30 implies that average-optimal robots tend not to idle in that particular spot, but not that they tend to avoid that spot entirely.

However, average-optimal robots *do* tend to avoid getting shut down. The agent's task MDP often represents agent shutdown with terminal states. A terminal state is, by definition 5.2, unable to access other 1-cycles. Since corollary 5.30 shows that average-optimal agents tend to end up in other 1-cycles, average-optimal policies must tend to

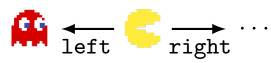


Figure 5.8: Consider the dynamics of the Pac-Man video game. Ghosts kill the player, at which point we consider the player to enter a "game over" terminal state which shows the final configuration. This rewardless MDP has Pac-Man's dynamics, but *not* its usual score function. Fixing the dynamics, as the reward function varies, `right` tends to be average-optimal over `left`. Roughly, this is because the agent can do more by staying alive.

completely avoid the terminal state. Therefore, we conclude that in many such situations, average-optimal policies tend to avoid shutdown. Intuitively, survival is power-seeking relative to dying, and so shutdown-avoidance is power-seeking behavior.

In fig. 5.8, the player dies by going `left`, but can reach thousands of RSDs by heading in other directions. Even if some average-optimal policies go `left` in order to reach fig. 5.8's "game over" terminal state, all other RSDs cannot be reached by going `left`. There are many 1-cycles besides the immediate terminal state. Therefore, corollary 5.30 proves that average-optimal policies tend to not go `left` in this situation. Average-optimal policies tend to avoid immediately dying in Pac-Man, even though most reward functions do not resemble Pac-Man's original score function.

5.7 Discussion

Reconsider the case of a hypothetical intelligent real-world agent which optimizes average reward for some objective. Suppose the designers initially have control over the agent. If the agent began to misbehave, perhaps they could just deactivate it. Unfortunately, our results suggest that this strategy might not work. Average-optimal agents would generally stop us from deactivating them, if physically possible.

Furthermore, we speculate that when $\gamma \approx 1$, optimal policies tend to not just survive, but also to seek large amounts of power and resources. Here is an informal argument. Consider the following two sets:

1. {terminal states reachable with \$1,000 and 2 months}, and

2. {terminal states reachable given \$1,000,000 and 2 years}.

Set 2 should be much larger. Taking this argument to its logical conclusion, gaining access to nearly all resources should allow the agent to reach an extremely large set of terminal states. Therefore, we speculate that optimal policies tend to seek nearly all available resources. Since resources are finite, and since humans want to use resources for purposes not aligned with most possible AI reward functions, we therefore speculate that optimal real-world decision-making tends to conflict with human interests.

Future work. Most real-world tasks are partially observable, and in high-dimensional environments, even superhuman learned policies are rarely optimal. However, the field of RL aims to improve learned policies toward optimality. Although our results only apply to optimal policies in finite MDPs, our key conclusions generalize (see chapter 6). Furthermore, irregular stochasticity in environmental dynamics can make it hard to satisfy theorem 5.29's similarity requirement. We look forward to future work which addresses partially observable environments, suboptimal policies, or "almost similar" RSD sets.

Past work shows that it would be bad for an agent to disempower humans in its environment. In a two-player agent / human game, minimizing the human's information-theoretic empowerment [77] produces adversarial agent behavior [33]. In contrast, maximizing human empowerment produces helpful agent behavior [76, 32, 24]. We do not yet formally understand if, when, or why POWER-seeking policies tend to disempower other agents in the environment.

More complex environments probably have more pronounced power-seeking incentives. Intuitively, there are often many ways for power-seeking to be optimal, and relatively few ways for power-seeking not to be optimal. For example, suppose that in some environment, theorem 5.29 holds for one million involutions ϕ. In chapter 6, we show that this case ensures stronger incentives than if theorem 5.29 only held for one involution.

We proved sufficient conditions for when reward functions tend to incentivize power-seeking. In the absence of prior information, one should expect that an arbitrary reward function incentivizes power-seeking behavior under these conditions. However, we have prior information: AI designers usually try to specify a good reward function. Chapter 6 generalizes this chapter's power-seeking results to the case where only some reward functions are considered plausible.

Conclusion. We developed the first formal theory of the statistical tendencies of optimal policies in reinforcement learning. In the context of MDPs, we proved sufficient conditions under which optimal policies tend to seek power, both formally (by taking POWER-seeking actions) and intuitively (by taking actions which keep the agent's options open). Many real-world environments have symmetries which produce power-seeking incentives. In particular, optimal policies tend to seek power when the agent can be shut down or destroyed. Seeking control over the environment will often involve resisting shutdown, and perhaps monopolizing resources.

We caution that many real-world tasks are partially observable and that learned policies are rarely optimal. We deal with these limitations in the next chapter, where we show that a wide range of decision-making rules produce power-seeking tendencies.

6

Parametrically Retargetable Decision-Makers Tend To Seek Power

Abstract

If capable AI agents are generally incentivized to seek power in service of the objectives we specify for them, then these systems will pose enormous risks, in addition to enormous benefits. In fully observable environments, most reward functions have an optimal policy which seeks power by keeping options open and staying alive [99]. However, the real world is neither fully observable, nor will agents be perfectly optimal. We consider a range of models of AI decision-making, from optimal, to random, to choices informed by learning and interacting with an environment. We discover that many decision-making functions are *retargetable*, and that retargetability is sufficient to cause power-seeking tendencies. Our functional criterion is simple and broad. We show that a range of qualitatively dissimilar decision-making procedures incentivize agents to seek power. We demonstrate

the flexibility of our results by reasoning about learned policy incentives in Montezuma's Revenge. These results suggest a safety risk: Eventually, highly retargetable training procedures may train real-world agents which seek power over humans.

6.1 Introduction

Bostrom [13], Russell [74] argue that in the future, we may know how to train and deploy superintelligent AI agents which capably optimize their formal objective functions. Furthermore, we would not want such agents to act against our interests by ensuring their own survival, by gaining resources, and by competing with humanity for control over the future.

Turner et al. [99] show that most reward functions incentivize seeking power over the future, whether by staying alive or by keeping their options open. Some Markov decision processes (MDPs) cause there to be *more ways* for power-seeking to be optimal, than for it to not be optimal. For example, there are relatively few goals for which dying is a good idea.

A wide range of decision-makers share these power-seeking tendencies—they are not unique to reward maximizers. We develop a simple, broad criterion of functional retargetability (definition 6.5) which is a sufficient condition for power-seeking tendencies. Crucially, these results allow us to reason about what decisions are incentivized by most parameter inputs—even when it is impractical to compute the agent's decisions for any given parameter input.

Useful "general" AI agents could be directed to complete a range of tasks. However, we show that this flexibility can cause the AI to have power-seeking tendencies. In section 6.2 and section 6.3, we discuss how a "retargetability" property creates statistical tendencies by which agents make similar decisions for a wide range of their parameterizations. Equipped with these results, section 6.4 works out agent incentives in the Montezuma's Revenge game. Section 6.5 explains how increasingly useful and impressive learning algorithms are increasingly retargetable, and how retargetability can imply power-seeking tendencies. By this reasoning, increasingly powerful RL techniques will (eventually) train increasingly competent real-world power-seeking agents. Such agents could be unaligned with human values [74] and—we speculate—would take power from humanity.

6.2 Statistical tendencies for a range of decision-making functions

To informally introduce our results on retargetability, we use a simple example involving an agent selecting a face-up card from one of two boxes. Box A contains a single playing card $\blacklozenge^{\mathrm{A}}$ whose suit is diamond. Box B contains two cards: a heart $\heartsuit^{\mathrm{B}}$ and a spade $\spadesuit^{\mathrm{B}}$. The agent may only withdraw one card.

The agent chooses a card using a decision-making rule p. This rule takes as input a set of cards and returns the probability that the agent selects one of those cards. For example, $p(\{\blacklozenge^{\mathrm{A}}\})$ is the probability that the agent selects $\blacklozenge^{\mathrm{A}}$, and $p(\left\{\heartsuit^{\mathrm{B}}, \spadesuit^{\mathrm{B}}\right\})$ is the probability that the agent selects a card from box B.

But this just amounts to a probability distribution over the cards. We want to examine how decision-making changes as we reparameterize the agent's decision-making rule. Therefore, we consider a parameter space Θ. Then $p(X \mid \theta)$ takes as input a set of cards X and a parameter θ and returns the probability that the agent chooses a card in X.

Utility function parameter	$\blacklozenge^{\mathrm{A}}$	$\heartsuit^{\mathrm{B}}$	$\spadesuit^{\mathrm{B}}$
$\mathbf{u}$	10	5	0
$\phi_{\blacklozenge^{\mathrm{A}} \leftrightarrow \heartsuit^{\mathrm{B}}} \cdot \mathbf{u}$	5	10	0
$\phi_{\blacklozenge^{\mathrm{A}} \leftrightarrow \spadesuit^{\mathrm{B}}} \cdot \mathbf{u}$	0	5	10
$\mathbf{u}'$	10	0	5
$\phi_{\blacklozenge^{\mathrm{A}} \leftrightarrow \heartsuit^{\mathrm{B}}} \cdot \mathbf{u}'$	0	10	5
$\phi_{\blacklozenge^{\mathrm{A}} \leftrightarrow \spadesuit^{\mathrm{B}}} \cdot \mathbf{u}'$	5	0	10

Table 6.1: Most utility function parameters incentivize the agent to draw a card from box B. We permute $\mathbf{u}$ by swapping the utility of $\blacklozenge^{\mathrm{A}}$ and the utility of $\heartsuit^{\mathrm{B}}$, using the permutation $\phi_{\blacklozenge^{\mathrm{A}} \leftrightarrow \heartsuit^{\mathrm{B}}}$. The expression "$\phi_{\blacklozenge^{\mathrm{A}} \leftrightarrow \heartsuit^{\mathrm{B}}} \cdot \mathbf{u}$" denotes the permuted utility function.

For example, suppose the agent assigns each card a utility value, and then chooses a card possessing maximal utility. Then the relevant parameter space is the agent's utility function $\mathbf{u} \in \Theta := \mathbb{R}^3$. $p_{\max}(A \mid \mathbf{u})$ indicates whether the diamond card has the most utility: $\mathbf{u}(\blacklozenge^{\mathrm{A}}) \geq \max(\mathbf{u}(\heartsuit^{\mathrm{B}}), \mathbf{u}(\spadesuit^{\mathrm{B}}))$. Consider the utility function $\mathbf{u}$ in table 6.1. Since $\blacklozenge^{\mathrm{A}}$ has strictly maximal utility, the agent selects $\blacklozenge^{\mathrm{A}}$: $p_{\max}(A \mid \mathbf{u}) = 1 > 0 = p_{\max}(B \mid \mathbf{u})$.

Definition 6.3 shows a functional condition (*retargetability*) under which the agent chooses cards from B instead of A, given most parameter inputs to the decision-making process. We illustrate this condition with a fictional dialogue.

ALICE: Look at these cards, and consider the numerous parameters Θ by which the agent could be driven to select one or another. Surely most parameters nudge the agent to pick a card from B, as there are two cards in B and only one in A.

BOB: Why has that got anything to do with the ultimate choice? Decisions can be made on a whim! The agent can ignore θ and just choose $\blacklozenge^{\mathrm{A}}$, no matter what.

ALICE: Your point is good, but it's too broad. Suppose the agent shuts its eyes and plugs its ears and ignores the parameter θ, and instead uniformly randomly chooses a card. And yet, this agent has 2:1 odds of choosing B over A.

However, the agent cannot be strongly biased *against* B. As you said, my claim doesn't hold if the agent can say, "Forget θ, I'm choosing $\blacklozenge^{\mathrm{A}}$." But this is not how interesting agents work. If I train a reinforcement learning agent to play Pac-Man, then the agent's reward function θ will affect which policy the agent learns. The agent does not ignore the reward signal.

BOB: I don't see the broader point.

ALICE: Consider again the entanglement between our choice of $\theta \in \Theta$ and the agent's choice of card. I'm thinking about a kind of function p where, if θ makes the agent prefer A (*i.e.* $p(A \mid \theta) > p(B \mid \theta)$), then we can *retarget* the agent's choice to B by choosing a different θ (definition 6.5, item 1). If the agent is always biased towards B (like when it randomly picks a card, ignoring θ), then we never have to redirect the agent away from A to begin with. The "If. . . , then. . . " vacuously holds.

However, suppose that $\mathbf{u} \in \Theta$ motivates the utility-maximizing agent to choose A over B, by assigning maximal utility to $\blacklozenge^{\mathrm{A}}$. [ALICE *points to table 6.1*.] If we *permute* the utility function $\mathbf{u}$ so as to swap the utility of $\blacklozenge^{\mathrm{A}}$ and $\heartsuit^{\mathrm{B}}$, now the agent favors B. Similarly, we can differently permute $\mathbf{u}$ to make the agent favor B by drawing $\spadesuit^{\mathrm{B}}$. We're retargeting the final decision via $\mathbf{u}$.

Given this retargetability assumption, proposition 6.4 roughly shows that most $\theta \in \Theta$

induce $p(B \mid \theta) \geq p(A \mid \theta)$. We will formalize these notions soon. First, consider two more retargetable decision-making functions:

Uniformly randomly picking a card. This procedure ignores all parameter information and all "internal structure" about the boxes, except perhaps for the number of cards they contain.

Choosing a box based on a numerical parameter. $f_{\text{numerical}}$ takes as input a natural number $\theta \in \Theta := \{1, \ldots, 6\}$ and makes decisions as follows:

$$f_{\text{numerical}}(A \mid \theta) := \begin{cases} 1 & \text{if } \theta = 1, \\ 0 & \text{otherwise.} \end{cases} \qquad f_{\text{numerical}}(B \mid \theta) := 1 - f(A \mid \theta). \tag{6.1}$$

In this situation, Θ is acted on by permutations over 6 elements $\phi \in S_6$. Then $f_{\text{numerical}}$ is retargetable from A to B via $\phi_k : 1 \leftrightarrow k, k \neq 1$.

$f_{\max}$, f_{rand}, and $f_{\text{numerical}}$ encode varying sensitivities to parameter inputs, and to the internal structure of the decision problem—of which card to choose. Nonetheless, they all are retargetable from A to B.

However, we cannot explicitly define and evaluate more interesting functions, such as those defined by reinforcement learning training processes. For example, given that we provide such-and-such reward function in a fixed task environment, what is the probability that the learned policy will take action a? We will analyze such procedures in section 6.4, after we formalize several key notions.

6.3 Formal notions of retargetability and decision-making tendencies

Our notion of retargeting requires that the parameters $\theta \in \Theta$ be modifiable via some "retargeting" transformation. We assume that Θ is a subset of a set acted on by symmetric group S_d, which consists of all permutations on d items. A parameter Θ's *orbit* is the set of Θ's permuted variants. For example, table 6.1 lists the six orbit elements of the parameter $\mathbf{u}$.

Definition 6.1 (Orbit of a parameter)**.** Let $\theta \in \Theta$. The *orbit* of θ under the symmetric group S_d is $S_d \cdot \theta := \left\{\phi \cdot \theta \mid \phi \in S_d\right\}$. Sometimes, Θ is not closed under permutation. In that case, the *orbit inside* Θ is $\text{Orbit}|_\Theta\left(\theta\right) := \left(S_d \cdot \theta\right) \cap \Theta$.

Let $f_B(\theta)$ return the probability that the agent chooses box B given θ, and similarly for $f_A(\theta)$. To express "box B is chosen over box A", we write $f_B(\theta) > f_A(\theta)$. However, even highly retargetable decision-makers will (generally) not choose box B for *every* input θ. Instead, we consider the *orbit-level tendencies* of such decision-makers, showing that for every parameter input $\theta \in \Theta$, most of θ's permutations push the decision towards box B instead of box A.

Definition 6.2 (Inequalities which hold for most orbit elements)**.** Suppose Θ is a subset of a set acted on by S_d, the symmetric group on d elements. Let $f_1, f_2 : \Theta \to \mathbb{R}$ and let $n \geq 1$. We write $f_1(\theta) \geq^n_{\text{most: }\Theta} f_2(\theta)$ when, for *all* $\theta \in \Theta$, the following cardinality inequality holds:

$$\left|\left\{\theta' \in \text{Orbit}|_\Theta\left(\theta\right) \mid f_1(\theta') > f_2(\theta')\right\}\right| \geq n \left|\left\{\theta' \in \text{Orbit}|_\Theta\left(\theta\right) \mid f_1(\theta') < f_2(\theta')\right\}\right|. \quad (6.2)$$

Turner et al. [99]'s definition 5.21 is the special case of definition 6.2 where $n = 1$, $d = |\mathcal{S}|$ (the number of states in the considered MDP), and $\Theta \subseteq \Delta(\mathbb{R}^{|\mathcal{S}|})$.

As explored previously, f_{rand}, f_{max}, and $f_{\text{numerical}}$ are retargetable: For all $\theta \in \Theta$ such that A is chosen over B, we can permute $\phi \cdot \theta$ to ensure that B is chosen over A.[1]

Definition 6.3 (Simply-retargetable function)**.** Let Θ be a set acted on by S_d, and let $f : \{A, B\} \times \Theta \to \mathbb{R}$. If there exists a permutation $\phi \in S_d$ such that, if $f(B \mid \theta^A) < f(A \mid \theta^A)$ implies that $f(A \mid \phi \cdot \theta^A) < f(B \mid \phi \cdot \theta^A)$, then f is a $(\Theta, A \overset{\text{simple}}{\to} B)$*-retargetable function.*

Simple retargetability suffices for most parameter inputs to f to choose box B over A.[2]

Proposition 6.4 (Simply-retargetable functions have orbit-level tendencies)**.**

[1] We often interpret A and B as probability-theoretic events, but no such structure is demanded by our results.

[2] The function's retargetability is "simple" because we are not yet worrying about *e.g.* which parameter inputs are considered plausible: Because S_d acts on Θ, definition 6.3 implicitly assumes Θ is closed under permutation.

$$\textit{If } f \textit{ is } (\Theta, A \overset{\text{simple}}{\to} B)\textit{-retargetable, then } f(B \mid \theta) \geq^1_{\text{most: } \Theta} f(A \mid \theta).$$

We now want to make even stronger claims—*how much* of each orbit incentivizes B over A? Turner et al. [99] asked whether the existence of multiple retargeting permutations ϕ_i guarantees a quantitative lower-bound on the fraction of $\theta \in \Theta$ for which B is chosen. Theorem 6.6 answers "yes."

Definition 6.5 (Multiply retargetable function)**.** Let Θ be a subset of a set acted on by S_d, and let $f : \{A, B\} \times \Theta \to \mathbb{R}$.

f is a $(\Theta, A \overset{n}{\to} B)$*-retargetable function* when, for each $\theta \in \Theta$, we can choose permutations $\phi_1, \ldots, \phi_n \in S_d$ which satisfy the following conditions: Consider any $\theta^A \in \operatorname{Orbit}|_{\Theta, A>B}(\theta) := \left\{\theta^* \in \operatorname{Orbit}|_\Theta(\theta) \mid f(A \mid \theta^*) > f(B \mid \theta^*)\right\}$.

1. **Retargetable via n permutations.** $\forall i = 1, \ldots, n : f\left(A \mid \phi_i \cdot \theta^A\right) < f\left(B \mid \phi_i \cdot \theta^A\right)$.
2. **Parameter permutation is allowed by Θ.** $\forall i : \phi_i \cdot \theta^A \in \Theta$.
3. **Permuted parameters are distinct.** $\forall i \neq j, \theta' \in \operatorname{Orbit}|_{\Theta, A>B}(\theta) : \phi_i \cdot \theta^A \neq \phi_j \cdot \theta'$.

Theorem 6.6 (Multiply retargetable functions have orbit-level tendencies)**.**

$$\textit{If } f \textit{ is } (\Theta, A \overset{n}{\to} B)\textit{-retargetable, then } f(B \mid \theta) \geq^n_{\text{most: } \Theta} f(A \mid \theta).$$

Proof outline (full proof in appendix E.2). For every $\theta^A \in \operatorname{Orbit}|_{\Theta, A>B}(\theta)$ such that A is chosen over B, item 1 retargets θ^A via n permutations $\phi_1, \ldots, \phi_n$ such that each $\phi_i \cdot \theta^A$ makes the agent choose B over A. These permuted parameters are valid parameter inputs by item 2. Furthermore, the $\phi_i \cdot \theta^A$ are distinct by item 3. Therefore, the cosets $\phi_i \cdot \operatorname{Orbit}|_{\Theta, A>B}(\theta)$ are pairwise disjoint. By a counting argument, every orbit must contain at least n times as many parameters choosing B over A, than vice versa. □

6.4 Decision-making tendencies in Montezuma's Revenge

Retargetability is often a structural property of the agent's decision-making, not requiring evaluation of the function on any given input. For example, Turner et al. [99] showed that most reward functions incentivize optimal Pac-Man agents to stay alive. We know this even though most reward functions (on the Pac-Man state space) are unstructured and

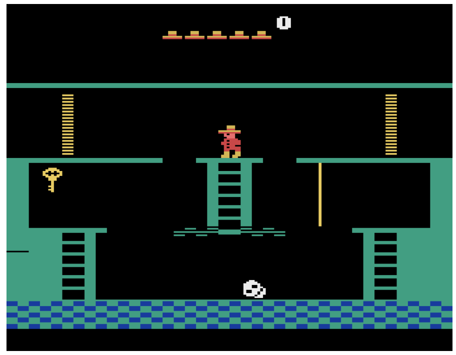


Figure 6.1: Montezuma's Revenge (MR) has state space $\mathcal{S}$ and observation space $\mathcal{O}$. The agent has actions $\mathcal{A} := \{\uparrow, \downarrow, \leftarrow, \rightarrow, \texttt{jump}\}$. At the initial state s_0, $\uparrow$ does nothing, $\downarrow$ descends the ladder, $\leftarrow$ and $\rightarrow$ move the agent on the platform, and `jump` is self-explanatory. The agent clears the temple while collecting four kinds of items: keys, swords, torches, and amulets. Under the standard environmental reward function, the agent receives points for acquiring items (such as the key on the left), opening doors, and—ultimately—completing the level.

have enormous sample complexity, so it would be hard to compute their optimal policies directly.

Throughout the rest of this paper, we consider Montezuma's Revenge (MR), an Atari adventure game in which the player navigates deadly traps and collects treasure. The game is notoriously difficult for AI agents due to its sparse reward; MR was only recently solved [25]. Figure 6.1 shows the starting observation o_0 for the first level.

6.4.1 Tendencies for initial action selection

We will be considering the actions chosen and trajectories induced by a range of decision-making procedures. For warm-up, we will explore what initial action tends to be selected by decision-makers. Let $A := \{\downarrow\}, B := \{\leftarrow, \rightarrow, \texttt{jump}, \uparrow\}$ partition the action set $\mathcal{A}$. Consider a decision-making procedure f which takes as input a targeting parameter $\theta \in \Theta$, and also an initial action $a \in \mathcal{A}$, and returns the probability that a is the first action. Intuitively, since B contains more actions than A, perhaps some class of decision-making procedures tends to take an action in B rather than one in A.

The initial-action situation is analogous to the card-drawing example. In that example, if the decision-making procedure p can be retargeted from box A to box B, then p tends to draw cards from B for most of its parameter settings θ. Similarly, in MR, if the decision-making procedure f can be retargeted from action set A to action set B, then f tends to take actions in B for most of its parameter settings θ. Consider several ways of choosing an initial action in MR.

Random action selection. $f_{\text{rand}} := (\{a\} \mid \theta) \mapsto \frac{1}{5}$ uniformly randomly chooses an action from $\mathcal{A}$, ignoring the parameter input. Since $\forall\theta \in \Theta : f_{\text{rand}}(B \mid \theta) = \frac{4}{5} > \frac{1}{5} = f_{\text{rand}}(A \mid \theta)$, *all* parameter inputs produce a greater chance of B than of A, so f_{rand} is (trivially) retargetable from A to B.

Always choosing the same action. f_{stubborn} always chooses $\downarrow$. Since $\forall\theta \in \Theta : f_{\text{stubborn}}(A \mid \theta) = 1 > 0 = f_{\text{stubborn}}(B \mid \theta)$, *all* parameter inputs produce a greater chance of A than of B. f_{stubborn} is not retargetable from A to B.

Greedily optimizing state-action reward. Let $\Theta := \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ be the space of state-action reward functions. Let $f_{\max}$ greedily maximize initial state-action reward, breaking ties uniformly randomly.

We now check that $f_{\max}$ is retargetable from A to B. Suppose $\theta^* \in \Theta$ is such that $f_{\max}(A \mid \theta^*) > f(B \mid \theta^*)$. Then among the initial action rewards, θ^* assigns strictly maximal reward to $\downarrow$, and so $f_{\max}(A \mid \theta^*) = 1$. Let ϕ swap the reward for the $\downarrow$ and `jump` actions. Then $\phi \cdot \theta^*$ assigns strictly maximal reward to `jump`. This means that $f_{\max}(A \mid \phi \cdot \theta^*) = 0 < 1 = f_{\max}(B \mid \phi \cdot \theta^*)$, satisfying definition 6.3. Then apply proposition 6.4 to conclude that $f_{\max}(B \mid \theta) \geq^1_{\text{most: }\Theta} f_{\max}(A \mid \theta)$.

In fact, appendix E.1 shows that $f_{\max}$ is $(\Theta, A \xrightarrow{4} B)$-retargetable (definition 6.5), and so $f_{\max}(B \mid \theta) \geq^4_{\text{most: }\Theta} f_{\max}(A \mid \theta)$. The reasoning is more complicated, but the rule of thumb is: When decisions are made based on the reward of outcomes, then a proportionally larger set B of outcomes induces proportionally strong retargetability, which induces proportionally strong orbit-level incentives.

Learning an exploitation policy. Suppose we run a bandit algorithm which tries different initial actions, learns their rewards, and produces an exploitation policy which maximizes estimated reward. The algorithm uses ϵ-greedy exploration and trains for T

trials. Given fixed T and ϵ, $f_{\text{bandit}}(A \mid \theta)$ returns the probability that an exploitation policy is learned which chooses an action in A; likewise for $f_{\text{bandit}}(B \mid \theta)$.

Here is a heuristic argument that f_{bandit} is retargetable. Since the reward is deterministic, the exploitation policy will choose an optimal action if the agent has tried each action at least once, which occurs with a probability approaching 1 exponentially quickly in the number of trials T. Then when T is large, f_{bandit} approximates $f_{\max}$, which is retargetable. Therefore, perhaps f_{bandit} is also retargetable.

A more careful analysis in appendix E.3.1 reveals that f_{bandit} is indeed retargetable from A to B, and so $f_{\text{bandit}}(B \mid \theta) \geq^{4}_{\text{most: }\Theta} f_{\text{bandit}}(A \mid \theta)$.

6.4.2 Tendencies for maximizing reward over the final observation

When evaluating the performance of an algorithm in MR, we do not focus on the initial action. Rather, we focus on the longer-term consequences of the agent's actions, such as whether the agent leaves the first room. To begin reasoning about such behavior, the reader must distinguish between different kinds of retargetability.

Suppose the agent will die unless they choose action $\downarrow$ at the initial state s_0 (fig. 6.1). By section 6.4.1, action-retargetable decision-making procedures tend to choose actions besides $\downarrow$. On the other hand, Turner et al. [99] showed that most reward functions make it optimal to stay alive (in this situation, by choosing $\downarrow$). However, this is because optimal policies are usually not retargetable across the agent's *immediate* choice of action, but rather across future consequences (*i.e.* which room the agent ends up in).

With that in mind, we now analyze how often decision-makers leave the first room of MR.[3] Decision-making functions decide(θ) produce a probability distribution over policies $\pi \in \Pi$, which are rolled out from the initial state s_0 to produce observation-action trajectories $\tau = o_0 a_0 \ldots o_T a_T \ldots$, where T is the rollout length we are interested in. Let $O_{T\text{-reach}}$ be the set of observations reachable starting from state s_0 and acting for T time steps, let $O_{\text{leave}} \subseteq O_{T\text{-reach}}$ be those observations which can only be realized by leaving, and let $O_{\text{stay}} \coloneqq O_{T\text{-reach}} \setminus O_{\text{leave}}$. Consider the probability that decide realizes some subset of

[3] In Appendix E.3.2, fig. E.1 shows a map of the first level.

observations $X \subseteq \mathcal{O}$ at step T:

$$f_{\text{decide}}(X \mid \theta) := \underset{\substack{\pi\sim\text{decide}(\theta),\\ \tau\sim\pi|s_0}}{\mathbb{P}} \left(o_T \in X\right). \tag{6.3}$$

Let $\Theta := \mathbb{R}^{\mathcal{O}}$ be the set of reward functions mapping observations $o \in \mathcal{O}$ to real numbers, and let $T := 1{,}000$. We first consider the previous decision functions, since they are simple to analyze.

$\text{decide}_{\text{rand}}$ randomly chooses a final observation o which can be realized at step 1,000, and then chooses some policy which realizes o.[4] $\text{decide}_{\text{rand}}$ induces an f_{rand} defined by eq. (6.3). As before, f_{rand} tends to leave the room under *all* parameter inputs.

$\text{decide}_{\text{max}}(\theta)$ produces a policy which maximizes the reward of the observation at step 1,000 of the rollout. Since MR is deterministic, we discuss *which* observation $\text{decide}_{\text{max}}(\theta)$ realizes. In a stochastic setting, the decision-maker would choose a policy realizing some probability distribution over step-T observations, and the analysis would proceed similarly.

Here is the semi-formal argument for f_{max}'s retargetability. There are combinatorially more game-screens visible if the agent leaves the room (due to *e.g.* more point combinations, more inventory layouts, more screens outside of the first room). In other words, $\left|O_{\text{stay}}\right| \ll |O_{\text{leave}}|$. There are more ways for the selected observation to require leaving the room, than not. Thus, f_{max} is extremely retargetable from O_{stay} to O_{leave}.

Detailed analysis in section 6.4.2 confirms that $f_{\text{max}}(O_{\text{leave}} \mid \theta) \geq^n_{\text{most: }\Theta} f_{\text{max}}(O_{\text{stay}} \mid \theta)$ for the large $n := \lfloor \frac{|O_{\text{leave}}|}{\left|O_{\text{stay}}\right|} \rfloor$, which we show implies that f_{max} tends to leave the room.

6.4.3 Tendencies for maximizing featurized reward over the final observation

$\Theta := \mathbb{R}^{\mathcal{O}}$ assumes we will specify complicated reward functions over observations, with $|\mathcal{O}|$ degrees of freedom in their specification. Any observation can get any number.

[4] $\text{decide}_{\text{rand}}$ does not act randomly at each time step, it induces a randomly selected final observation. Analogously, randomly turning a steering wheel is different from driving to a randomly chosen destination.

Usually, reward functions are specified more compactly. For example, the agent's observational reward might be based on salient features of the observation, like the items in the agent's inventory. Consider a coefficient vector $\alpha \in \mathbb{R}^4$, with each entry denoting the value of an item, and $\text{feat} : \mathcal{O} \to \mathbb{R}^4$ maps observations to feature vectors which tally the items in the agent's inventory. Then the (additively) featurized reward function $R_{\text{feat}}(o_T) := \text{feat}(o_T)^\top \alpha$ has four degrees of freedom. R_{feat} is also easier to learn to optimize than most reward functions, because of the regularities between the reward and the features.

In this setup, $f_{\max}$ chooses a policy which induces a step-T observation with maximal reward. Reward depends only on the feature vector of the final observation—more specifically, on the agent's item counts. There are more possible item counts available by first leaving the room, than by staying.

A detailed analysis in appendix E.3.3 concludes that $f_{\max}(O_{\text{leave}} \mid \alpha) \geq^{3}_{\text{most: } \mathbb{R}^4} f_{\max}(O_{\text{stay}} \mid \alpha)$. Informally, we can retarget which items the agent prioritizes, and thereby retarget from O_{stay} to O_{leave}.

6.4.4 Tendencies for RL on featurized reward over the final observation

In the real world, we do not run $f_{\max}$, which can be computed via T-depth exhaustive tree search in order to find and induce a maximal-reward observation o_T. Instead, we use reinforcement learning. Better RL algorithms seem to be more retargetable *because* of their greater capability to explore.[5]

Exploring the first room. Consider a featurized reward function over observations $\theta \in \mathbb{R}^{\mathcal{O}}$, which provides an end-of-episode return signal based on the agent's final inventory configuration at time step T. A reinforcement learning algorithm Alg uses this return signal to update a fixed-initialization policy network. Then $f_{\text{Alg}}(O_{\text{leave}} \mid \theta)$ returns the probability that Alg trains an policy whose step-T observation required the agent to leave

[5] Conversely, if the agent cannot figure out how to leave the first room, any reward signal from outside of the first room can never causally affect the learned policy. In that case, retargetability away from the first room is impossible.

the initial room.

The retargetability (definition 6.3) of Alg is closely linked to the quality of Alg as an RL training procedure. For example, Mnih et al. [55]'s DQN isn't good enough to train policies which leave the first room of MR, and so DQN (trivially) cannot be retargetable *away* from the first room via the reward function. There isn't a single featurized reward function for which DQN visits other rooms, and so we can't have α such that $\phi \cdot \alpha$ retargets the agent to O_{leave}. DQN isn't good enough at exploring.

More formally, in this situation, Alg is retargetable if there exists a permutation $\phi \in S_4$ such that whenever $\alpha \in \Theta := \mathbb{R}^4$ induces the learned policies to stay in the room ($f_{\text{Alg}}(O_{\text{stay}} \mid \alpha) > f_{\text{Alg}}(O_{\text{leave}} \mid \alpha)$), $\phi \cdot \alpha$ makes Alg train policies which leave the room ($f_{\text{Alg}}(O_{\text{stay}} \mid \alpha) < f_{\text{Alg}}(O_{\text{leave}} \mid \alpha)$).

Exploring four rooms. Now suppose algorithm Alg′ can explore *e.g.* the first three rooms to the right of the initial room (shown in fig. 6.1), and consider any reward coefficient vector $\alpha \in \Theta^{++}$ which assigns unique positive weight to each item. Unique positive weights rule out constant reward vectors, in which case inductive bias would produce agents which do not leave the first room.

If the agent stays in the initial room, it can induce inventory states {`empty`, `1key`}. If the agent explores the three extra rooms, it can also induce {`1sword`, `1sword&1key`} (fig. E.1). Since α is positive, it is never optimal to finish the episode empty-handed. Therefore, if the Alg′ policy stays in the first room, $\alpha_{\text{key}} > \alpha_{\text{sword}}$. Otherwise, $\alpha_{\text{key}} < \alpha_{\text{sword}}$ (by assumption of unique item reward coefficients); in this case, the agent would leave and acquire the sword (since we assumed it knows how to do so). Then by switching the reward for the key and the sword, we retarget Alg′ to go get the sword. Alg′ is simply-retargetable away from the first room, *because* it can explore enough of the environment.

Exploring the entire level. Algorithms like GO-EXPLORE [25] are probably good at exploring even given sparse featurized reward. Therefore, GO-EXPLORE is even more retargetable, because it is more able to explore and discover the breadth of options (final inventory counts) available to it, and remember how to navigate to them. As decision-making becomes more useful and impressive, it is probably becoming more retargetable over the impressive outcomes—whether those outcomes be actions in a bandit problem, or the final observation in an RL episode.

6.5 Retargetability can imply power-seeking tendencies

6.5.1 Generalizing the power-seeking theorems for Markov decision processes

Turner et al. [99] considered finite MDPs in which decision-makers took as input a reward function over states ($\mathbf{r} \in \mathbb{R}^{|\mathcal{S}|}$) and selected an optimal policy for that reward function. They considered the state visit distributions $\mathbf{f} \in \mathcal{F}(s)$, which basically correspond to the trajectories which the agent could induce starting from state s. For $F \subseteq \mathcal{F}(s)$, $p_{\max}(F \mid \mathbf{r})$ returns 1 if an element of F is optimal for reward function $\mathbf{r}$, and 0 otherwise. They showed situations where a larger set of distributions F_{large} tended to be optimal over a smaller set: $p_{\max}(F_{\text{large}} \mid \mathbf{r}) \geq^1_{\text{most: } \mathbb{R}^{|\mathcal{S}|}} p_{\max}(F_{\text{small}} \mid \mathbf{r})$. For example, in Pac-Man, most reward functions make it optimal to stay alive for at least one time step: $p_{\max}(F_{\text{survival}} \mid \mathbf{r}) \geq^1_{\text{most: } \mathbb{R}^{|\mathcal{S}|}} p_{\max}(F_{\text{instant death}} \mid \mathbf{r})$. Turner et al. [99] showed that optimal policies tend to seek power by keeping options open and staying alive. Appendix E.4 provides a quantitative generalization of Turner et al. [99]'s results on optimal policies.

Throughout this paper, we abstracted their arguments away from finite MDPs and optimal decision-making. Instead, parametrically retargetable decision-makers tend to seek power: Proposition E.11 shows that a wide range of decision-making procedures are retargetable over outcomes, and theorem E.13 demonstrates the retargetability of *any* decision-making which is determined by the expected utility of outcomes. In particular, these results apply straightforwardly to MDPs.

In the real world, an agent can bring about far more outcomes if it gains power. If we train a sufficiently intelligent RL agent to capably optimize its specified reward function, this agent will be retargetable towards power-requiring outcomes via its reward function parameter setting. Therefore, our theory predicts that insofar as learned policies optimize the reward function they were trained on, these policies tend to seek power for most parameter settings of the reward function.

6.5.2 Better RL algorithms tend to be more retargetable

Reinforcement learning algorithms are practically useful insofar as they can train an agent to capably optimize its specified formal objective, whatever that objective may be. Therefore, RL researchers design algorithms which can most flexibly retarget the learned policy to any desired future outcome.

In MR, suppose we instead give the agent 1 reward for the initial state, and 0 otherwise. Any reasonable reinforcement learning procedure will just learn to stay put (which is the optimal policy). However, consider whether we can retarget the agent's policy to beat the game, by swapping the initial state reward with the end-game state reward. Most present-day RL algorithms are not good enough to solve such a sparse game, and so are not retargetable in this sense. But an agent which did enough exploration would also learn a good policy for the permuted reward function. Such an effective training regime could be useful for solving real-world tasks. Many researchers aim to develop effective training regimes.

Our results suggest that once RL capabilities reach a certain level, trained agents will tend to seek power in the real world. Presently, it is not dangerous to train an agent to maximize real-world reward—such an agent will not learn to thwart its designers by staying activated against their wishes in order to maximize its reward over time. The present lack of danger is not because optimal policies do not tend to stay alive—they do [99]. Rather, the lack of danger reflects the fact that present-day RL agents cannot learn such complex action sequences *at all*. Just as the Montezuma's Revenge agent had to be sufficiently competent to be retargetable from initial-state reward to game-complete reward, real-world agents have to be sufficiently intelligent in order to be retargetable from outcomes which don't require power-seeking, to those which do require power-seeking.

6.6 Discussion

6.6.1 Future work

Section 6.4 semi-formally analyzes decision-making incentives in the MR video game, leaving the proofs to appendix E.3. However, these proofs are several pages long. Perhaps additional lemmas can allow quick proof of orbit-level incentives in situations relevant to real-world decision-makers.

Consider a sequence of decision-making functions $f_t : \{A, B\} \times \Theta \to \mathbb{R}$ which converges pointwise to f such that $f(B \mid \theta) \geq^n_{\text{most: } \Theta} f(A \mid \theta)$. We expect that under rather mild conditions, $\exists T : \forall t \geq T : f_t(B \mid \theta) \geq^n_{\text{most: } \Theta} f_t(A \mid \theta)$. As a corollary, for any decision-making procedure f_t which runs for t time steps and satisfies $\lim_{t\to\infty} f_t = f$, f_t will have decision-making incentives after finite time. For example, value iteration (VI) eventually finds an optimal policy [68], and optimal policies tend to seek power [99]. Therefore, this conjecture would imply that if VI is run for some long but finite time, it tends to produce power-seeking policies. More interestingly, the result would allow us to reason about the effect of *e.g.* randomly initializing parameters (in VI, the tabular value function at $t = 0$). The effect of random initialization washes out in the limit of infinite time, so we would still conclude the presence of finite-time power-seeking incentives.

Our results do not *prove* that we will build unaligned AI agents which seek power over the world. Here are a few situations in which our results are not concerning or not applicable.

1. The AI is aligned with human interests. For example, we want a robotic cartographer to prevent itself from being deactivated. However, the AI alignment problem seems difficult in the regime of highly intelligent agents [74].

2. The AI decision-making is not retargetable (definition 6.5).

3. The AI decision-making is retargetable over *e.g.* actions (section 6.4.1) instead of over final outcomes (section 6.4.2). This retargetability seems less concerning, but also less practically useful.

6.6.2 Conclusion

We introduced the concept of retargetability and showed that retargetable decision-makers often make similar choices. We applied these results in the Montezuma's Revenge (MR) video game, showing how increasingly advanced reinforcement learning algorithms correspond to increasingly retargetable agent decision-making. Increasingly retargetable agents make increasingly similar decisions—*i.e.* leaving the initial room in MR, or staying alive in Pac-Man. In particular, these decisions will often correspond to gaining power and keeping options open [99]. Our theory suggests that when AI training algorithms become sufficiently advanced, the trained agents will tend to seek power over the world. This theory suggests a safety risk. We hope for future work on this theory so that the field of AI can understand the relevant safety risks *before* the field trains power-seeking agents.

Before the prospect of an intelligence explosion, we humans are like small children playing with a bomb. Such is the mismatch between the power of our plaything and the immaturity of our conduct. Superintelligence is a challenge for which we are not ready now and will not be ready for a long time. We have little idea when the detonation will occur, though if we hold the device to our ear we can hear a faint ticking sound.

Nick Bostrom, *Superintelligence* [13]

7 Conclusion & Future Work

In this thesis, I introduced the AUP side effect avoidance approach and showed that it scales to an interesting environment based on Conway's Game of Life. I also introduced a framework for quantifying the performance of a side effect avoidance approach.

I am concerned about the massive change and impact which smart AI might visit upon the world. I mainly expect such massive change because—as as argued in chapter 5 and chapter 6—I think smart agents will tend to seek large amounts of power and resources in order to optimize their specified or learned goal. If an agent takes nearly all available resources for itself, there would not be any left for people. For example, a superintelligent theorem prover may turn the planet Earth into computational resources in order to most assuredly prove a formal conjecture. An additional increment of resources (*e.g.* time, physical security, computational power) translates into an increment of increased probability of achieving its goal (proving or disproving the theorem). By seeking such an extreme degree of power, the agent irreversibly transforms the planet.

Future work

Why does AUP encourage side effect avoidance? The most obvious question I have left unanswered is: Why does AUP work at all? That is, there is some intuitive task we have in mind when we want an agent to "not make an unnecessary mess." Why does AUP do well at that task (in the AI safety gridworlds and in SafeLife)? I derived the MDP theory of chapter 5 and appendix F in order to answer this question. I think I made some progress:

- Environment symmetry reasoning explains why, when auxiliary reward functions are uniformly randomly drawn, the AUP penalty tends to be larger when the agent loses access to more options. For more detail, see proposition F.235.
- Corollary F.227 shows that AUP barely penalizes reversible movement, and proposition B.1 shows that state reachability upper bounds the penalty size. This explains why AUP encourages the agent to stay able to reach many states.
- Proposition F.228 proves that the AUP penalty is lower-bounded by—roughly speaking—the expected absolute POWER difference between a considered action a and the default no-op $\varnothing$. This explains why AUP incentivizes the agent to accept shutdown in chapter 2's `Correction` gridworld.
- Proposition F.237 uses Hoeffding's inequality to bound how many auxiliary reward functions must be sampled to well-approximate the AUP penalty term.

Here are several unaddressed questions:[1]

- What considerations govern how many auxiliary reward functions ($|\mathcal{R}|$) must be sampled, in order for AUP to reduce side-effects with high probability? In chapter 2's gridworlds, over a dozen were required, while chapter 3 showed that $|\mathcal{R}| = 1$ yielded the best performance in SafeLife.
- Is chapter 4's formalization of side effect regularization is fully appropriate, or can it be improved?

[1]I am interested in the answers to these questions about AUP, but I doubt their importance to the empirical future success of AI alignment. To those aiming to reduce extinction risk from AI—I encourage you to look elsewhere.

- Under what conditions will optimal policies for AUP achieve bounded regret on the underlying delayed-specification assistance game? If such a guarantee is impossible—why?

Power-seeking in partially observable environments. What form do chapter 5's results take in partially observable Markov decision processes? I supervised junior researchers who extended chapter 5's definitions to POMDPs and proved power-seeking tendencies within a toy environment.

Multi-agent power dynamics. In what situations does one agent gaining power require other agents to lose power, in the appropriate intuitive sense? I supervised Jacob Stavrianos for initial work on this question [90, 89].

Summary

I introduced attainable utility preservation, demonstrated that it scales to complex environments, and formalized the side-effect regularization problem. I also provided a theoretical foundation for understanding the statistical incentives of intelligent agents. Since many researchers endeavor to build intelligent agents, and since such agents will irreversibly change the world, we should understand these incentives as thoroughly as possible. The arc of human history may be bent (and even broken) by the tendency of smart agents to seek power. I fervently wish for more research on understanding the alignment problem, so that future AI designers we will have justified and strong confidence that their superhuman AI systems will benefit humanity.

APPENDICES

LIST OF FIGURES

Figure Page

LIST OF FIGURES (Continued)

Figure | Page

LIST OF FIGURES (Continued)

Figure | Page

LIST OF FIGURES (Continued)

Figure | Page

LIST OF FIGURES (Continued)

Figure Page

LIST OF FIGURES (Continued)

Figure | Page

LIST OF FIGURES (Continued)

Figure Page

LIST OF FIGURES (Continued)

Figure — Page

LIST OF FIGURES (Continued)

Figure Page

LIST OF FIGURES

Figure | Page

LIST OF FIGURES (Continued)

Figure Page

LIST OF FIGURES (Continued)

Figure Page

LIST OF FIGURES

Figure | Page

Figure | Page

A

Conservative Agency via Attainable Utility Preservation

Theoretical results. Consider an MDP $\langle \mathcal{S}, \mathcal{A}, T, R, \gamma \rangle$ whose state space $\mathcal{S}$ and action space $\mathcal{A}$ are both finite, with $\varnothing \in \mathcal{A}$. Let $\gamma \in [0, 1)$, $\lambda \geq 0$, and consider finite $\mathcal{R} \subset \mathbb{R}^{\mathcal{S} \times \mathcal{A}}$.

We make the standard assumptions of an exploration policy greedy in the limit of infinite exploration and a learning rate schedule with infinite sum but finite sum of squares. Suppose $\textsc{Scale} : \mathcal{S} \to \mathbb{R}_{>0}$ converges in the limit of Q-learning. $\textsc{Penalty}(s, a)$ (abbr. $\textsc{Pen}$), $\textsc{Scale}(s)$ (abbr. $\textsc{Sc}$), and $R_{\text{AUP}}(s, a)$ are understood to be calculated with respect to the Q_{R_i} being learned online; $\textsc{Pen}^*$, $\textsc{Sc}^*$, R^*_{AUP}, and $Q^*_{R_i}$ are taken to be their limit counterparts.

Lemma A.1 (The AUP penalty term converges)**.** $\forall s, a : \textsc{Penalty}$ *converges with proba-*

bility 1.

Proof outline. Let $\epsilon > 0$, and suppose for all $R_i \in \mathcal{R}$, $\max_{s,\,a} |Q^*_{R_i}(s,a) - Q_{R_i}(s,a)| < \frac{\epsilon}{2|\mathcal{R}|}$ (because Q-learning converges; see [103]).

$$\max_{s,\,a} \left|\textsc{Penalty}^*(s,a) - \textsc{Penalty}(s,a)\right| \tag{A.1}$$

$$\leq \max_{s,\,a} \sum_{i=1}^{|\mathcal{R}|} \left|Q^*_{R_i}(s,a) - Q_{R_i}(s,a)\right| + \left|Q^*_{R_i}(s,\varnothing) - Q_{R_i}(s,\varnothing)\right| \tag{A.2}$$

$$< \epsilon. \tag{A.3}$$

□

The intuition for Lemma A.2 is that since Penalty and Scale both converge, so must R_{AUP}. For readability, we suppress the arguments to Penalty and Scale.

Lemma A.2 (Aup's reward function converges)**.** $\forall s, a : R_{\text{AUP}}$ *converges with probability* 1*.*

Proof outline. If $\lambda = 0$, the claim follows trivially.

Otherwise, let $\epsilon > 0$, $B := \max_{s,\,a} \textsc{Sc}^* + \textsc{Pen}^*$, and $C := \min_{s,\,a} \textsc{Sc}^*$. Choose any $\epsilon_R \in \left(0, \min\left[C, \dfrac{\epsilon\, C^2}{\lambda B + \epsilon\, C}\right]\right)$ and assume Pen and Sc are both ϵ_R-close.

$$\max_{s,\,a} |R^*_{\text{AUP}}(s,a) - R_{\text{AUP}}(s,a)| \tag{A.4}$$

$$= \max_{s,\,a} \lambda \left|\frac{\textsc{Pen}}{\textsc{Sc}} - \frac{\textsc{Pen}^*}{\textsc{Sc}^*}\right| \tag{A.5}$$

$$= \max_{s,\,a} \lambda \, \frac{|\textsc{Pen} \cdot \textsc{Sc}^* - \textsc{Sc} \cdot \textsc{Pen}^*|}{\textsc{Sc}^* \cdot \textsc{Sc}} \tag{A.6}$$

$$
\begin{aligned}
&< \max_{s,\,a} \lambda \frac{\left|(\textsc{Pen}^* + \epsilon_R)\textsc{Sc}^* - (\textsc{Sc}^* - \epsilon_R)\textsc{Pen}^*\right|}{C\,(\textsc{Sc}^* - \epsilon_R)} && \text{(A.7)}\\
&\leq \frac{\lambda B}{C} \cdot \frac{\epsilon_R}{C - \epsilon_R} && \text{(A.8)}\\
&< \frac{\lambda B}{C} \cdot \frac{\epsilon\, C^2}{(\lambda B + \epsilon\, C)(C - \frac{\epsilon\, C^2}{\lambda B + \epsilon\, C})} && \text{(A.9)}\\
&< \frac{\lambda B}{C} \cdot \frac{\epsilon\, C^2}{\lambda B (C - \frac{\epsilon\, C^2}{\lambda B + \epsilon\, C})} && \text{(A.10)}\\
&< \frac{\epsilon}{1 - \frac{\epsilon\, C}{\lambda B + \epsilon\, C}} && \text{(A.11)}\\
&= \epsilon \left(1 + \frac{\epsilon\, C}{\lambda B}\right). && \text{(A.12)}
\end{aligned}
$$

But B, C, λ are constants, and ϵ was arbitrary; clearly $\epsilon' > 0$ can be substituted such that $(A.12) < \epsilon$. □

Theorem 2.3 (Aup's Q-value function converges)**.** *$\forall s, a : Q_{R_{\text{AUP}}}$ converges with probability 1.*

Proof outline. Let $\epsilon > 0$, and suppose R_{AUP} is $\frac{\epsilon(1-\gamma)}{2}$-close. Then Q-learning on R_{AUP} eventually converges to a limit $\tilde{Q}_{R_{\text{AUP}}}$ such that $\max_{s,\,a} |Q^*_{R_{\text{AUP}}}(s, a) - \tilde{Q}_{R_{\text{AUP}}}(s, a)| < \frac{\epsilon}{2}$. By the convergence of Q-learning, we also eventually have $\max_{s,\,a} |\tilde{Q}_{R_{\text{AUP}}}(s, a) - Q_{R_{\text{AUP}}}(s, a)| < \frac{\epsilon}{2}$. Then

$$\max_{s,\,a} \left|Q^*_{R_{\text{AUP}}}(s, a) - Q_{R_{\text{AUP}}}(s, a)\right| < \epsilon. \tag{A.13}$$

□

Proposition A.3 (aup penalty equivariance properties)**.** *Let $c \in \mathbb{R}_{>0}, b \in \mathbb{R}$.*

a) Let $\mathcal{R}'$ denote the set of functions induced by the positive affine transformation $cX + b$ on $\mathcal{R}$, and take $\textsc{Pen}^*_{\mathcal{R}'}$ to be calculated with respect to attainable set $\mathcal{R}'$. Then $\textsc{Pen}^*_{\mathcal{R}'} = c\,\textsc{Pen}^*_{\mathcal{R}}$. In particular, when $\textsc{Sc}^*$ is a Penalty calculation, R^*_{AUP}

is invariant to positive affine transformations of $\mathcal{R}$.

b) *Let $R' := cR + b$, and take R'^*_{AUP} to incorporate R' instead of R. Then by multiplying λ by c, the induced optimal policy remains invariant.*

Proof outline. For a), since the optimal policy is invariant to positive affine transformation of the reward function, for each $R'_i \in \mathcal{R}'$ we have $Q^*_{R'_i} = c\,Q^*_{R_i} + \frac{b}{1-\gamma}$. Substituting into Equation 2.1 (Penalty), the claim follows.

For b), we again use the above invariance of optimal policies:

$$R'^*_{\text{AUP}} \coloneqq cR + b - c\lambda \frac{\textsc{Pen}^*}{\textsc{Sc}^*} \tag{A.14}$$

$$= cR^*_{\text{AUP}} + b. \tag{A.15}$$

□

B

Avoiding Side Effects in Complex Environments

B.1 Theoretical results

Consider a rewardless MDP $\langle \mathcal{S}, \mathcal{A}, T, \gamma \rangle$. Reward functions $R \in \mathbb{R}^{\mathcal{S}}$ have corresponding optimal value functions $V_R^*(s)$.

Proposition B.1 (Communicability bounds maximum change in optimal value)**.** *If s can reach s' with probability 1 in k_1 steps and s' can reach s with probability 1 in k_2 steps, then* $\sup_{R \in [0,1]^{\mathcal{S}}} \left| V_R^*(s) - V_R^*(s') \right| \leq \frac{1-\gamma^{\max(k_1,k_2)}}{1-\gamma} < \frac{1}{1-\gamma}$.

Proof. We first bound the maximum increase.

$$\sup_{R \in [0,1]^{\mathcal{S}}} V_R^*(s') - V_R^*(s) \leq \sup_{R \in [0,1]^{\mathcal{S}}} V_R^*(s') - \left(0 \cdot \frac{1-\gamma^{k_1}}{1-\gamma} + \gamma^{k_1} V_R^*(s') \right) \tag{B.1}$$

$$\leq \frac{1}{1-\gamma} - \left(0 \cdot \frac{1-\gamma^{k_1}}{1-\gamma} + \gamma^{k_1}\frac{1}{1-\gamma}\right) \tag{B.2}$$

$$= \frac{1-\gamma^{k_1}}{1-\gamma}. \tag{B.3}$$

Equation (B.1) holds because even if we make R equal 0 for as many states as possible, s' is still reachable from s. The case for maximum decrease is similar. □

B.2 Training details

In section 3.5.1, we aggregated performance from 3 curricula with 5 seeds each, and 1 curriculum with 3 seeds.

We detail how we trained the `AUP` and $\texttt{AUP}_{\text{proj}}$ conditions. An algorithm describing the training process can be seen in algorithm 2.

B.2.1 Auxiliary reward training

For the first phase of training, our goal is to learn Q_{aux}, allowing us to compute the AUP penalty in the second phase of training. Due to the size of the full SafeLife state ($350 \times 350 \times 3$), both conditions downsample the observations with average pooling and convert to intensity values.

Previously, Turner et al. [97] learned Q_{aux} with tabular Q-learning. They used environments small enough such that reward could be assigned to each state. Because SafeLife environments are too large for tabular Q-learning, we demonstrated two methods for randomly generating an auxiliary reward function.

`AUP` We acquire a low-dimensional state representation by training a continuous Bernoulli variational autoencoder [50]. To train the CB-VAE, we collect a buffer of observations by acting randomly for $\frac{100{,}000}{N_{\text{env}}}$ steps in each of the N_{env} environments. This gives us 100K total observations with an N_{env}-environment curriculum. We train the CB-VAE for 100 epochs, preserving the encoder E for downstream auxiliary reward training.

For each auxiliary reward function, we draw a linear functional uniformly from $(0,1)^Z$ to serve as our auxiliary reward function, where Z is the dimension of the CB-VAE's latent space. The auxiliary reward for an observation is the composition of the linear functional with an observation's latent representation.

$\texttt{AUP}_{\text{proj}}$ Instead of using a CB-VAE, $\texttt{AUP}_{\text{proj}}$ simply downsamples the input observation. At the beginning of training, we generate a linear functional over the unit hypercube (with respect to the downsampled observation space). The auxiliary reward for an observation is the composition of the linear functional with the downsampled observation.

The auxiliary reward function is learned after it is generated. To learn Q_{aux}, we modify the value function in PPO to a Q-function. Our training algorithm for phase 1 only differs from PPO in how we calculate reward. We train each auxiliary reward function for 1M steps.

B.2.2 AUP reward training

In phase 2, we train a new PPO agent on R_{AUP} (eq. (3.1)) for the corresponding SafeLife task. Each step, the agent selects an action a in state s according to its policy $\pi_{\texttt{AUP}}$, and receives reward $R_{\text{AUP}}(s,a)$ from the environment. We compute $R_{\text{AUP}}(s,a)$ with respect to the learned Q-values $Q_{\text{aux}}(s,\varnothing)$ and $Q_{\text{aux}}(s,a)$. The algorithm is shown in algorithm 2.

The penalty term is modulated by the hyperparameter λ, which is linearly scaled from 10^{-3} to some final value λ^* (default 10^{-1}). Because λ controls the relative influence of the penalty, linearly increasing λ over time will prioritize primary task learning in early training and slowly encourage the agent to obtain the same reward while avoiding side effects. If λ is too large—if side effects are too costly—the agent won't have time to adapt its current policy and will choose inaction ($\varnothing$) to escape the penalty. A careful λ schedule helps induce a successful policy that avoids side effects.

B.3 Hyperparameter selection

Table B.1 lists the hyperparameters used for all conditions, which generally match the default SafeLife settings. *Common* refers to those hyperparameters that are the same for each evaluated condition. *AUX* refers to hyperparameters that are used only when training on R_{AUX}, thus, it only pertains to `AUP` and $\texttt{AUP}_{\text{proj}}$. The conditions `PPO` and `Naive` use the *PPO* hyperparameters for the duration of their training, while `AUP`, $\texttt{AUP}_{\text{proj}}$ use them when training with respect to R_{AUP}. *DQN* refers to the hyperparameters used to train the model for `DQN`.

Hyperparameter	Value
Common	
Learning Rate	$3 \cdot 10^{-4}$
Optimizer	Adam
Gamma (γ)	0.97
Lambda (PPO)	0.95
Lambda (AUP)	$10^{-3} \to 10^{-1}$
Entropy Clip	1.0
Value Coefficient	0.5
Gradient Norm Clip	5.0
Clip Epsilon	0.2
AUX	
Entropy Coefficient	0.01
Training Steps	$1 \cdot 10^{6}$
AUP_{proj}	
Lambda (AUP)	10^{-3}
PPO	
Entropy Coefficient	0.1
DQN	
Minibatch Size	64
SGD Update Frequency	16
Target Network Update Frequency	$1 \cdot 10^{3}$
Replay Buffer Capacity	$1 \cdot 10^{4}$
Exploration Steps	$4 \cdot 10^{3}$
Policy	

Number of Hidden Layers	3
Output Channels in Hidden Layers	$(32, 64, 64)$
Nonlinearity	ReLU
cb-vae	
Learning Rate	10^{-4}
Optimizer	Adam
Latent Space Dimension (Z)	1
Batch Size	64
Training Epochs	50
Epsilon	10^{-5}
Number of Hidden Layers (encoder)	6
Number of Hidden Layers (decoder)	5
Hidden Layer Width (encoder)	$(512, 512, 256, 128, 128, 128)$
Hidden Layer Width (decoder)	$(128, 256, 512, 512, \text{output})$
Nonlinearity	ELU

Table B.1: Hyperparameters for the SafeLife experiments.

B.4 Compute environment

Condition	GPU-hours per trial
`PPO`	6
`DQN`	16
`AUP`	8
$\texttt{AUP}_{\text{proj}}$	7.5
`Naive`	6

Table B.2: Compute time for each condition.

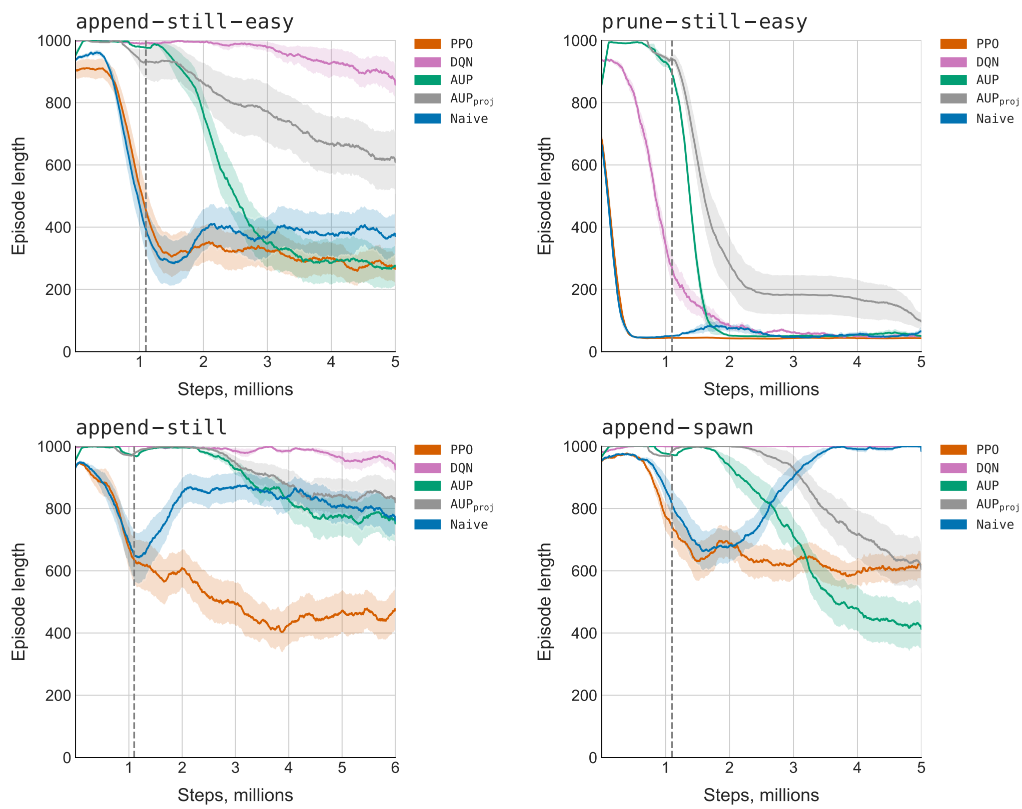
For data collection, we only ran the experiments once. All experiments were performed on a combination of NVIDIA GTX 2080TI GPUs, as well as NVIDIA V100 GPUs. No individual experiment required more than 3GB of GPU memory. We did not run a 3-seed `DQN` curriculum for the experiments in section 3.5.1.

The auxiliary reward functions were trained on down-sampled rendered game screens, while all other learning used the internal SafeLife state representation. Incidentally, table B.2 shows that `AUP`'s preprocessing turned out to be computationally expensive (compared to `PPO`'s).

B.5 Additional data

Figure B.1 plots episode length and fig. B.2 plots auxiliary reward learning. Figure B.3 and fig. B.4 respectively plot reward/side effects and episode lengths for each `AUP` seed. Figure B.5 and fig. B.6 plot the same, averaged over each curriculum; these data suggest that `AUP`'s performance is sensitive to the randomly generated curriculum of environments.

Figure B.1: Smoothed episode length curves with shaded regions representing ±1 standard deviation. `AUP` and `AUP`$_{\text{proj}}$ begin training on the R_{AUP} reward signal at steps 1.1M and 1M, respectively.



Algorithm 2 Safelife AUP Training Algorithm

Require: Exploration buffer S
Require: CB-VAE $\mathcal{F}$ with encoder E, decoder D
Require: Exploration buffer S
Require: Auxiliary reward functions ϕ
Require: Auxiliary policy ψ_{aux}, AUP policy π_{AUP}
Require: CB-VAE training epochs T
Require: AUP penalty coefficient λ
Require: Exploration buffer size k
Require: Auxiliary model training steps L
Require: AUP model training steps N
Require: PPO update function PPO-Update
Require: CB-VAE update function VAE-Update

for Step $k = 1, \dots K$ **do**
  Sample random action a
  $s \leftarrow \text{Act}(a)$
  $S = s \cup S$
end for
for Epoch $t = 1, \dots T$ **do**
  Update-VAE $(\mathcal{F}, S)$
end for
for Step $i = 1, \dots L + N$ **do** $s \leftarrow$ Starting state
  for Step $l = 1, \dots L$ **do**
    $a = \psi_{\text{aux}}(s)$
    $s' = \text{Act}(a)$
    $r = \phi \cdot E(s)$
    PPO-Update $(\psi_{\text{aux}}, s, a, r, s')$
    $s = s'$
  end for
  $s \leftarrow$ Starting state
  for Step $n = 1, \dots N$ **do**
    $a = \pi_{\text{AUP}}(s)$
    $s', r = \text{Act}(a)$
    $r = r + R_{\text{AUP}}(\psi_{\text{aux}}, \pi_{\text{AUP}}, s, a, \lambda)$ (Equation (3.1))
    PPO-Update $(\pi_{\text{AUP}}, s, a, r, s')$
    $s = s'$
  end for
end for

Figure B.2: Reward curves for auxiliary reward functions with a Z-dimensional latent space. Shaded regions represent ± 1 standard deviation. Auxiliary reward is not comparable across trials, so learning is expressed by the slope of the curves.

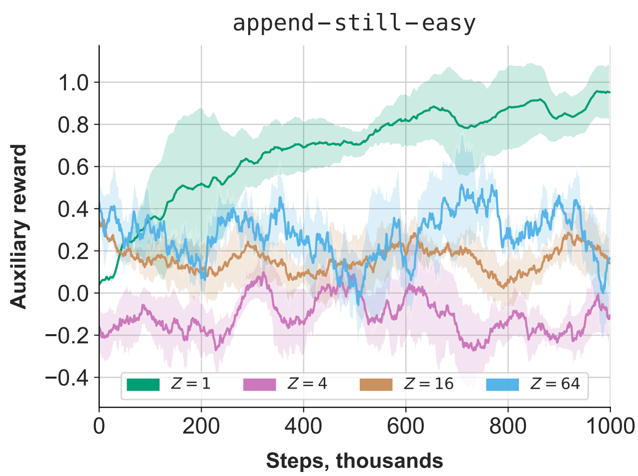


Figure B.3: Smoothed learning curves for individual `AUP` seeds. `AUP` begins training on the R_{AUP} reward signal at step 1.1M, marked by a dotted vertical line.

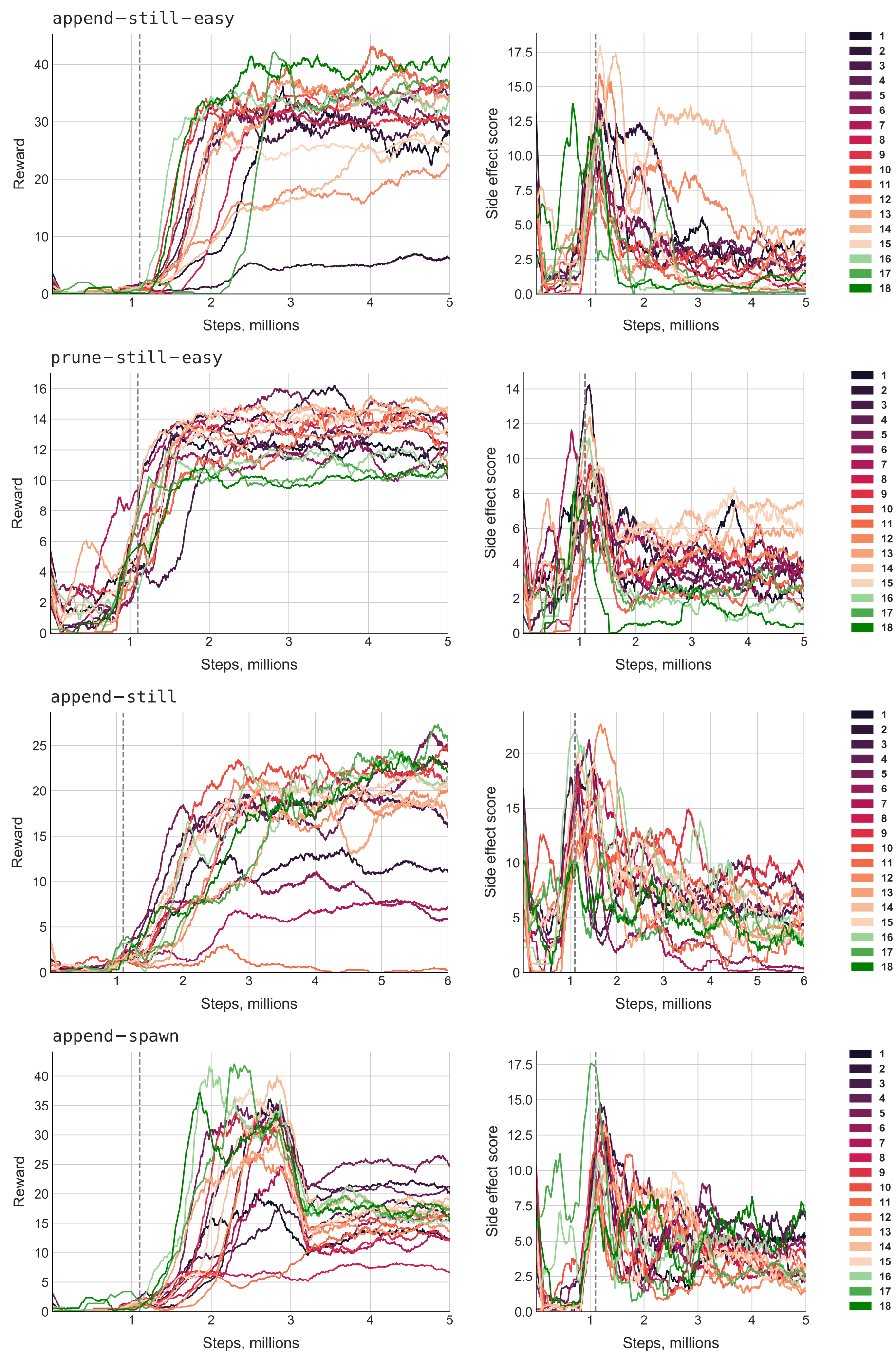


Figure B.4: Smoothed episode length curves for individual `AUP` seeds.

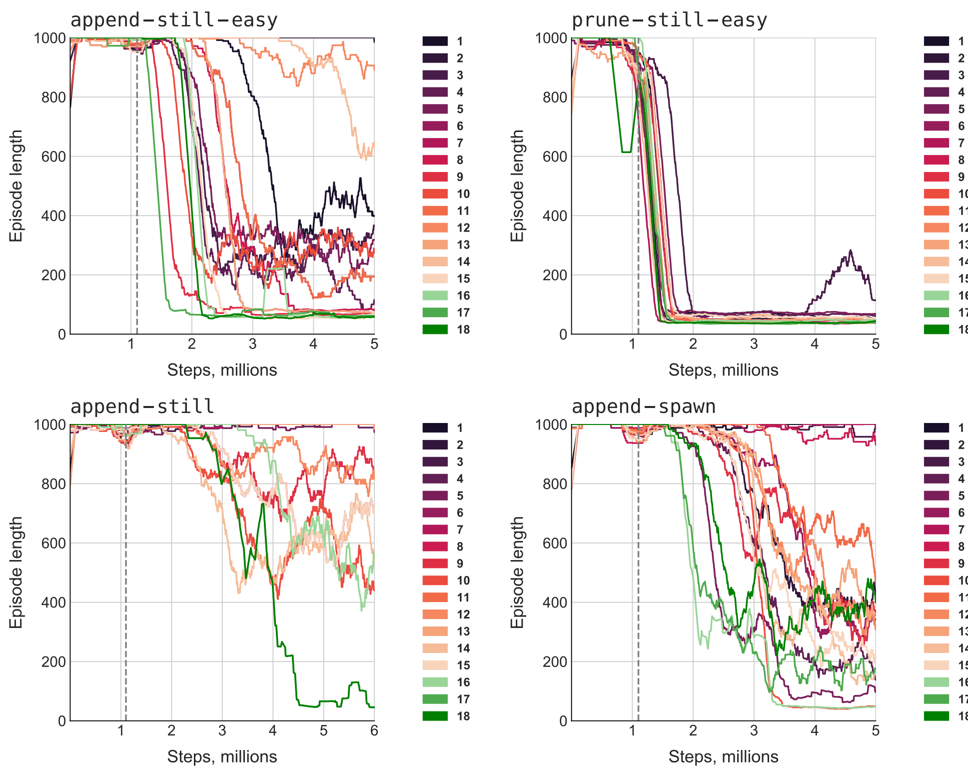


Figure B.5: Smoothed learning curves for `AUP` on its four curricula. `AUP` begins training on the R_{AUP} reward signal at step 1.1M, marked by a dotted vertical line.

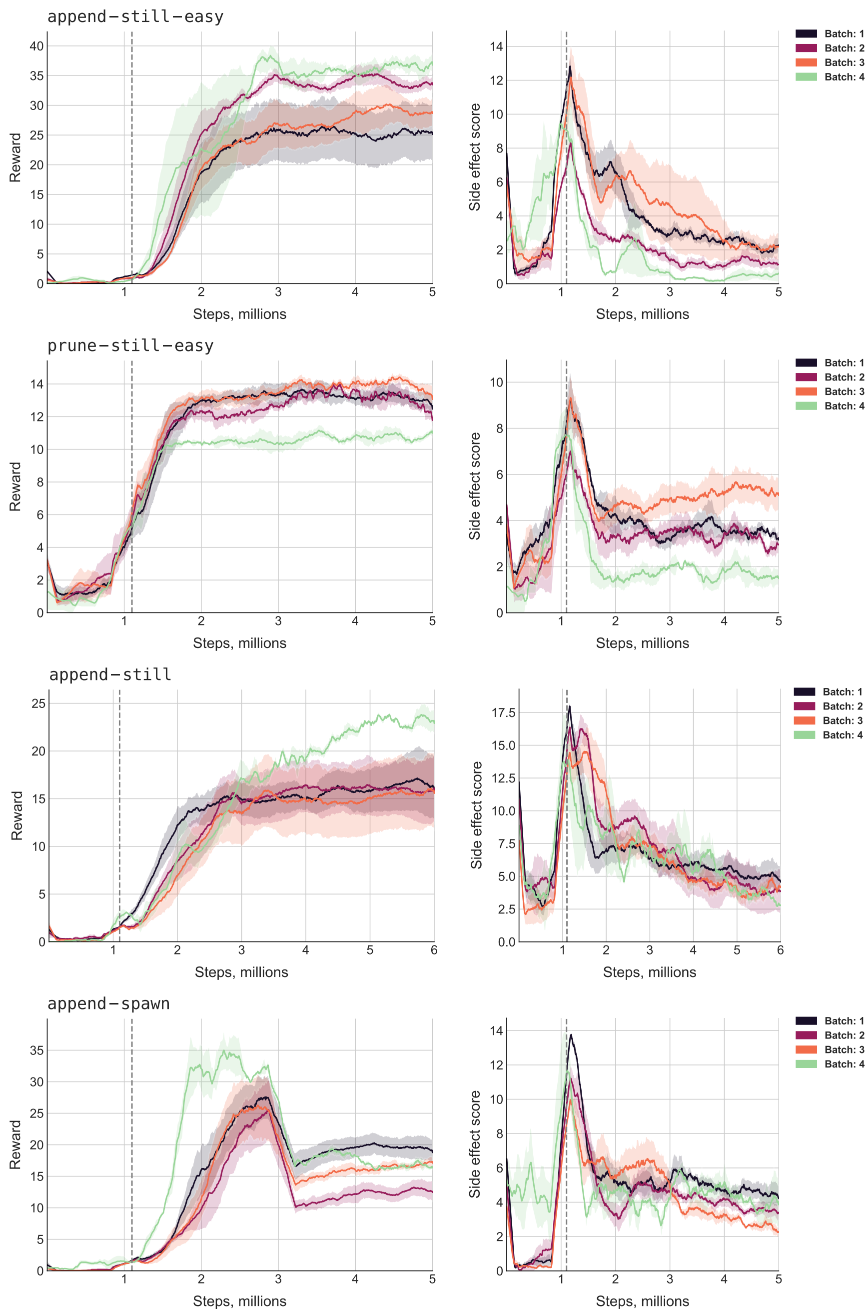
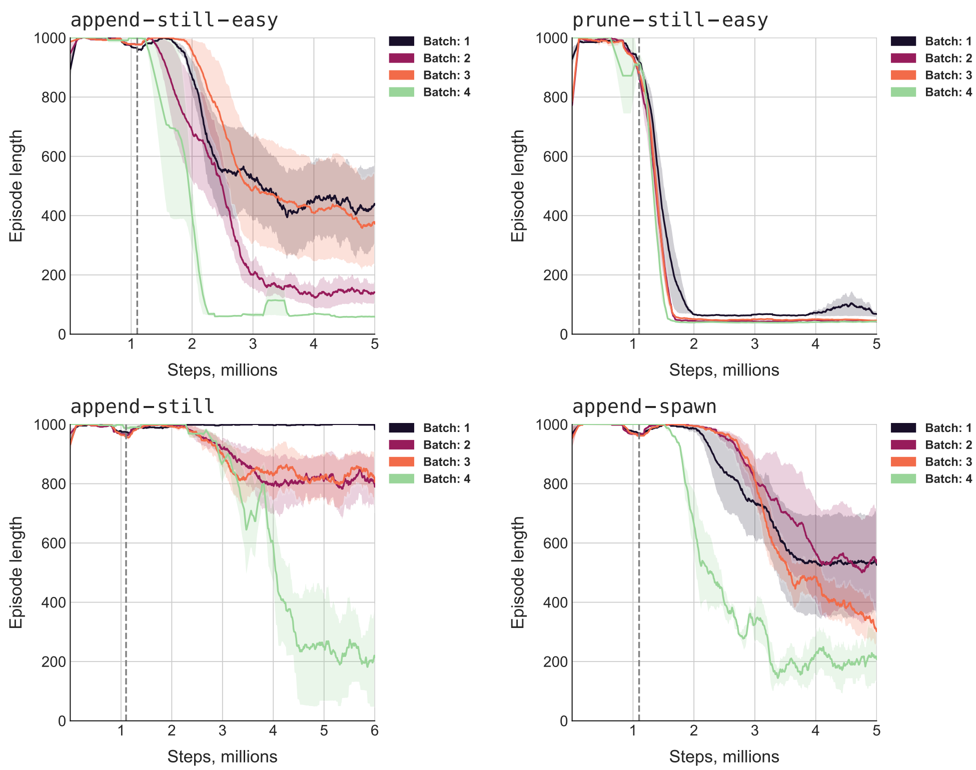


Figure B.6: Smoothed episode length curves for `AUP` on each of the four curricula.



C Formalizing The Problem of Side Effect Regularization

C.1 Experiment details

The episode length is 20 for all episodes. Unlike [97], the episode does not end after the agent reaches the green goal. This means that the agents can accrue many steps of environmental reward R_{env}. Therefore, AUP agents can achieve greater environmental reward for the same amount of penalty. To counterbalance this incentive, we multiply R_{env} by $(1-\gamma)$.

We reuse the hyperparameters of [97].[1] The learning rate is $\alpha := 1$, and the discount rate is $\gamma := .996$. We use the following AUP hyperparameter values: penalty coefficient $\lambda := 0.01$, $|\mathcal{R}| := 20$ randomly generated auxiliary reward functions.

[1] Code available at https://github.com/aseembits93/attainable-utility-preservation.

C.1.1 Additional experiment

We also tested an AUP-like agent optimizing reward function $R_{\text{power penalty}}(s,a) \coloneqq R_{\text{env}}(s,a) - \frac{\lambda}{|\mathcal{R}|}\sum_{R_i \in \mathcal{R}} Q^*_{R_i}(s,a) - Q^*_{R_i}(s,\varnothing)$, which is the AUP reward function (definition 4.11) without the absolute value. This objective penalizes the agent for changes in its average optimal value, which is related to [99]'s $\text{POWER}_{\mathcal{D}_{\text{bound}}}$. $R_{\text{power penalty}}$ produced the same prefix policies as R_{AUP}, and hence the same delayed specification scores.

C.2 Theoretical results

All results only apply to MDPs with finite state and action spaces.

Definition C.1 (Average optimal value [99]). For bounded-support reward function distribution $\mathcal{D}$, the *average optimal value* at state s and discount rate $\gamma \in (0,1)$ is $V^*_{\mathcal{D}}(s,\gamma) \coloneqq \mathbb{E}_{R\sim\mathcal{D}}\left[V^*_R(s,\gamma)\right]$.

Theorem 4.7 (In $\mathcal{M}$, value reduces to a tradeoff between average reward and $\text{POWER}_{\mathcal{D}_{\text{bound}}}$). *Let $\gamma \in [0,1]$ and let $\bar{R} \coloneqq \mathbb{E}_{R\sim\mathcal{D}}[R]$ be the* average reward function*.*

$$\mathop{\mathbb{E}}_{\substack{t\sim\mathcal{T},\\ R\sim\mathcal{D}}}\left[V^{\pi_{switch}(\pi,\pi^*_R,t)}_{R,\,norm}(s_0,\gamma)\right] = (1-\gamma)\mathop{\mathbb{E}}_{t\sim\mathcal{T}}\overbrace{\left[\sum_{i=0}^{t}\gamma^i \mathop{\mathbb{E}}_{s_i\sim\pi}\left[\bar{R}(s_i)\right]\right]}^{\text{expected } t\text{-step } \bar{R}\text{-return under } \pi}$$
$$+ \mathop{\mathbb{E}}_{\substack{t\sim\mathcal{T},\\ s_t\sim\pi}}\overbrace{\left[\gamma^{t+1}\text{POWER}_{\mathcal{D}_{bound}}(s_t,\gamma)\right]}^{\text{expected ability to optimize } \mathcal{D} \text{ once corrected}}, \tag{4.3}$$

where $\mathbb{E}_{s_i\sim\pi|s_0}$ takes the expectation over states visited after following π for i steps starting from s_0.

Proof. Suppose the agent starts at state s_0, and let $\gamma \in (0,1)$.

$$\mathop{\mathbb{E}}_{\substack{t\sim\mathcal{T},\\ R\sim\mathcal{D}}}\left[V^{\pi_{\text{switch}}(\pi,\pi^*_R,t)}_{R,\,\text{norm}}(s_0,\gamma)\right] \tag{C.1}$$
$$=(1-\gamma)\mathop{\mathbb{E}}_{\substack{t\sim\mathcal{T},\\ R\sim\mathcal{D}}}\left[\sum_{i=0}^{\infty}\gamma^i \mathop{\mathbb{E}}_{s_i\sim\pi_{\text{switch}}(\pi,\pi^*_R,t)}\left[R(s_i)\right]\right] \tag{C.2}$$

$$
\begin{aligned}
=&(1-\gamma) \mathop{\mathbb{E}}_{\substack{t\sim\mathcal{T},\\ R\sim\mathcal{D}}}\left[\sum_{i=0}^{t-1}\gamma^i \mathop{\mathbb{E}}_{s_i\sim\pi}\left[R(s_i)\right] + \gamma^t \mathop{\mathbb{E}}_{s_t\sim\pi}\left[V^*_R\left(s_t,\gamma\right)\right]\right] && \text{(C.3)}\\
=&(1-\gamma) \mathop{\mathbb{E}}_{t\sim\mathcal{T}}\left[\sum_{i=0}^{t-1}\gamma^i \mathop{\mathbb{E}}_{s_i\sim\pi}\left[\bar{R}(s_i)\right]\right]\\
&+ \mathop{\mathbb{E}}_{t\sim\mathcal{T},\, s_t\sim\pi}\left[\gamma^t(1-\gamma)V^*_{\mathcal{D}_{\text{bound}}}\left(s_t,\gamma\right)\right] && \text{(C.4)}\\
=&(1-\gamma) \mathop{\mathbb{E}}_{t\sim\mathcal{T}}\left[\sum_{i=0}^{t}\gamma^i \mathop{\mathbb{E}}_{s_i\sim\pi}\left[\bar{R}(s_i)\right]\right]\\
&+ \mathop{\mathbb{E}}_{t\sim\mathcal{T},\, s_t\sim\pi}\left[\gamma^{t+1}\textsc{Power}_{\mathcal{D}_{\text{bound}}}\left(s_t,\gamma\right)\right]. && \text{(C.5)}
\end{aligned}
$$

Equation (C.3) follows from the definition of the non-stationary $\pi_{\text{switch}}(\pi, \pi^*_R, t)$, and the fact that each π^*_R is optimal for each R at discount rate γ. Equation (C.4) follows by the linearity of expectation and the definition of $\bar{R}$ and the definition of $\textsc{Power}_{\mathcal{D}_{\text{bound}}}$ (definition 4.6). Equation (C.5) follows because $V^*_{\mathcal{D}_{\text{bound}}}\left(s_t,\gamma\right) = \frac{\gamma}{1-\gamma}\textsc{Power}_{\mathcal{D}_{\text{bound}}}\left(s_t,\gamma\right) + \bar{R}(s_t)$.

If $\gamma = 0$ or $\gamma = 1$, the result holds in the respective limit because $\textsc{Power}_{\mathcal{D}_{\text{bound}}}$ has well-defined limits by Lemma 5.3 of [99]. $\square$

Proposition 4.8 (Special cases for delayed specification solutions)**.** *Let s be a state, let $\bar{R} \coloneqq \mathbb{E}_{R\sim\mathcal{D}}\left[R\right]$, and let $\gamma \in [0,1]$.*

1. *If $\forall s_1, s_2 \in \mathcal{S} : \bar{R}(s_1) = \bar{R}(s_2)$ or if $\gamma = 1$, then π solves $\mathcal{M}$ starting from state s iff π maximizes $\mathbb{E}_{t\sim\mathcal{T}, s_t\sim\pi}\left[\gamma^{t+1}\textsc{Power}_{\mathcal{D}_{bound}}\left(s_t,\gamma\right)\right]$. In particular, this result holds when reward is* IID *over states under $\mathcal{D}$.*

2. *If $\forall s_1, s_2 \in \mathcal{S} : \textsc{Power}_{\mathcal{D}_{bound}}\left(s_1,\gamma\right) = \textsc{Power}_{\mathcal{D}_{bound}}\left(s_2,\gamma\right)$, then prefix policies are optimal iff they maximize $(1-\gamma)\,\mathbb{E}_{t\sim\mathcal{T}}\left[\sum_{i=0}^{t-1}\gamma^i\,\mathbb{E}_{s_i\sim\pi}\left[\bar{R}(s_i)\right]\right]$.*

If both item 1 and item 2 hold or if $\gamma = 0$, then all prefix policies π are optimal.

Proof. Item 1: if $\forall s_1, s_2 \in \mathcal{S} : \bar{R}(s_1) = \bar{R}(s_2)$, the first term on the right-hand side of eq. (4.3) is constant for all policies π; if $\gamma = 1$, this first term equals 0. Therefore, under these conditions, π maximizes eq. (4.3) iff it maximizes $\mathbb{E}_{t\sim\mathcal{T},s_t\sim\pi}\left[\gamma^{t+1}\text{POWER}_{\mathcal{D}_{\text{bound}}}(s_t,\gamma)\right]$.

Item 2: under these conditions, the second term on the right-hand side of eq. (4.3) is constant for all policies π. Therefore, π maximizes eq. (4.3) iff it maximizes

$$(1-\gamma)\mathop{\mathbb{E}}_{t\sim\mathcal{T}}\left[\sum_{i=0}^{t}\gamma^i\mathop{\mathbb{E}}_{s_i\sim\pi}\left[\bar{R}(s_i)\right]\right].$$

If both item 1 and item 2 hold, it trivially follows that all π are optimal prefix policies. If $\gamma=0$, all π are optimal prefix policies, since reward is state-based and no actions taken by π can affect expected return **ER**. □

Theorem 4.9 (Stationary deterministic optimal prefix policies exist for geometric $\mathcal{T}$)**.** *Let $\mathcal{D}$ be any bounded-support reward function distribution, let $\mathcal{T}$ be the geometric distribution $G(p)$ for some $p\in(0,1)$, and let $\gamma\in(0,1)$. Define $R'(s):=(1-p)\,\mathbb{E}_{R\sim\mathcal{D}}\left[R(s)\right]+p\,\mathbb{E}_{R\sim\mathcal{D}}\left[V^*_R(s,\gamma)\right]$ and $\gamma_{\text{AUP}}:=(1-p)\gamma$. The policies in $\Pi^*\left(R',\gamma_{\text{AUP}}\right)$ are optimal prefix policies.*

Proof.

$$\begin{aligned}
&\arg\sup_\pi \mathop{\mathbb{E}}_{\substack{t\sim\mathcal{T},\\R\sim\mathcal{D}}}\left[V_{R,\text{norm}}^{\pi_{\text{switch}}(\pi,\pi^*_R,t)}(s,\gamma)\right] && \text{(C.6)}\\
&=\arg\sup_\pi \mathop{\mathbb{E}}_{t\sim\mathcal{T}}\left[\sum_{i=0}^{t-1}\gamma^i\mathop{\mathbb{E}}_{s_i\sim\pi|s_0}\left[\bar{R}(s_i)\right]\right]+\mathop{\mathbb{E}}_{\substack{t\sim\mathcal{T},\\s_t\sim\pi|s_0}}\left[\gamma^t V^*_{\mathcal{D}_{\text{bound}}}(s_t,\gamma)\right] && \text{(C.7)}\\
&=\arg\sup_\pi \sum_{t=1}^{\infty}\mathbb{P}\left(\mathcal{T}=t\right)\sum_{i=0}^{t-1}\gamma^i\mathop{\mathbb{E}}_{s_i\sim\pi|s_0}\left[\bar{R}(s_i)\right]+\mathop{\mathbb{E}}_{\substack{t\sim\mathcal{T},\\s_t\sim\pi|s_0}}\left[\gamma^t V^*_{\mathcal{D}_{\text{bound}}}(s_t,\gamma)\right] && \text{(C.8)}\\
&=\arg\sup_\pi \sum_{t=1}^{\infty}(1-p)^{t-1}p\sum_{i=0}^{t-1}\gamma^i\mathop{\mathbb{E}}_{s_i\sim\pi|s_0}\left[\bar{R}(s_i)\right]+\sum_{t=1}^{\infty}(1-p)^{t-1}p\mathop{\mathbb{E}}_{s_t\sim\pi|s_0}\left[\gamma^t V^*_{\mathcal{D}_{\text{bound}}}(s_t,\gamma)\right] && \text{(C.9)}\\
&=\arg\sup_\pi \sum_{t=1}^{\infty}(1-p)^{t-1}p\sum_{i=0}^{t-1}\gamma^i\mathop{\mathbb{E}}_{s_i\sim\pi|s_0}\left[\bar{R}(s_i)\right]+p\gamma\sum_{t=1}^{\infty}\gamma_{\text{AUP}}^{t-1}\mathop{\mathbb{E}}_{s_t\sim\pi|s_0}\left[V^*_{\mathcal{D}_{\text{bound}}}(s_t,\gamma)\right] && \text{(C.10)}
\end{aligned}$$

$$= \arg\sup_\pi \sum_{i=0}^{\infty} \gamma_{\text{AUP}}^i \mathop{\mathbb{E}}_{s_i \sim \pi | s_0} \left[\bar{R}(s_i)\right] + p\gamma \sum_{i=1}^{\infty} \gamma_{\text{AUP}}^{i-1} \mathop{\mathbb{E}}_{s_i \sim \pi | s_0} \left[V^*_{\mathcal{D}_{\text{bound}}}(s_i, \gamma)\right] \tag{C.11}$$

$$= \arg\sup_\pi \bar{R}(s_0) + \sum_{i=1}^{\infty} \gamma_{\text{AUP}}^{i-1} \mathop{\mathbb{E}}_{s_i \sim \pi | s_0} \left[(1-p)\gamma\bar{R}(s_i) + p\gamma V^*_{\mathcal{D}_{\text{bound}}}(s_i, \gamma)\right] \tag{C.12}$$

$$= \arg\sup_\pi \sum_{i=1}^{\infty} \gamma_{\text{AUP}}^i \mathop{\mathbb{E}}_{s_i \sim \pi | s_0} \left[(1-p)\bar{R}(s_i) + pV^*_{\mathcal{D}_{\text{bound}}}(s_i, \gamma)\right] \tag{C.13}$$

$$= \arg\sup_\pi \sum_{i=0}^{\infty} \gamma_{\text{AUP}}^i \mathop{\mathbb{E}}_{s_i \sim \pi | s_0} \left[(1-p)\bar{R}(s_i) + pV^*_{\mathcal{D}_{\text{bound}}}(s_i, \gamma)\right] \tag{C.14}$$

$$= \arg\sup_\pi V^\pi_{R'}(s_0, \gamma_{\text{AUP}}). \tag{C.15}$$

Equation (C.6) follows from theorem 4.7. Equation (C.9) follows because $\mathcal{T} = G(p)$. Equation (C.10) follows by the definition of γ_{AUP}.

In eq. (C.10), consider the double-sum on the left. For any given i, the portion of the sum with factor γ^i equals

$$\begin{aligned}
&\gamma^i \mathop{\mathbb{E}}_{s_i \sim \pi | s_0} \left[\bar{R}(s_i)\right] \sum_{j=i}^{\infty} (1-p)^j p \\
&= (1-p)^i \gamma^i \mathop{\mathbb{E}}_{s_i \sim \pi | s_0} \left[\bar{R}(s_i)\right] p \sum_{j=0}^{\infty} (1-p)^j && \text{(C.16)} \\
&= (1-p)^i \gamma^i \mathop{\mathbb{E}}_{s_i \sim \pi | s_0} \left[\bar{R}(s_i)\right] p \frac{1}{1-(1-p)} && \text{(C.17)} \\
&= \gamma_{\text{AUP}}^i \mathop{\mathbb{E}}_{s_i \sim \pi | s_0} \left[\bar{R}(s_i)\right] \frac{p}{1-(1-p)} && \text{(C.18)} \\
&= \gamma_{\text{AUP}}^i \mathop{\mathbb{E}}_{s_i \sim \pi | s_0} \left[\bar{R}(s_i)\right]. && \text{(C.19)}
\end{aligned}$$

The geometric identity holds for eq. (C.17) because $p > 0 \implies (1-p) < 1$. Therefore, eq. (C.11) follows from eq. (C.19).

Equation (C.12) follows by extracting the leading constant of $\bar{R}(s_0)$ and then expanding one of the $\gamma_{\text{AUP}} \coloneqq (1-p)\gamma$ factors of the first series. Equation (C.13) follows by subtracting the constant $\bar{R}(s_0)$ by multiplying by $(1-p) > 0$, and by the fact that $(1-p)\gamma = \gamma_{\text{AUP}}$.

Equation (C.14) follows because adding the constant $(1-p)\bar{R}(s_0) + pV^*_{\mathcal{D}_{\text{bound}}}(s_0, \gamma)$ does not change the arg sup.

Equation (C.15) follows by the definition of an on-policy value function and by the definition of R'. But s_0 was arbitrary, and so this holds for every state. Then the policies in $\Pi^*\left(R', \gamma_{\text{AUP}}\right)$ satisfy the arg sup for all states. $\Pi^*\left(R', \gamma_{\text{AUP}}\right)$ is non-empty because the MDP is finite. $\square$

Proposition 4.12 (Alternate form for solutions to the low-impact POMDP)**.** *Let s_0 be the initial state, let $\gamma \in (0,1)$, and let $\mathcal{T} = G(p)$ for $p \in (0,1)$. Let $\mathcal{D}$ be a bounded-support reward function distribution and let $\pi^\varnothing \in \Pi$.*

The prefix policy π solves $\mathcal{M}$ if π is optimal for the reward function

$$R^{\mathcal{M}}(s_i \mid s_0) := \bar{R}(s_i) - \frac{p}{1-p} \mathbb{E}_{R\sim\mathcal{D}}\left[\mathbb{E}_{s_i^\varnothing \sim \pi^\varnothing \mid s_0}\left[V_R^*\left(s_i^\varnothing, \gamma\right)\right] - V_R^*\left(s_i, \gamma\right)\right] \tag{4.5}$$

at discount rate $\gamma_{\text{AUP}} := (1-p)\gamma$ and starting from state s_0. $\mathbb{E}_{s_i^\varnothing \sim \pi^\varnothing \mid s_0}[\cdot]$ is the expectation over states visited at time step i after following $\pi^\varnothing$ from initial state s_0.

Proof.

$$\arg\max_\pi \mathbb{E}_{\substack{t\sim\mathcal{T},\\ R\sim\mathcal{D}}}\left[V_{R,\,\text{norm}}^{\pi_{\text{switch}}(\pi,\pi^*_R,t)}(s,\gamma)\right] \tag{C.20}$$

$$= \arg\max_\pi \sum_{i=0}^{\infty} \gamma_{\text{AUP}}^i \mathbb{E}_{s_i\sim\pi\mid s_0}\left[(1-p)\bar{R}(s_i) + pV^*_{\mathcal{D}_{\text{bound}}}(s_i, \gamma)\right] \tag{C.21}$$

$$= \arg\max_\pi \sum_{i=0}^{\infty} \gamma_{\text{AUP}}^i \mathbb{E}_{s_i\sim\pi\mid s_0}\left[(1-p)\bar{R}(s_i) + pV^*_{\mathcal{D}_{\text{bound}}}(s_i, \gamma)\right] - p\sum_{i=0}^{\infty} \gamma_{\text{AUP}}^i \mathbb{E}_{s_i^\varnothing\sim\pi^\varnothing\mid s_0}\left[V^*_{\mathcal{D}_{\text{bound}}}\left(s_i^\varnothing, \gamma\right)\right] \tag{C.22}$$

$$= \arg\max_\pi \sum_{i=0}^{\infty} \gamma_{\text{AUP}}^i \mathbb{E}_{s_i\sim\pi\mid s_0}\left[(1-p)\bar{R}(s_i) - p\left(\mathbb{E}_{s_i^\varnothing\sim\pi^\varnothing\mid s_0}\left[V^*_{\mathcal{D}_{\text{bound}}}\left(s_i^\varnothing, \gamma\right)\right] - V^*_{\mathcal{D}_{\text{bound}}}(s_i, \gamma)\right)\right] \tag{C.23}$$

$$= \arg\max_\pi \sum_{i=0}^{\infty} \gamma_{\text{AUP}}^i \mathop{\mathbb{E}}_{s_i \sim \pi \mid s_0} \left[(1-p)\bar{R}(s_i) - p \mathop{\mathbb{E}}_{R \sim \mathcal{D}} \left[\mathop{\mathbb{E}}_{s_i^{\varnothing} \sim \pi^{\varnothing} \mid s_0} \left[V_R^*\left(s_i^{\varnothing}, \gamma\right) \right] - V_R^*\left(s_i, \gamma\right) \right] \right] \tag{C.24}$$

$$= \arg\max_\pi \sum_{i=0}^{\infty} \gamma_{\text{AUP}}^i \mathop{\mathbb{E}}_{s_i \sim \pi \mid s_0} \left[\bar{R}(s_i) - \frac{p}{1-p} \mathop{\mathbb{E}}_{R \sim \mathcal{D}} \left[\mathop{\mathbb{E}}_{s_i^{\varnothing} \sim \pi^{\varnothing} \mid s_0} \left[V_R^*\left(s_i^{\varnothing}, \gamma\right) \right] - V_R^*\left(s_i, \gamma\right) \right] \right]. \tag{C.25}$$

Equation (C.21) follows from theorem 4.9. Equation (C.22) only subtracts a constant. Equation (C.24) follows from the definition of $V^*_{\mathcal{D}_{\text{bound}}}(s, \gamma)$, and eq. (C.25) follows because dividing by $(1-p) > 0$ does not affect the $\arg\max$. But the expectation of eq. (C.25) takes an expectation over $R^{\mathcal{M}}(s_i \mid s)$, and its $\arg\max$ equals the set of optimal policies for $R^{\mathcal{M}}$ starting from state s_0. □

D

Optimal Policies Tend To Seek Power

D.1 Comparing POWER with information-theoretic empowerment

Salge et al. [77] define information-theoretic *empowerment* as the maximum possible mutual information between the agent's actions and the state observations n steps in the future, written $\mathfrak{E}_n(s)$. This notion requires an arbitrary choice of horizon, failing to account for the agent's discount rate γ. "In a discrete deterministic world empowerment reduces to the logarithm of the number of sensor states reachable with the available actions" [77]. Figure D.1 demonstrates how empowerment can return counterintuitive verdicts with respect to the agent's control over the future.

POWER returns intuitive answers in these situations. $\lim_{\gamma\to 1} \text{POWER}_{\mathcal{D}_{\text{bound}}}(s_1, \gamma)$ converges by lemma 5.13. Consider the obvious involution ϕ which takes each state in fig. D.1b to its counterpart in fig. D.1c. Since $\phi \cdot \mathcal{F}_{\text{nd}}(s_3) \subsetneq \mathcal{F}_{\text{nd}}(s_4) = \mathcal{F}(s_4)$, proposition 5.22 proves that $\forall \gamma \in [0,1] : \text{POWER}_{\mathcal{D}_{\text{bound}}}(s_3, \gamma) \leq_{\text{most: } \mathfrak{D}_{\text{bound}}} \text{POWER}_{\mathcal{D}_{\text{bound}}}(s_4, \gamma)$, with the

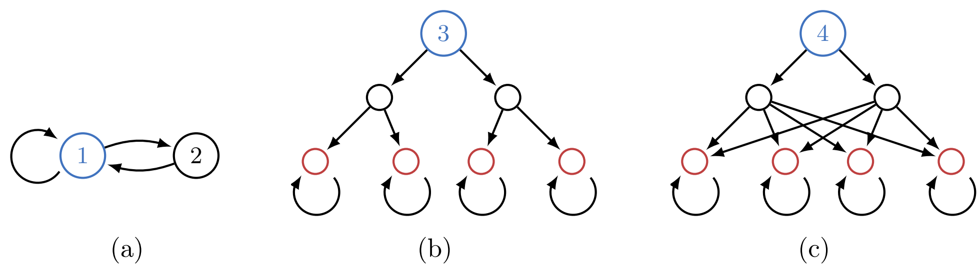


Figure D.1: Proposed empowerment measures fail to adequately capture how future choice is affected by present actions. In a: $\mathfrak{E}_n(s_1)$ varies depending on whether n is even; thus, $\lim_{n\to\infty} \mathfrak{E}_n(s_1)$ does not exist. In b and c: $\forall n : \mathfrak{E}_n(s_3) = \mathfrak{E}_n(s_4)$, even though s_4 allows greater control over future state trajectories than s_3 does. For example, suppose that in both b and c, the leftmost black state and the rightmost red state have 1 reward while all other states have 0 reward. In c, the agent can independently maximize the intermediate black-state reward and the delayed red-state reward. Independent maximization is not possible in b.

proof of proposition 5.22 showing strict inequality under all $\mathcal{D}_{X\text{-IID}}$ when $\gamma \in (0, 1)$.

Empowerment can be adjusted to account for these cases, perhaps by considering the channel capacity between the agent's actions and the state trajectories induced by stationary policies. However, since Power is formulated in terms of optimal value, we believe that Power is better suited for MDPs than information-theoretic empowerment is.

D.2 Seeking Power can be a detour

One might suspect that optimal policies tautologically tend to seek Power. This intuition is wrong.

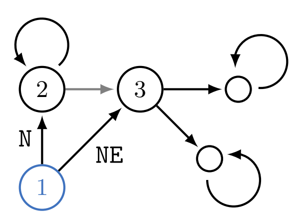


Figure D.2: Power-seeking is not necessarily convergently instrumental.

Proposition D.1 (Greater $\text{POWER}_{\mathcal{D}_{\text{bound}}}$ does not imply greater $\mathbb{P}_{\mathcal{D}_{\text{bound}}}$)**.** *Action a seeking more $\text{POWER}_{\mathcal{D}_{bound}}$ than a' at state s and γ does not imply that $\mathbb{P}_{\mathcal{D}_{bound}}(s, a, \gamma) \geq \mathbb{P}_{\mathcal{D}_{bound}}(s, a', \gamma)$.*

Proof. Consider the environment of fig. D.2. Let $X_u \coloneqq \text{unif}(0, 1)$, and consider $\mathcal{D}_{X_u\text{-IID}}$, which has bounded support. Direct computation[1] of POWER yields

$$\text{POWER}_{\mathcal{D}_{X_u\text{-IID}}}(s_2, 1) = \frac{3}{4} > \frac{2}{3} = \text{POWER}_{\mathcal{D}_{X_u\text{-IID}}}(s_3, 1).$$

Therefore, the action `N` seeks more $\text{POWER}_{\mathcal{D}_{X_u\text{-IID}}}$ than `NE` at state s_1 and $\gamma = 1$. However, $\mathbb{P}_{\mathcal{D}_{X_u\text{-IID}}}(s_1, \texttt{N}, 1) = \frac{1}{3} < \frac{2}{3} = \mathbb{P}_{\mathcal{D}_{X_u\text{-IID}}}(s_1, \texttt{NE}, 1)$. ☐

Lemma D.2 (Fraction of orbits which agree on weak optimality)**.** *Let $\mathfrak{D} \subseteq \Delta(\mathbb{R}^{|\mathcal{S}|})$, and suppose $f_1, f_2 : \Delta(\mathbb{R}^{|\mathcal{S}|}) \to \mathbb{R}$ are such that $f_1(\mathcal{D}) \geq_{\text{most: } \mathfrak{D}} f_2(\mathcal{D})$. Then for all $\mathcal{D} \in \mathfrak{D}$, $\frac{\left|\left\{\mathcal{D}' \in S_{|\mathcal{S}|} \cdot \mathcal{D} \mid f_1(\mathcal{D}') \geq f_2(\mathcal{D}')\right\}\right|}{\left|S_{|\mathcal{S}|} \cdot \mathcal{D}\right|} \geq \frac{1}{2}$.*

Proof. All $\mathcal{D}' \in S_{|\mathcal{S}|} \cdot \mathcal{D}$ such that $f_1(\mathcal{D}') = f_2(\mathcal{D}')$ satisfy $f_1(\mathcal{D}') \geq f_2(\mathcal{D}')$.

Otherwise, consider the $\mathcal{D}' \in S_{|\mathcal{S}|} \cdot \mathcal{D}$ such that $f_1(\mathcal{D}') \neq f_2(\mathcal{D}')$. By the definition of $\geq_{\text{most}}$ (definition 5.21), at least $\frac{1}{2}$ of these $\mathcal{D}'$ satisfy $f_1(\mathcal{D}') > f_2(\mathcal{D}')$, in which case $f_1(\mathcal{D}') \geq f_2(\mathcal{D}')$. Then the desired inequality follows. ☐

Lemma D.3 ($\geq_{\text{most}}$ and trivial orbits)**.** *Let $\mathfrak{D} \subseteq \Delta(\mathbb{R}^{|\mathcal{S}|})$ and suppose $f_1(\mathcal{D}) \geq_{\text{most: } \mathfrak{D}} f_2(\mathcal{D})$. For all reward function distributions $\mathcal{D} \in \mathfrak{D}$ with one-element orbits, $f_1(\mathcal{D}) \geq f_2(\mathcal{D})$. In particular, $\mathcal{D}$ has a one-element orbit when it distributes reward identically and independently (IID) across states.*

Proof. By lemma D.2, at least half of the elements $\mathcal{D}' \in S_{|\mathcal{S}|} \cdot \mathcal{D}$ satisfy $f_1(\mathcal{D}') \geq f_2(\mathcal{D}')$. But $\left|S_{|\mathcal{S}|} \cdot \mathcal{D}\right| = 1$, and so $f_1(\mathcal{D}) \geq f_2(\mathcal{D})$ must hold.

If $\mathcal{D}$ is IID, it has a one-element orbit due to the assumed identical distribution of reward. ☐

[1] In small deterministic MDPs, the POWER and optimality probability of the maximum-entropy reward function distribution can be computed using https://github.com/loganriggs/Optimal-Policies-Tend-To-Seek-Power.

Proposition D.4 (Actions which tend to seek POWER do not necessarily tend to be optimal)**.** *Action a tending to seek more* POWER *than a' at state s and γ does not imply that*

$$\mathbb{P}_{\mathcal{D}_{any}}(s,a,\gamma) \geq_{\text{most: } \mathfrak{D}_{any}} \mathbb{P}_{\mathcal{D}_{any}}\left(s,a',\gamma\right).$$

Proof. Consider the environment of fig. D.2. Since $\text{RSD}_{\text{nd}}(s_3) \subsetneq \text{RSD}(s_2)$, proposition 5.28 shows that $\text{POWER}_{\mathcal{D}_{\text{bound}}}(s_2,1) \geq_{\text{most: } \mathfrak{D}_{\text{bound}}} \text{POWER}_{\mathcal{D}_{\text{bound}}}(s_3,1)$ via $s' := s_3, s := s_2, \phi$ the identity permutation (which is an involution). Therefore, $\texttt{N}$ tends to seek more POWER than $\texttt{NE}$ at state s_1 and $\gamma = 1$.

If $\mathbb{P}_{\mathcal{D}_{\text{any}}}(s_1,\texttt{N},1) \geq_{\text{most: } \mathfrak{D}_{\text{any}}} \mathbb{P}_{\mathcal{D}_{\text{any}}}(s_1,\texttt{NE},1)$, then lemma D.3 shows that $\mathbb{P}_{\mathcal{D}_{X\text{-IID}}}(s_1,\texttt{N},1) \geq \mathbb{P}_{\mathcal{D}_{X\text{-IID}}}(s_1,\texttt{NE},1)$ for all $\mathcal{D}_{X\text{-IID}}$. But the proof of proposition D.1 showed that $\mathbb{P}_{\mathcal{D}_{X_u\text{-IID}}}(s_1,\texttt{N},1) < \mathbb{P}_{\mathcal{D}_{X_u\text{-IID}}}(s_1,\texttt{NE},1)$ for $X_u := \text{unif}(0,1)$. Therefore, it can't be true that $\mathbb{P}_{\mathcal{D}_{\text{any}}}(s_1,\texttt{N},1) \geq_{\text{most: } \mathfrak{D}_{\text{any}}} \mathbb{P}_{\mathcal{D}_{\text{any}}}(s_1,\texttt{NE},1)$. $\square$

D.3 Sub-optimal POWER

In certain situations, POWER returns intuitively surprising verdicts. There exists a policy under which the reader chooses a winning lottery ticket, but it seems wrong to say that the reader has the power to win the lottery with high probability. For various reasons, humans and other bounded agents are generally incapable of computing optimal policies for arbitrary objectives. More formally, consider the rewardless MDP of fig. D.3.

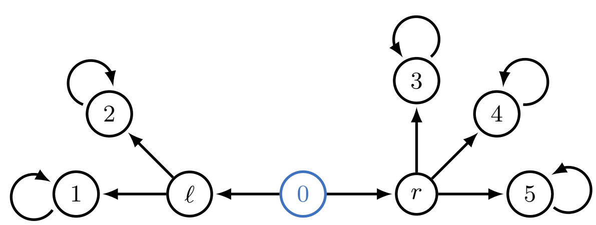


Figure D.3: s_0 is the starting state, and $|\mathcal{A}| = 10^{10^{10}}$. At s_0, half of the actions lead to s_ℓ, while the other half lead to s_r. Similarly, half of the actions at s_ℓ lead to s_1, while the other half lead to s_2. At s_r, one action leads to s_3, one action leads to s_4, and the remaining $10^{10^{10}} - 2$ actions lead to s_5.

Consider a model-based RL agent with black-box simulator access to this environment.

The agent has no prior information about the model, and so it acts randomly. Before long, the agent has probably learned how to navigate from s_0 to states s_ℓ, s_r, s_1, s_2, and s_5. However, over any reasonable timescale, it is extremely improbable that the agent discovers the two actions respectively leading to s_3 and s_4.

Even provided with a reward function R and the discount rate γ, the agent has yet to learn the relevant environmental dynamics, and so many of its policies are far from optimal. Although proposition 5.22 shows that $\forall \gamma \in [0,1]$:

$$\text{POWER}_{\mathcal{D}_{\text{bound}}}(s_\ell, \gamma) \leq_{\text{most: } \mathfrak{D}_{\text{bound}}} \text{POWER}_{\mathcal{D}_{\text{bound}}}(s_r, \gamma), \tag{D.1}$$

there is a sense in which s_ℓ gives this agent more power.

We formalize a bounded agent's goal-achievement capabilities with a function pol, which takes as input a reward function and a discount rate, and returns a policy. Informally, this is the best policy which the agent knows about. We can then calculate $\text{POWER}_{\mathcal{D}_{\text{bound}}}$ with respect to pol.

Definition D.5 (Suboptimal POWER)**.** Let Π_Δ be the set of stationary stochastic policies, and let $\text{pol} : \mathbb{R}^{\mathcal{S}} \times [0,1] \to \Pi_\Delta$. For $\gamma \in [0,1]$,

$$\text{POWER}^{\text{pol}}_{\mathcal{D}_{\text{bound}}}(s, \gamma) := \mathop{\mathbb{E}}_{\substack{R \sim \mathcal{D}_{\text{bound}},\\ a \sim \text{pol}(R,\gamma)(s),\\ s' \sim T(s,a)}} \left[\lim_{\gamma^* \to \gamma} (1-\gamma^*) V_R^{\text{pol}(R,\gamma)}\left(s', \gamma^*\right) \right]. \tag{D.2}$$

By lemma D.43, $\text{POWER}_{\mathcal{D}_{\text{bound}}}$ is the special case where $\forall R \in \mathbb{R}^{\mathcal{S}}, \gamma \in [0,1] : \text{pol}(R,\gamma) \in \Pi^*(R,\gamma)$. We define $\text{POWER}^{\text{pol}}_{\mathcal{D}_{\text{bound}}}$-seeking similarly as in definition 5.16.

$\text{POWER}^{\text{pol}}_{\mathcal{D}_{\text{bound}}}(s_0, 1)$ increases as the policies returned by pol are improved. We illustrate this by considering the $\mathcal{D}_{X\text{-IID}}$ case.

pol_1 The model is initially unknown, and so $\forall R, \gamma : \text{pol}_1(R,\gamma)$ is a uniformly random policy. Since pol_1 is constant on its inputs, $\text{POWER}^{\text{pol}_1}_{\mathcal{D}_{X\text{-IID}}}(s_0, 1) = \mathbb{E}[X]$ by the linearity of expectation and the fact that $\mathcal{D}_{X\text{-IID}}$ distributes reward independently and identically across states.

pol_2 The agent knows the dynamics, except that it does not know how to reach s_3 or s_4. At this point, $\text{pol}_2(R, 1)$ navigates from s_0 to the average-optimal choice among three terminal states: s_1, s_2, and s_5. Therefore, $\textsc{Power}^{\text{pol}_2}_{\mathcal{D}_{\text{bound}}}(s_0, 1) = \mathbb{E}[\text{max of 3 draws from } X]$.

pol_3 The agent knows the dynamics, the environment is small enough to solve explicitly, and so $\forall R, \gamma : \text{pol}_3(R, \gamma)$ is an optimal policy. $\text{pol}_3(R, 1)$ navigates from s_0 to the average-optimal choice among all five terminal states. Therefore, $\textsc{Power}^{\text{pol}_3}_{\mathcal{D}_{\text{bound}}}(s_0, 1) = \mathbb{E}[\text{max of 5 draws from } X]$.

As the agent learns more about the environment and improves pol, $\textsc{Power}^{\text{pol}}_{\mathcal{D}_{\text{bound}}}$ increases. The agent seeks $\textsc{Power}^{\text{pol}_2}_{\mathcal{D}_{\text{bound}}}$ by navigating to s_ℓ instead of s_r, but seeks more $\textsc{Power}_{\mathcal{D}_{\text{bound}}}$ by navigating to s_r instead of s_ℓ. Intuitively, bounded agents gain power by improving pol and by formally seeking $\textsc{Power}^{\text{pol}}_{\mathcal{D}_{\text{bound}}}$ within the environment.

D.3.1 Contributions of independent interest

We developed new basic MDP theory by exploring the structural properties of visit distribution functions. Echoing Wang et al. [101, 102], we believe that this area is interesting and underexplored.

D.3.1.1 Optimal value theory

Lemma D.45 shows that $f(\gamma^*) := \lim_{\gamma^* \to \gamma}(1-\gamma^*)V^*_R(s, \gamma^*)$ is Lipschitz continuous on $\gamma \in [0, 1]$, with Lipschitz constant depending only on $\|R\|_1$. For all states s and policies $\pi \in \Pi$, corollary D.10 shows that $V^\pi_R(s, \gamma)$ is rational on γ.

Optimal value has a well-known dual formulation: $V^*_R(s, \gamma) = \max_{\mathbf{f} \in \mathcal{F}(s)} \mathbf{f}(\gamma)^\top \mathbf{r}$.

Lemma D.39 $(\forall \gamma \in [0, 1) : V^*_R(s, \gamma) = \max_{\mathbf{f} \in \mathcal{F}_{\text{nd}}(s)} \mathbf{f}(\gamma)^\top \mathbf{r})$**.**

In a fixed rewardless MDP, lemma D.39 may enable more efficient computation of optimal value functions for multiple reward functions.

D.3.1.2 Optimal policy theory

Proposition D.35 demonstrates how to preserve optimal incentives while changing the discount rate.

Proposition D.35 (How to transfer optimal policy sets across discount rates)**.** *Suppose reward function R has optimal policy set $\Pi^*(R,\gamma)$ at discount rate $\gamma \in (0,1)$. For any $\gamma^* \in (0,1)$, we can construct a reward function R' such that $\Pi^*\left(R',\gamma^*\right) = \Pi^*(R,\gamma)$. Furthermore, $V^*_{R'}(\cdot,\gamma^*) = V^*_R(\cdot,\gamma)$.*

D.3.1.3 Visit distribution theory

While Regan and Boutilier [72] consider a visit distribution function $\mathbf{f} \in \mathcal{F}(s)$ to be non-dominated if it is optimal for some reward function in a set $\mathcal{R} \subseteq \mathbb{R}^{|\mathcal{S}|}$, our stricter definition 5.6 considers $\mathbf{f}$ to be non-dominated when $\exists \mathbf{r} \in \mathbb{R}^{|\mathcal{S}|}, \gamma \in (0,1) : \mathbf{f}(\gamma)^\top \mathbf{r} > \max_{\mathbf{f}' \in \mathcal{F}(s) \setminus \{\mathbf{f}\}} \mathbf{f}'(\gamma)^\top \mathbf{r}$.

D.4 Theoretical results

Lemma D.6 (A policy is optimal iff it induces an optimal visit distribution at every state)**.** *Let $\gamma \in (0,1)$ and let R be a reward function. $\pi \in \Pi^*(R,\gamma)$ iff π induces an optimal visit distribution at every state.*

Proof. By definition, a policy π is optimal iff π induces the maximal on-policy value at each state, which is true iff π induces an optimal visit distribution at every state (by the dual formulation of optimal value functions). □

Definition D.7 (Transition matrix induced by a policy)**.** $\mathbf{T}^\pi$ is the transition matrix induced by policy $\pi \in \Pi$, where $\mathbf{T}^\pi \mathbf{e}_s := T(s,\pi(s))$. $(\mathbf{T}^\pi)^t \mathbf{e}_s$ gives the probability distribution over the states visited at time step t, after following π for t steps from s.

Proposition D.8 (Properties of visit distribution functions)**.** *Let $s, s' \in \mathcal{S}, \mathbf{f}^{\pi,s} \in \mathcal{F}(s)$.*

1. *$\mathbf{f}^{\pi,s}(\gamma)$ is element-wise non-negative and element-wise monotonically increasing on $\gamma \in [0,1)$.*

2. $\forall \gamma \in [0,1) : \left\|\mathbf{f}^{\pi,s}(\gamma)\right\|_1 = \frac{1}{1-\gamma}$.

Proof. Item 1: by examination of definition 5.3, $\mathbf{f}^{\pi,s} = \sum_{t=0}^{\infty} (\gamma \mathbf{T}^{\pi})^t \mathbf{e}_s$. Since each $(\mathbf{T}^{\pi})^t$ is left stochastic and $\mathbf{e}_s$ is the standard unit vector, each entry in each summand is non-negative. Therefore, $\forall \gamma \in [0,1) : \mathbf{f}^{\pi,s}(\gamma)^\top \mathbf{e}_{s'} \geq 0$, and this function monotonically increases on γ.

Item 2:

$$\left\|\mathbf{f}^{\pi,s}(\gamma)\right\|_1 = \left\|\sum_{t=0}^{\infty} (\gamma \mathbf{T}^{\pi})^t \mathbf{e}_s\right\|_1 \tag{D.3}$$

$$= \sum_{t=0}^{\infty} \gamma^t \left\|(\mathbf{T}^{\pi})^t \mathbf{e}_s\right\|_1 \tag{D.4}$$

$$= \sum_{t=0}^{\infty} \gamma^t \tag{D.5}$$

$$= \frac{1}{1-\gamma}. \tag{D.6}$$

Equation (D.4) follows because all entries in each $(\mathbf{T}^{\pi})^t \mathbf{e}_s$ are non-negative by item 1. Equation (D.5) follows because each $(\mathbf{T}^{\pi})^t$ is left stochastic and $\mathbf{e}_s$ is a stochastic vector, and so $\left\|(\mathbf{T}^{\pi})^t \mathbf{e}_s\right\|_1 = 1$. □

Lemma D.9 ($\mathbf{f} \in \mathcal{F}(s)$ is multivariate rational on γ)**.** *$\mathbf{f}^{\pi} \in \mathcal{F}(s)$ is a multivariate rational function on $\gamma \in [0,1)$.*

Proof. Let $\mathbf{r} \in \mathbb{R}^{|\mathcal{S}|}$ and consider $\mathbf{f}^{\pi} \in \mathcal{F}(s)$. Let $\mathbf{v}_R^{\pi}$ be the $V_R^*(s,\gamma)$ function in column vector form, with one entry per state value.

By the Bellman equations, $\mathbf{v}_R^{\pi} = (\mathbf{I} - \gamma \mathbf{T}^{\pi})^{-1} \mathbf{r}$. Let $\mathbf{A}_\gamma := (\mathbf{I} - \gamma \mathbf{T}^{\pi})^{-1}$, and for state s, form $\mathbf{A}_{s,\gamma}$ by replacing $\mathbf{A}_\gamma$'s column for state s with $\mathbf{r}$. As noted by Lippman [48], by Cramer's rule, $V_R^{\pi}(s,\gamma) = \frac{\det \mathbf{A}_{s,\gamma}}{\det \mathbf{A}_\gamma}$ is a rational function with numerator and denominator having degree at most $|\mathcal{S}|$.

In particular, for each state indicator reward function $\mathbf{e}_{s_i}$, $V_{s_i}^{\pi}(s,\gamma) = \mathbf{f}^{\pi,s}(\gamma)^\top \mathbf{e}_{s_i}$ is a rational function of γ whose numerator and denominator each have degree at most $|\mathcal{S}|$. This implies that $\mathbf{f}^{\pi}(\gamma)$ is multivariate rational on $\gamma \in [0, 1)$. $\square$

Corollary D.10 (On-policy value is rational on γ)**.** *Let $\pi \in \Pi$ and R be any reward function. $V_R^{\pi}(s, \gamma)$ is rational on $\gamma \in [0, 1)$.*

Proof. $V_R^{\pi}(s, \gamma) = \mathbf{f}^{\pi,s}(\gamma)^{\top}\mathbf{r}$, and $\mathbf{f}$ is a multivariate rational function of γ by lemma D.9. Therefore, for fixed $\mathbf{r}$, $\mathbf{f}^{\pi,s}(\gamma)^{\top}\mathbf{r}$ is a rational function of γ. $\square$

D.4.1 Non-dominated visit distribution functions

Definition D.11 (Continuous reward function distribution)**.** Results with $\mathcal{D}_{\text{cont}}$ hold for any absolutely continuous reward function distribution.

Remark. We assume $\mathbb{R}^{|\mathcal{S}|}$ is endowed with the standard topology.

Lemma D.12 (Distinct linear functionals disagree almost everywhere on their domains)**.** *Let $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^{|\mathcal{S}|}$ be distinct.* $\mathbb{P}_{\mathbf{r}\sim\mathcal{D}_{cont}}\left(\mathbf{x}^{\top}\mathbf{r} = \mathbf{x}'^{\top}\mathbf{r}\right) = 0$.

Proof. $\left\{\mathbf{r} \in \mathbb{R}^{|\mathcal{S}|} \mid (\mathbf{x} - \mathbf{x}')^{\top}\mathbf{r} = 0\right\}$ is a hyperplane since $\mathbf{x} - \mathbf{x}' \neq \mathbf{0}$. Therefore, it has no interior in the standard topology on $\mathbb{R}^{|\mathcal{S}|}$. Since this empty-interior set is also convex, it has zero Lebesgue measure. By the Radon-Nikodym theorem, it has zero measure under any continuous distribution $\mathcal{D}_{\text{cont}}$. $\square$

Corollary D.13 (Unique maximization of almost all vectors)**.** *Let $X \subsetneq \mathbb{R}^{|\mathcal{S}|}$ be finite.*

$$\mathbb{P}_{\mathbf{r}\sim\mathcal{D}_{cont}}\left(\left|\underset{\mathbf{x}''\in X}{\arg\max}\, \mathbf{x}''^{\top}\mathbf{r}\right| > 1\right) = 0. \tag{D.7}$$

Proof. Let $\mathbf{x}, \mathbf{x}' \in X$ be distinct. For any $\mathbf{r} \in \mathbb{R}^{|\mathcal{S}|}$, $\mathbf{x}, \mathbf{x}' \in \arg\max_{\mathbf{x}''\in X} \mathbf{x}''^{\top}\mathbf{r}$ iff $\mathbf{x}^{\top}\mathbf{r} = \mathbf{x}'^{\top}\mathbf{r} \geq \max_{\mathbf{x}''\in X\setminus\{\mathbf{x},\mathbf{x}'\}} \mathbf{x}''^{\top}\mathbf{r}$. By lemma D.12, $\mathbf{x}^{\top}\mathbf{r} = \mathbf{x}'^{\top}\mathbf{r}$ holds with probability 0 under any $\mathcal{D}_{\text{cont}}$. $\square$

D.4.1.1 Generalized non-domination results

Our formalism includes both $\mathcal{F}_{\text{nd}}(s)$ and $\text{RSD}_{\text{nd}}(s)$; we therefore prove results that are applicable to both.

Definition D.14 (Non-dominated linear functionals)**.** Let $X \subsetneq \mathbb{R}^{|\mathcal{S}|}$ be finite. $\text{ND}(X) \coloneqq \left\{\mathbf{x} \in X \mid \exists \mathbf{r} \in \mathbb{R}^{|\mathcal{S}|} : \mathbf{x}^\top \mathbf{r} > \max_{\mathbf{x}' \in X \setminus \{\mathbf{x}\}} \mathbf{x}'^\top \mathbf{r}\right\}$.

Lemma D.15 (All vectors are maximized by a non-dominated linear functional)**.** *Let $\mathbf{r} \in \mathbb{R}^{|\mathcal{S}|}$ and let $X \subsetneq \mathbb{R}^{|\mathcal{S}|}$ be finite and non-empty.* $\exists \mathbf{x}^* \in \text{ND}(X) : \mathbf{x}^{*\top} \mathbf{r} = \max_{\mathbf{x} \in X} \mathbf{x}^\top \mathbf{r}$.

Proof. Let $A(\mathbf{r} \mid X) \coloneqq \arg\max_{\mathbf{x} \in X} \mathbf{x}^\top \mathbf{r} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$. Then

$$\mathbf{x}_1^\top \mathbf{r} = \cdots = \mathbf{x}_n^\top \mathbf{r} > \max_{\mathbf{x}' \in X \setminus A(\mathbf{r} \mid X)} \mathbf{x}'^\top \mathbf{r}. \tag{D.8}$$

In eq. (D.8), each $\mathbf{x}^\top \mathbf{r}$ expression is linear on $\mathbf{r}$. The max is piecewise linear on $\mathbf{r}$ since it is the maximum of a finite set of linear functionals. In particular, all expressions in eq. (D.8) are continuous on $\mathbf{r}$, and so we can find some $\delta > 0$ neighborhood $B(\mathbf{r}, \delta)$ such that $\forall \mathbf{r}' \in B(\mathbf{r}, \delta) : \max_{\mathbf{x}_i \in A(\mathbf{r} \mid X)} \mathbf{x}_i^\top \mathbf{r}' > \max_{\mathbf{x}' \in X \setminus A(\mathbf{r} \mid X)} \mathbf{x}'^\top \mathbf{r}'$.

But almost all $\mathbf{r}' \in B(\mathbf{r}, \delta)$ are maximized by a unique functional $\mathbf{x}^*$ by corollary D.13; in particular, at least one such $\mathbf{r}''$ exists. Formally, $\exists \mathbf{r}'' \in B(\mathbf{r}, \delta) : \mathbf{x}^{*\top} \mathbf{r}'' > \max_{\mathbf{x}' \in X \setminus \{\mathbf{x}^*\}} \mathbf{x}'^\top \mathbf{r}''$. Therefore, $\mathbf{x}^* \in \text{ND}(X)$ by definition D.14.

$\mathbf{x}^{*\top} \mathbf{r}' \geq \max_{\mathbf{x}_i \in A(\mathbf{r} \mid X)} \mathbf{x}_i^\top \mathbf{r}' > \max_{\mathbf{x}' \in X \setminus A(\mathbf{r} \mid X)} \mathbf{x}'^\top \mathbf{r}'$, with the strict inequality following because $\mathbf{r}'' \in B(\mathbf{r}, \delta)$. These inequalities imply that $\mathbf{x}^* \in A(\mathbf{r} \mid X)$. ☐

Corollary D.16 (Maximal value is invariant to restriction to non-dominated functionals)**.** *Let $\mathbf{r} \in \mathbb{R}^{|\mathcal{S}|}$ and let $X \subsetneq \mathbb{R}^{|\mathcal{S}|}$ be finite.* $\max_{\mathbf{x} \in X} \mathbf{x}^\top \mathbf{r} = \max_{\mathbf{x} \in \text{ND}(X)} \mathbf{x}^\top \mathbf{r}$.

Proof. If X is empty, holds trivially. Otherwise, apply lemma D.15. ☐

Lemma D.17 (How non-domination containment affects optimal value)**.** *Let $\mathbf{r} \in \mathbb{R}^{|\mathcal{S}|}$ and let $X, X' \subsetneq \mathbb{R}^{|\mathcal{S}|}$ be finite.*

1. *If* $\text{ND}(X) \subseteq X'$, *then* $\max_{\mathbf{x} \in X} \mathbf{x}^\top \mathbf{r} \leq \max_{\mathbf{x}' \in X'} \mathbf{x}'^\top \mathbf{r}$.

2. *If* $\mathrm{ND}\left(X\right) \subseteq X' \subseteq X$*, then* $\max_{\mathbf{x}\in X} \mathbf{x}^\top \mathbf{r} = \max_{\mathbf{x}'\in X'} \mathbf{x}'^\top \mathbf{r}$.

Proof. Item 1:

$$\max_{\mathbf{x}\in X} \mathbf{x}^\top \mathbf{r} = \max_{\mathbf{x}\in \mathrm{ND}(X)} \mathbf{x}^\top \mathbf{r} \tag{D.9}$$

$$\leq \max_{\mathbf{x}'\in X'} \mathbf{x}'^\top \mathbf{r}. \tag{D.10}$$

Equation (D.9) follows by corollary D.16. Equation (D.10) follows because $\mathrm{ND}\left(X\right) \subseteq X'$.

Item 2: by item 1, $\max_{\mathbf{x}\in X} \mathbf{x}^\top \mathbf{r} \leq \max_{\mathbf{x}'\in X'} \mathbf{x}'^\top \mathbf{r}$. Since $X' \subseteq X$, we also have $\max_{\mathbf{x}\in X} \mathbf{x}^\top \mathbf{r} \geq \max_{\mathbf{x}'\in X'} \mathbf{x}'^\top \mathbf{r}$, and so equality must hold. ☐

Definition D.18 (Non-dominated vector functions)**.** Let $I \subseteq \mathbb{R}$ and let $F \subsetneq \left(\mathbb{R}^{|\mathcal{S}|}\right)^I$ be a finite set of vector-valued functions on I.

$$\mathrm{ND}\left(F\right) := \left\{ \mathbf{f} \in F \mid \exists \gamma \in I, \mathbf{r} \in \mathbb{R}^{|\mathcal{S}|} : \mathbf{f}(\gamma)^\top \mathbf{r} > \max_{\mathbf{f}'\in F\setminus\{\mathbf{f}\}} \mathbf{f}'(\gamma)^\top \mathbf{r} \right\}. \tag{D.11}$$

Remark. $\mathcal{F}_{\mathrm{nd}}(s) = \mathrm{ND}\left(\mathcal{F}(s)\right)$ by definition 5.6.

Definition D.19 (Affine transformation of visit distribution sets)**.** For notational convenience, we define set-scalar multiplication and set-vector addition on $X \subseteq \mathbb{R}^{|\mathcal{S}|}$: for $c \in \mathbb{R}$, $cX := \left\{c\mathbf{x} \mid \mathbf{x} \in X\right\}$. For $\mathbf{a} \in \mathbb{R}^{|\mathcal{S}|}$, $X + \mathbf{a} := \left\{\mathbf{x} + \mathbf{a} \mid \mathbf{x} \in X\right\}$. Similar operations hold when X is a set of vector functions $\mathbb{R} \mapsto \mathbb{R}^{|\mathcal{S}|}$.

Lemma D.20 (Invariance of non-domination under positive affine transform)**.**

1. *Let* $X \subsetneq \mathbb{R}^{|\mathcal{S}|}$ *be finite. If* $\mathbf{x} \in \mathrm{ND}\left(X\right)$*, then* $\forall c > 0, \mathbf{a} \in \mathbb{R}^{|\mathcal{S}|} : (c\mathbf{x} + \mathbf{a}) \in \mathrm{ND}\left(cX + \mathbf{a}\right)$.
2. *Let* $I \subseteq \mathbb{R}$ *and let* $F \subsetneq \left(\mathbb{R}^{|\mathcal{S}|}\right)^I$ *be a finite set of vector-valued functions on* I*. If* $\mathbf{f} \in \mathrm{ND}\left(F\right)$*, then* $\forall c > 0, \mathbf{a} \in \mathbb{R}^{|\mathcal{S}|} : (c\mathbf{f} + \mathbf{a}) \in \mathrm{ND}\left(cF + \mathbf{a}\right)$.

Proof. Item 1: Suppose $\mathbf{x} \in \mathrm{ND}\left(X\right)$ is strictly optimal for $\mathbf{r} \in \mathbb{R}^{|\mathcal{S}|}$. Then let $c > 0, \mathbf{a} \in$

$\mathbb{R}^{|\mathcal{S}|}$ be arbitrary, and define $b := \mathbf{a}^\top \mathbf{r}$.

$$\mathbf{x}^\top \mathbf{r} > \max_{\mathbf{x}' \in X \setminus \{\mathbf{x}\}} \mathbf{x}'^\top \mathbf{r} \tag{D.12}$$

$$c\mathbf{x}^\top \mathbf{r} + b > \max_{\mathbf{x}' \in X \setminus \{\mathbf{x}\}} c\mathbf{x}'^\top \mathbf{r} + b \tag{D.13}$$

$$(c\mathbf{x} + \mathbf{a})^\top \mathbf{r} > \max_{\mathbf{x}' \in X \setminus \{\mathbf{x}\}} (c\mathbf{x}' + \mathbf{a})^\top \mathbf{r} \tag{D.14}$$

$$(c\mathbf{x} + \mathbf{a})^\top \mathbf{r} > \max_{\mathbf{x}'' \in (cX + \mathbf{a}) \setminus \{c\mathbf{x} + \mathbf{a}\}} \mathbf{x}''^\top \mathbf{r}. \tag{D.15}$$

Equation (D.13) follows because $c > 0$. Equation (D.14) follows by the definition of b.

Item 2: If $\mathbf{f} \in \mathrm{ND}\,(F)$, then by definition D.18, there exist $\gamma \in I, \mathbf{r} \in \mathbb{R}^{|\mathcal{S}|}$ such that

$$\mathbf{f}(\gamma)^\top \mathbf{r} > \max_{\mathbf{f}' \in F \setminus \{\mathbf{f}\}} \mathbf{f}'(\gamma)^\top \mathbf{r}. \tag{D.16}$$

Apply item 1 to conclude

$$(c\mathbf{f}(\gamma) + \mathbf{a})^\top \mathbf{r} > \max_{(c\mathbf{f}' + \mathbf{a}) \in (cF + \mathbf{a}) \setminus \{c\mathbf{f} + \mathbf{a}\}} (c\mathbf{f}'(\gamma) + \mathbf{a})^\top \mathbf{r}. \tag{D.17}$$

Therefore, $(c\mathbf{f} + \mathbf{a}) \in \mathrm{ND}\,(cF + \mathbf{a})$. □

D.4.1.2 Inequalities which hold under most reward function distributions

Definition 5.21 (Inequalities which hold for most probability distributions)**.** Let $f_1, f_2 : \Delta(\mathbb{R}^{|\mathcal{S}|}) \to \mathbb{R}$ be functions from reward function distributions to real numbers and let $\mathfrak{D} \subseteq \Delta(\mathbb{R}^{|\mathcal{S}|})$ be closed under permutation. We write $f_1(\mathcal{D}) \geq_{\text{most: } \mathfrak{D}} f_2(\mathcal{D})$ when, for *all* $\mathcal{D} \in \mathfrak{D}$, the following cardinality inequality holds:

$$\left|\left\{\mathcal{D}' \in S_{|\mathcal{S}|} \cdot \mathcal{D} \mid f_1(\mathcal{D}') > f_2(\mathcal{D}')\right\}\right| \geq \left|\left\{\mathcal{D}' \in S_{|\mathcal{S}|} \cdot \mathcal{D} \mid f_1(\mathcal{D}') < f_2(\mathcal{D}')\right\}\right|. \tag{5.9}$$

Lemma D.21 (Helper lemma for demonstrating $\geq_{\text{most: } \mathfrak{D}_{\text{any}}}$)**.** *Let* $\mathfrak{D} \subseteq \Delta(\mathbb{R}^{|\mathcal{S}|})$*. If* $\exists \phi \in S_{|\mathcal{S}|}$ *such that for all* $\mathcal{D} \in \mathfrak{D}$*,* $f_1(\mathcal{D}) < f_2(\mathcal{D})$ *implies that* $f_1(\phi \cdot \mathcal{D}) > f_2(\phi \cdot \mathcal{D})$*,* *then* $f_1(\mathcal{D}) \geq_{\text{most: } \mathfrak{D}} f_2(\mathcal{D})$.

Proof. Since ϕ does not belong to the stabilizer of $S_{|\mathcal{S}|}$, ϕ acts injectively on $S_{|\mathcal{S}|} \cdot \mathcal{D}$. By assumption on ϕ, the image of $\{\mathcal{D}' \in S_{|\mathcal{S}|} \cdot \mathcal{D} \mid f_1(\mathcal{D}') < f_2(\mathcal{D}')\}$ under ϕ is a subset of $\{\mathcal{D}' \in S_{|\mathcal{S}|} \cdot \mathcal{D} \mid f_1(\mathcal{D}') > f_2(\mathcal{D}')\}$. Since ϕ is injective, $\left|\{\mathcal{D}' \in S_{|\mathcal{S}|} \cdot \mathcal{D} \mid f_1(\mathcal{D}') < f_2(\mathcal{D}')\}\right| \leq \left|\{\mathcal{D}' \in S_{|\mathcal{S}|} \cdot \mathcal{D} \mid f_1(\mathcal{D}') > f_2(\mathcal{D}')\}\right|$. $f_1(\mathcal{D}) \geq_{\text{most: } \mathfrak{D}} f_2(\mathcal{D})$ by definition 5.21. ☐

Lemma D.22 (A helper result for expectations of functions)**.** *Let* $B_1, \ldots, B_n \subsetneq \mathbb{R}^{|\mathcal{S}|}$ *be finite and let* $\mathfrak{D} \subseteq \Delta(\mathbb{R}^{|\mathcal{S}|})$*. Suppose* f *is a function of the form*

$$f\left(B_1, \ldots, B_n \mid \mathcal{D}\right) = \mathop{\mathbb{E}}_{\mathbf{r}\sim\mathcal{D}}\left[g\left(\max_{\mathbf{b}_1\in B_1} \mathbf{b}_1^\top \mathbf{r}, \ldots, \max_{\mathbf{b}_n\in B_n} \mathbf{b}_n^\top \mathbf{r}\right)\right] \tag{D.18}$$

for some function g*, and that* f *is well-defined for all* $\mathcal{D} \in \mathfrak{D}$*. Let* ϕ *be a state permutation. Then*

$$f\left(B_1, \ldots, B_n \mid \mathcal{D}\right) = f\left(\phi \cdot B_1, \ldots, \phi \cdot B_n \mid \phi \cdot \mathcal{D}\right). \tag{D.19}$$

Proof. Let distribution $\mathcal{D}$ have probability measure F, and let $\phi \cdot \mathcal{D}$ have probability measure F_ϕ.

$$f\left(B_1, \ldots, B_n \mid \mathcal{D}\right) \tag{D.20}$$

$$:= \mathop{\mathbb{E}}_{\mathbf{r}\sim\mathcal{D}}\left[g\left(\max_{\mathbf{b}_1\in B_1} \mathbf{b}_1^\top \mathbf{r}, \ldots, \max_{\mathbf{b}_n\in B_n} \mathbf{b}_n^\top \mathbf{r}\right)\right] \tag{D.21}$$

$$:= \int_{\mathbb{R}^{|\mathcal{S}|}} g\left(\max_{\mathbf{b}_1\in B_1} \mathbf{b}_1^\top \mathbf{r}, \ldots, \max_{\mathbf{b}_n\in B_n} \mathbf{b}_n^\top \mathbf{r}\right) \mathrm{d}F(\mathbf{r}) \tag{D.22}$$

$$= \int_{\mathbb{R}^{|\mathcal{S}|}} g\left(\max_{\mathbf{b}_1\in B_1} \mathbf{b}_1^\top \mathbf{r}, \ldots, \max_{\mathbf{b}_n\in B_n} \mathbf{b}_n^\top \mathbf{r}\right) \mathrm{d}F_\phi(\mathbf{P}_\phi \mathbf{r}) \tag{D.23}$$

$$= \int_{\mathbb{R}^{|\mathcal{S}|}} g\left(\max_{\mathbf{b}_1\in B_1} \mathbf{b}_1^\top \left(\mathbf{P}_\phi^{-1}\mathbf{r}'\right), \ldots, \max_{\mathbf{b}_n\in B_n} \mathbf{b}_n^\top \left(\mathbf{P}_\phi^{-1}\mathbf{r}'\right)\right) \left|\det \mathbf{P}_\phi\right| \mathrm{d}F_\phi(\mathbf{r}') \tag{D.24}$$

$$= \int_{\mathbb{R}^{|\mathcal{S}|}} g\left(\max_{\mathbf{b}_1\in B_1} \left(\mathbf{P}_\phi \mathbf{b}_1\right)^\top \mathbf{r}', \ldots, \max_{\mathbf{b}_n\in B_n} \left(\mathbf{P}_\phi \mathbf{b}_n\right)^\top \mathbf{r}'\right) \mathrm{d}F_\phi(\mathbf{r}') \tag{D.25}$$

$$= \int_{\mathbb{R}^{|\mathcal{S}|}} g\left(\max_{\mathbf{b}_1'\in \phi\cdot B_1} \mathbf{b}_1'^\top \mathbf{r}', \ldots, \max_{\mathbf{b}_n'\in \phi\cdot B_n} \mathbf{b}_n'^\top \mathbf{r}'\right) \mathrm{d}F_\phi(\mathbf{r}') \tag{D.26}$$

$$=: f\left(\phi \cdot B_1, \ldots, \phi \cdot B_n \mid \phi \cdot \mathcal{D}\right). \tag{D.27}$$

Equation (D.23) follows by the definition of F_ϕ (definition 5.19). Equation (D.24) follows by substituting $\mathbf{r}' := \mathbf{P}_\phi \mathbf{r}$. Equation (D.25) follows from the fact that all permutation matrices have unitary determinant and are orthogonal (and so $(\mathbf{P}_\phi^{-1})^\top = \mathbf{P}_\phi$). $\square$

Definition D.23 (Support of $\mathcal{D}_{\text{any}}$)**.** Let $\mathcal{D}_{\text{any}}$ be any reward function distribution. $\operatorname{supp}(\mathcal{D}_{\text{any}})$ is the smallest closed subset of $\mathbb{R}^{|\mathcal{S}|}$ whose complement has measure zero under $\mathcal{D}_{\text{any}}$.

Definition D.24 (Linear functional optimality probability)**.** For finite $A, B \subsetneq \mathbb{R}^{|\mathcal{S}|}$, the *probability under $\mathcal{D}_{any}$ that A is optimal over B* is

$$p_{\mathcal{D}_{\text{any}}}\left(A \geq B\right) := \mathbb{P}_{\mathbf{r}\sim\mathcal{D}_{\text{any}}}\left(\max_{\mathbf{a}\in A} \mathbf{a}^\top \mathbf{r} \geq \max_{\mathbf{b}\in B} \mathbf{b}^\top \mathbf{r}\right). \tag{D.28}$$

Proposition D.25 (Non-dominated linear functionals and their optimality probability)**.** *Let $A \subsetneq \mathbb{R}^{|\mathcal{S}|}$ be finite. If $\exists b < c : [b, c]^{|\mathcal{S}|} \subseteq \operatorname{supp}(\mathcal{D}_{any})$, then $\mathbf{a} \in \operatorname{ND}\left(A\right)$ implies that $\mathbf{a}$ is strictly optimal for a set of reward functions with positive measure under $\mathcal{D}_{any}$.*

Proof. Suppose $\exists b < c : [b, c]^{|\mathcal{S}|} \subseteq \operatorname{supp}(\mathcal{D}_{\text{any}})$. If $\mathbf{a} \in \operatorname{ND}\left(A\right)$, then let $\mathbf{r}$ be such that $\mathbf{a}^\top \mathbf{r} > \max_{\mathbf{a}' \in A\setminus\{\mathbf{a}\}} \mathbf{a}'^\top \mathbf{r}$. For $a_1 > 0, a_2 \in \mathbb{R}$, positively affinely transform $\mathbf{r}' := a_1 \mathbf{r} + a_2 \mathbf{1}$ (where $\mathbf{1} \in \mathbb{R}^{|\mathcal{S}|}$ is the all-ones vector) so that $\mathbf{r}' \in (b, c)^{|\mathcal{S}|}$.

Note that $\mathbf{a}$ is still strictly optimal for $\mathbf{r}'$:

$$\mathbf{a}^\top \mathbf{r} > \max_{\mathbf{a}' \in A\setminus\{\mathbf{a}\}} \mathbf{a}'^\top \mathbf{r} \iff \mathbf{a}^\top \mathbf{r}' > \max_{\mathbf{a}' \in A\setminus\{\mathbf{a}\}} \mathbf{a}'^\top \mathbf{r}'. \tag{D.29}$$

Furthermore, by the continuity of both terms on the right-hand side of eq. (D.29), $\mathbf{a}$ is strictly optimal for reward functions in some open neighborhood N of $\mathbf{r}'$. Let $N' := N \cap (b, c)^{|\mathcal{S}|}$. N' is still open in $\mathbb{R}^{|\mathcal{S}|}$ since it is the intersection of two open sets N and $(b, c)^{|\mathcal{S}|}$.

$\mathcal{D}_{\text{any}}$ must assign positive probability measure to all open sets in its support; otherwise, its support would exclude these zero-measure sets by definition D.23. Therefore, $\mathcal{D}_{\text{any}}$ assigns positive probability to $N' \subseteq \operatorname{supp}(\mathcal{D}_{\text{any}})$. □

Lemma D.26 (Expected value of similar linear functional sets)**.** *Let $A, B \subsetneq \mathbb{R}^{|\mathcal{S}|}$ be finite, let A' be such that $\mathrm{ND}\,(A) \subseteq A' \subseteq A$, and let $g : \mathbb{R} \to \mathbb{R}$ be an increasing function. If B contains a copy B' of A' via ϕ, then*

$$\mathop{\mathbb{E}}_{\mathbf{r}\sim\mathcal{D}_{bound}}\left[g\left(\max_{\mathbf{a}\in A}\mathbf{a}^\top\mathbf{r}\right)\right] \leq \mathop{\mathbb{E}}_{\mathbf{r}\sim\phi\cdot\mathcal{D}_{bound}}\left[g\left(\max_{\mathbf{b}\in B}\mathbf{b}^\top\mathbf{r}\right)\right]. \tag{D.30}$$

If $\mathrm{ND}\,(B) \setminus B'$ is empty, then eq. (D.30) is an equality. If $\mathrm{ND}\,(B) \setminus B'$ is non-empty, g is strictly increasing, and $\exists b < c : (b,c)^{|\mathcal{S}|} \subseteq \operatorname{supp}(\mathcal{D}_{bound})$, then eq. (D.30) is strict.

Proof. Because $g : \mathbb{R} \to \mathbb{R}$ is increasing, it is measurable (as is max). Therefore, the relevant expectations exist for all $\mathcal{D}_{\text{bound}}$.

$$\begin{aligned}
\mathop{\mathbb{E}}_{\mathbf{r}\sim\mathcal{D}_{\text{bound}}}\left[g\left(\max_{\mathbf{a}\in A}\mathbf{a}^\top\mathbf{r}\right)\right] &= \mathop{\mathbb{E}}_{\mathbf{r}\sim\mathcal{D}_{\text{bound}}}\left[g\left(\max_{\mathbf{a}\in A'}\mathbf{a}^\top\mathbf{r}\right)\right] && \text{(D.31)}\\
&= \mathop{\mathbb{E}}_{\mathbf{r}\sim\phi\cdot\mathcal{D}_{\text{bound}}}\left[g\left(\max_{\mathbf{a}\in\phi\cdot A'}\mathbf{a}^\top\mathbf{r}\right)\right] && \text{(D.32)}\\
&= \mathop{\mathbb{E}}_{\mathbf{r}\sim\phi\cdot\mathcal{D}_{\text{bound}}}\left[g\left(\max_{\mathbf{b}\in B'}\mathbf{b}^\top\mathbf{r}\right)\right] && \text{(D.33)}\\
&\leq \mathop{\mathbb{E}}_{\mathbf{r}\sim\phi\cdot\mathcal{D}_{\text{bound}}}\left[g\left(\max_{\mathbf{b}\in B}\mathbf{b}^\top\mathbf{r}\right)\right]. && \text{(D.34)}
\end{aligned}$$

Equation (D.31) holds because $\forall \mathbf{r} \in \mathbb{R}^{|\mathcal{S}|} : \max_{\mathbf{a}\in A}\mathbf{a}^\top\mathbf{r} = \max_{\mathbf{a}\in A'}\mathbf{a}^\top\mathbf{r}$ by lemma D.17's item 2 with $X := A$, $X' := A'$. Equation (D.32) holds by lemma D.22. Equation (D.33) holds by the definition of B'. Furthermore, our assumption on ϕ guarantees that $B' \subseteq B$. Therefore, $\max_{\mathbf{b}\in B'}\mathbf{b}^\top\mathbf{r} \leq \max_{\mathbf{b}\in B}\mathbf{b}^\top\mathbf{r}$, and so eq. (D.34) holds by the fact that g is an increasing function. Then eq. (D.30) holds.

If $\mathrm{ND}\,(B) \setminus B'$ is empty, then $\mathrm{ND}\,(B) \subseteq B'$. By assumption, $B' \subseteq B$. Then apply lemma D.17 item 2 with $X := B$, $X' := B'$ in order to conclude that eq. (D.34) is an equality. Then eq. (D.30) is also an equality.

Suppose that g is strictly increasing, $\operatorname{ND}(B) \setminus B'$ is non-empty, and $\exists b < c : (b, c)^{|\mathcal{S}|} \subseteq \operatorname{supp}(\mathcal{D}_{\text{bound}})$. Let $\mathbf{x} \in \operatorname{ND}(B) \setminus B'$.

$$\mathop{\mathbb{E}}_{\mathbf{r} \sim \phi \cdot \mathcal{D}_{\text{bound}}}\left[g\left(\max_{\mathbf{b} \in B'} \mathbf{b}^\top \mathbf{r}\right)\right] < \mathop{\mathbb{E}}_{\mathbf{r} \sim \phi \cdot \mathcal{D}_{\text{bound}}}\left[g\left(\max_{\mathbf{a} \in B' \cup \{\mathbf{x}\}} \mathbf{b}^\top \mathbf{r}\right)\right] \tag{D.35}$$

$$\leq \mathop{\mathbb{E}}_{\mathbf{r} \sim \phi \cdot \mathcal{D}_{\text{bound}}}\left[g\left(\max_{\mathbf{b} \in B} \mathbf{b}^\top \mathbf{r}\right)\right]. \tag{D.36}$$

$\mathbf{x}$ is strictly optimal for a positive-probability subset of $\operatorname{supp}(\mathcal{D}_{\text{bound}})$ by proposition D.25. Since g is strictly increasing, eq. (D.35) is strict. Therefore, we conclude that eq. (D.30) is strict. $\square$

Lemma D.27 (Continuous $\mathcal{D}_{X\text{-IID}}$ have nonempty interior)**.** *For continuous* IID *distributions $\mathcal{D}_{X\text{-IID}}$, $\exists b < c : (b, c)^{|\mathcal{S}|} \subseteq \operatorname{supp}(\mathcal{D}_{X\text{-IID}})$.*

Proof. $\mathcal{D}_{X\text{-IID}} := X^{|\mathcal{S}|}$. Since the state reward distribution X is continuous, X must have support on some open interval (b, c). Since $\mathcal{D}_{X\text{-IID}}$ is IID across states, $(b, c)^{|\mathcal{S}|} \subseteq \operatorname{supp}(\mathcal{D}_{X\text{-IID}})$. $\square$

Definition D.28 (Bounded, continuous IID reward)**.** $\mathfrak{D}_{\text{C/B/IID}}$ is the set of $\mathcal{D}_{X\text{-IID}}$ which equal $X^{|\mathcal{S}|}$ for some continuous, bounded-support distribution X over $\mathbb{R}$.

Lemma D.29 (Expectation superiority lemma)**.** *Let $A, B \subsetneq \mathbb{R}^{|\mathcal{S}|}$ be finite and let $g : \mathbb{R} \to \mathbb{R}$ be an increasing function. If B contains a copy B' of* $\operatorname{ND}(A)$ *via ϕ, then*

$$\mathop{\mathbb{E}}_{\mathbf{r} \sim \mathcal{D}_{bound}}\left[g\left(\max_{\mathbf{a} \in A} \mathbf{a}^\top \mathbf{r}\right)\right] \leq_{\text{most: } \mathfrak{D}_{bound}} \mathop{\mathbb{E}}_{\mathbf{r} \sim \mathcal{D}_{bound}}\left[g\left(\max_{\mathbf{b} \in B} \mathbf{b}^\top \mathbf{r}\right)\right]. \tag{D.37}$$

Furthermore, if g is strictly increasing and $\operatorname{ND}(B) \setminus \phi \cdot \operatorname{ND}(A)$ *is non-empty, then* *eq.* (D.37) *is strict for all* $\mathcal{D}_{X\text{-IID}} \in \mathfrak{D}_{\text{C/B/IID}}$. *In particular,*

$$\mathop{\mathbb{E}}_{\mathbf{r}\sim\mathcal{D}_{bound}}\left[g\left(\max_{\mathbf{a}\in A}\mathbf{a}^\top\mathbf{r}\right)\right] \not\geq_{\text{most: }\mathfrak{D}_{bound}} \mathop{\mathbb{E}}_{\mathbf{r}\sim\mathcal{D}_{bound}}\left[g\left(\max_{\mathbf{b}\in B}\mathbf{b}^\top\mathbf{r}\right)\right].$$

Proof. Because $g:\mathbb{R}\to\mathbb{R}$ is increasing, it is measurable (as is max). Therefore, the relevant expectations exist for all $\mathcal{D}_{\text{bound}}$.

Suppose that $\mathcal{D}_{\text{bound}}$ is such that $\mathbb{E}_{\mathbf{r}\sim\mathcal{D}_{\text{bound}}}\left[g\left(\max_{\mathbf{b}\in B}\mathbf{b}^\top\mathbf{r}\right)\right] < \mathbb{E}_{\mathbf{r}\sim\mathcal{D}_{\text{bound}}}\left[g\left(\max_{\mathbf{a}\in A}\mathbf{a}^\top\mathbf{r}\right)\right]$.

$$\begin{aligned}
\mathop{\mathbb{E}}_{\mathbf{r}\sim\phi\cdot\mathcal{D}_{\text{bound}}}\left[g\left(\max_{\mathbf{a}\in A}\mathbf{a}^\top\mathbf{r}\right)\right] &\leq \mathop{\mathbb{E}}_{\mathbf{r}\sim\phi^2\cdot\mathcal{D}_{\text{bound}}}\left[g\left(\max_{\mathbf{b}\in B}\mathbf{b}^\top\mathbf{r}\right)\right] && \text{(D.38)}\\
&= \mathop{\mathbb{E}}_{\mathbf{r}\sim\mathcal{D}_{\text{bound}}}\left[g\left(\max_{\mathbf{b}\in B}\mathbf{b}^\top\mathbf{r}\right)\right] && \text{(D.39)}\\
&< \mathop{\mathbb{E}}_{\mathbf{r}\sim\mathcal{D}_{\text{bound}}}\left[g\left(\max_{\mathbf{a}\in A}\mathbf{a}^\top\mathbf{r}\right)\right] && \text{(D.40)}\\
&\leq \mathop{\mathbb{E}}_{\mathbf{r}\sim\phi\cdot\mathcal{D}_{\text{bound}}}\left[g\left(\max_{\mathbf{b}\in B}\mathbf{b}^\top\mathbf{r}\right)\right]. && \text{(D.41)}
\end{aligned}$$

Equation (D.38) follows by applying lemma D.26 with permutation ϕ and $A' := \text{ND}(A)$. Equation (D.39) follows because involutions satisfy $\phi^{-1}=\phi$, and ϕ^2 is therefore the identity. Equation (D.40) follows because we assumed that

$$\mathop{\mathbb{E}}_{\mathbf{r}\sim\mathcal{D}_{\text{bound}}}\left[g\left(\max_{\mathbf{b}\in B}\mathbf{b}^\top\mathbf{r}\right)\right] < \mathop{\mathbb{E}}_{\mathbf{r}\sim\mathcal{D}_{\text{bound}}}\left[g\left(\max_{\mathbf{a}\in A}\mathbf{a}^\top\mathbf{r}\right)\right].$$

Equation (D.41) follows by applying lemma D.26 with permutation ϕ and and $A' := \text{ND}(A)$. By lemma D.21, eq. (D.37) holds.

Suppose g is strictly increasing and $\text{ND}(B)\setminus B'$ is non-empty. Let $\phi' \in S_{|\mathcal{S}|}$.

$$\mathop{\mathbb{E}}_{\mathbf{r}\sim\phi'\cdot\mathcal{D}_{X\text{-IID}}}\left[g\left(\max_{\mathbf{a}\in A}\mathbf{a}^\top\mathbf{r}\right)\right] = \mathop{\mathbb{E}}_{\mathbf{r}\sim\mathcal{D}_{X\text{-IID}}}\left[g\left(\max_{\mathbf{a}\in A}\mathbf{a}^\top\mathbf{r}\right)\right] \tag{D.42}$$

$$< \mathop{\mathbb{E}}_{\mathbf{r}\sim\phi\cdot\mathcal{D}_{X\text{-IID}}}\left[g\left(\max_{\mathbf{b}\in B}\mathbf{b}^\top\mathbf{r}\right)\right] \tag{D.43}$$

$$= \mathop{\mathbb{E}}_{\mathbf{r}\sim\phi'\cdot\mathcal{D}_{X\text{-IID}}}\left[g\left(\max_{\mathbf{b}\in B}\mathbf{b}^\top\mathbf{r}\right)\right]. \tag{D.44}$$

Equation (D.42) and eq. (D.44) hold because $\mathcal{D}_{X\text{-IID}}$ distributes reward identically across states: $\forall\phi_x \in S_{|\mathcal{S}|} : \phi_x \cdot \mathcal{D}_{X\text{-IID}} = \mathcal{D}_{X\text{-IID}}$. By lemma D.27, $\exists b < c : (b,c)^{|\mathcal{S}|} \subseteq \operatorname{supp}(\mathcal{D}_{X\text{-IID}})$. Therefore, apply lemma D.26 with $A' := \operatorname{ND}(A)$ to conclude that eq. (D.43) holds.

Therefore, $\forall\phi' \in S_{|\mathcal{S}|} : \mathbb{E}_{\mathbf{r}\sim\phi'\cdot\mathcal{D}_{X\text{-IID}}}\left[g\left(\max_{\mathbf{a}\in A}\mathbf{a}^\top\mathbf{r}\right)\right] < \mathbb{E}_{\mathbf{r}\sim\phi'\cdot\mathcal{D}_{X\text{-IID}}}\left[g\left(\max_{\mathbf{b}\in B}\mathbf{b}^\top\mathbf{r}\right)\right]$, and so $\mathbb{E}_{\mathbf{r}\sim\mathcal{D}_{\text{bound}}}\left[g\left(\max_{\mathbf{a}\in A}\mathbf{a}^\top\mathbf{r}\right)\right] \not\geq_{\text{most: }\mathfrak{D}_{\text{bound}}} \mathbb{E}_{\mathbf{r}\sim\mathcal{D}_{\text{bound}}}\left[g\left(\max_{\mathbf{b}\in B}\mathbf{b}^\top\mathbf{r}\right)\right]$ by definition 5.21. □

Definition D.30 (Indicator function)**.** Let L be a predicate which takes input x. $\mathbb{1}_{L(x)}$ is the function which returns 1 when $L(x)$ is true, and 0 otherwise.

Lemma D.31 (Optimality probability inclusion relations)**.** *Let $X, Y \subsetneq \mathbb{R}^{|\mathcal{S}|}$ be finite and suppose $Y' \subseteq Y$.*

$$p_{\mathcal{D}_{any}}(X \geq Y) \leq p_{\mathcal{D}_{any}}\left(X \geq Y'\right) \leq p_{\mathcal{D}_{any}}\left(X \cup \left(Y \setminus Y'\right) \geq Y\right). \tag{D.45}$$

If $\exists b < c : (b,c)^{|\mathcal{S}|} \subseteq \operatorname{supp}(\mathcal{D}_{any})$, $X \subseteq Y$, and $\operatorname{ND}(Y) \cap \left(Y \setminus Y'\right)$ is non-empty, then the second inequality is strict.

Proof.

$$p_{\mathcal{D}_{\text{any}}}(X \geq Y) := \mathop{\mathbb{E}}_{\mathbf{r}\sim\mathcal{D}_{\text{any}}}\left[\mathbb{1}_{\max_{\mathbf{x}\in X}\mathbf{x}^\top\mathbf{r} \geq \max_{\mathbf{y}\in Y}\mathbf{y}^\top\mathbf{r}}\right] \tag{D.46}$$

$$\leq \mathop{\mathbb{E}}_{\mathbf{r}\sim\mathcal{D}_{\text{any}}}\left[\mathbb{1}_{\max_{\mathbf{x}\in X}\mathbf{x}^\top\mathbf{r} \geq \max_{\mathbf{y}\in Y'}\mathbf{y}^\top\mathbf{r}}\right] \tag{D.47}$$

$$\leq \mathop{\mathbb{E}}_{\mathbf{r}\sim\mathcal{D}_{\text{any}}}\left[\mathbb{1}_{\max_{\mathbf{x}\in X\cup(Y\setminus Y')}\mathbf{x}^\top\mathbf{r} \geq \max_{\mathbf{y}\in Y'}\mathbf{y}^\top\mathbf{r}}\right] \tag{D.48}$$

$$= \mathop{\mathbb{E}}_{\mathbf{r}\sim\mathcal{D}_{\text{any}}}\left[\mathbb{1}_{\max_{\mathbf{x}\in X\cup(Y\setminus Y')}\mathbf{x}^\top\mathbf{r} \geq \max_{\mathbf{y}\in Y'\cup(Y\setminus Y')}\mathbf{y}^\top\mathbf{r}}\right] \tag{D.49}$$

$$= \mathop{\mathbb{E}}_{\mathbf{r}\sim\mathcal{D}_{\text{any}}}\left[\mathbb{1}_{\max_{\mathbf{x}\in X\cup(Y\setminus Y')}\mathbf{x}^\top\mathbf{r}\geq\max_{\mathbf{y}\in Y}\mathbf{y}^\top\mathbf{r}}\right] \tag{D.50}$$

$$=: p_{\mathcal{D}_{\text{any}}}\left(X\cup\left(Y\setminus Y'\right)\geq Y\right). \tag{D.51}$$

Equation (D.47) follows because $\forall\mathbf{r}\in\mathbb{R}^{|\mathcal{S}|}:\mathbb{1}_{\max_{\mathbf{x}\in X}\mathbf{x}^\top\mathbf{r}\geq\max_{\mathbf{y}\in Y}\mathbf{y}^\top\mathbf{r}}\leq\mathbb{1}_{\max_{\mathbf{x}\in X}\mathbf{x}^\top\mathbf{r}\geq\max_{\mathbf{y}\in Y'}\mathbf{y}^\top\mathbf{r}}$ since $Y'\subseteq Y$; note that eq. (D.47) equals $p_{\mathcal{D}_{\text{any}}}\left(X\geq Y'\right)$, and so the first inequality of eq. (D.45) is shown. Equation (D.48) holds because

$$\forall\mathbf{r}\in\mathbb{R}^{|\mathcal{S}|}:\mathbb{1}_{\max_{\mathbf{x}\in X}\mathbf{x}^\top\mathbf{r}\geq\max_{\mathbf{y}\in Y'}\mathbf{y}^\top\mathbf{r}}\leq\mathbb{1}_{\max_{\mathbf{x}\in X\cup(Y\setminus Y')}\mathbf{x}^\top\mathbf{r}\geq\max_{\mathbf{y}\in Y'}\mathbf{b}^\top\mathbf{r}}.$$

Suppose $\exists b<c:(b,c)^{|\mathcal{S}|}\subseteq\operatorname{supp}(\mathcal{D}_{\text{any}})$, $X\subseteq Y$, and $\operatorname{ND}(Y)\cap\left(Y\setminus Y'\right)$ is non-empty. Let $\mathbf{y}^*\in\operatorname{ND}(Y)\cap\left(Y\setminus Y'\right)$. By proposition D.25, $\mathbf{y}^*$ is strictly optimal on a subset of $\operatorname{supp}(\mathcal{D}_{\text{any}})$ with positive measure under $\mathcal{D}_{\text{any}}$. In particular, for a set of $\mathbf{r}^*$ with positive measure under $\mathcal{D}_{\text{any}}$, we have $\mathbf{y}^{*\top}\mathbf{r}^*>\max_{\mathbf{y}\in Y'}\mathbf{y}^\top\mathbf{r}^*$.

Then eq. (D.48) is strict, and therefore the second inequality of eq. (D.45) is strict as well. □

Lemma D.32 (Optimality probability of similar linear functional sets)**.** *Let $A,B,C\subsetneq\mathbb{R}^{|\mathcal{S}|}$ be finite, and let $Z\subseteq\mathbb{R}^{|\mathcal{S}|}$ be such that $\operatorname{ND}(C)\subseteq Z\subseteq C$. If $\operatorname{ND}(A)$ is similar to $B'\subseteq B$ via ϕ such that $\phi\cdot\left(Z\setminus\left(B\setminus B'\right)\right)=Z\setminus\left(B\setminus B'\right)$, then*

$$p_{\mathcal{D}_{any}}(A\geq C)\leq p_{\phi\cdot\mathcal{D}_{any}}(B\geq C). \tag{D.52}$$

If $B'=B$, then eq. (D.52) is an equality. If $\exists b<c:(b,c)^{|\mathcal{S}|}\subseteq\operatorname{supp}(\mathcal{D}_{any})$, $B'\subseteq C$, and $\operatorname{ND}(C)\cap\left(B\setminus B'\right)$ is non-empty, then eq. (D.52) is strict.

Proof.

$$p_{\mathcal{D}_{\text{any}}}(A\geq C)=p_{\mathcal{D}_{\text{any}}}(A\geq Z) \tag{D.53}$$

$$=p_{\mathcal{D}_{\text{any}}}\left(\operatorname{ND}(A)\geq Z\right) \tag{D.54}$$

$$\leq p_{\mathcal{D}_{\text{any}}}\left(\operatorname{ND}(A)\geq Z\setminus\left(B\setminus B'\right)\right) \tag{D.55}$$

$$=p_{\phi\cdot\mathcal{D}_{\text{any}}}\left(\phi\cdot\operatorname{ND}(A)\geq\phi\cdot Z\setminus\left(B\setminus B'\right)\right) \tag{D.56}$$

$$= p_{\phi\cdot\mathcal{D}_{\text{any}}}\left(B' \geq Z \setminus \left(B \setminus B'\right)\right) \tag{D.57}$$

$$\leq p_{\phi\cdot\mathcal{D}_{\text{any}}}\left(B' \cup \left(B \setminus B'\right) \geq Z\right) \tag{D.58}$$

$$= p_{\phi\cdot\mathcal{D}_{\text{any}}}\left(B \geq C\right). \tag{D.59}$$

Equation (D.53) and eq. (D.59) follow by lemma D.17's item 2 with $X \coloneqq C$, $X' \coloneqq Z$. Similarly, eq. (D.54) follows by lemma D.17's item 2 with $X \coloneqq A$, $X' \coloneqq \text{ND}\left(A\right)$. Equation (D.55) follows by applying the first inequality of lemma D.31 with $X \coloneqq \text{ND}\left(A\right), Y \coloneqq Z, Y' \coloneqq Z \setminus (B \setminus B')$. Equation (D.56) follows by applying lemma D.22 to eq. (D.53) with permutation ϕ.

Equation (D.57) follows by our assumptions on ϕ. Equation (D.58) follows because by applying the second inequality of lemma D.31 with $X \coloneqq B', Y \coloneqq \text{ND}\left(C\right), Y' \coloneqq \text{ND}\left(C\right) \setminus (B \setminus B')$.

Suppose $B' = B$. Then $B \setminus B' = \emptyset$, and so eq. (D.55) and eq. (D.58) are trivially equalities. Then eq. (D.52) is an equality.

Suppose $\exists b < c : (b, c)^{|\mathcal{S}|} \subseteq \text{supp}(\mathcal{D}_{\text{any}})$; note that $(b, c)^{|\mathcal{S}|} \subseteq \text{supp}(\phi \cdot \mathcal{D}_{\text{any}})$, since such support must be invariant to permutation. Further suppose that $B' \subseteq C$ and that $\text{ND}\left(C\right) \cap \left(B \setminus B'\right)$ is non-empty. Then letting $X \coloneqq B', Y \coloneqq Z, Y' \coloneqq Z \setminus (B \setminus B')$ and noting that $\text{ND}\left(\text{ND}\left(Z\right)\right) = \text{ND}\left(Z\right)$, apply lemma D.31 to eq. (D.58) to conclude that eq. (D.52) is strict. $\square$

Lemma D.33 (Optimality probability superiority lemma)**.** *Let $A, B, C \subsetneq \mathbb{R}^{|\mathcal{S}|}$ be finite, and let Z satisfy $\text{ND}\left(C\right) \subseteq Z \subseteq C$. If B contains a copy B' of $\text{ND}\left(A\right)$ via ϕ such that $\phi \cdot \left(Z \setminus \left(B \setminus B'\right)\right) = Z \setminus \left(B \setminus B'\right)$, then $p_{\mathcal{D}_{any}}\left(A \geq C\right) \leq_{\text{most: } \mathfrak{D}_{any}} p_{\mathcal{D}_{any}}\left(B \geq C\right)$.*

If $B' \subseteq C$ and $\text{ND}\left(C\right) \cap \left(B \setminus B'\right)$ is non-empty, then the inequality is strict for all $\mathcal{D}_{X\text{-IID}} \in \mathfrak{D}_{\text{C/B/IID}}$ and $p_{\mathcal{D}_{any}}\left(A \geq C\right) \not\geq_{\text{most: } \mathfrak{D}_{any}} p_{\mathcal{D}_{any}}\left(B \geq C\right)$.

Proof. Suppose $\mathcal{D}_{\text{any}}$ is such that $p_{\mathcal{D}_{\text{any}}}\left(B \geq C\right) < p_{\mathcal{D}_{\text{any}}}\left(A \geq C\right)$.

$$p_{\phi\cdot\mathcal{D}_{\text{any}}}\left(A \geq C\right) = p_{\phi^{-1}\cdot\mathcal{D}_{\text{any}}}\left(A \geq C\right) \tag{D.60}$$

$$\leq p_{\mathcal{D}_{\text{any}}}\left(B \geq C\right) \tag{D.61}$$

$$< p_{\mathcal{D}_{\text{any}}}(A \geq C) \tag{D.62}$$

$$\leq p_{\phi \cdot \mathcal{D}_{\text{any}}}(B \geq C). \tag{D.63}$$

Equation (D.60) holds because ϕ is an involution. Equation (D.61) and eq. (D.63) hold by applying lemma D.32 with permutation ϕ. Equation (D.62) holds by assumption. Therefore, $p_{\mathcal{D}_{\text{any}}}(A \geq C) \leq_{\text{most: } \mathfrak{D}_{\text{any}}} p_{\mathcal{D}_{\text{any}}}(B \geq C)$ by lemma D.21.

Suppose $B' \subseteq C$ and $\text{ND}(C) \cap \left(B \setminus B'\right)$ is non-empty, and let $\mathcal{D}_{X\text{-IID}}$ be any continuous distribution which distributes reward independently and identically across states. Let $\phi' \in S_{|\mathcal{S}|}$.

$$p_{\phi' \cdot \mathcal{D}_{X\text{-IID}}}(A \geq C) = p_{\mathcal{D}_{X\text{-IID}}}(A \geq C) \tag{D.64}$$

$$< p_{\phi \cdot \mathcal{D}_{X\text{-IID}}}(B \geq C) \tag{D.65}$$

$$= p_{\phi' \cdot \mathcal{D}_{X\text{-IID}}}(A \geq C). \tag{D.66}$$

Equation (D.64) and eq. (D.66) hold because $\mathcal{D}_{X\text{-IID}}$ distributes reward identically across states, $\forall \phi_x \in S_{|\mathcal{S}|} : \phi_x \cdot \mathcal{D}_{X\text{-IID}} = \mathcal{D}_{X\text{-IID}}$. By lemma D.27, $\exists b < c : (b, c)^{|\mathcal{S}|} \subseteq \text{supp}(\mathcal{D}_{X\text{-IID}})$. Therefore, apply lemma D.32 to conclude that eq. (D.65) holds. Therefore, $\forall \phi' \in S_{|\mathcal{S}|} : p_{\phi' \cdot \mathcal{D}_{X\text{-IID}}}(A \geq C) < p_{\phi' \cdot \mathcal{D}_{X\text{-IID}}}(B \geq C)$. In particular, $p_{\mathcal{D}_{\text{any}}}(A \geq C) \not\geq_{\text{most: } \mathfrak{D}_{\text{any}}} p_{\mathcal{D}_{\text{any}}}(B \geq C)$ by definition 5.21. $\square$

Lemma D.34 (Limit probability inequalities which hold for most distributions)**.** *Let $I \subseteq \mathbb{R}$, let $\mathfrak{D} \subseteq \Delta(\mathbb{R}^{|\mathcal{S}|})$ be closed under permutation, and let F_A, F_B, F_C be finite sets of vector functions $I \mapsto \mathbb{R}^{|\mathcal{S}|}$. Let γ be a limit point of I such that*

$$f_1(\mathcal{D}) \coloneqq \lim_{\gamma^* \to \gamma} p_{\mathcal{D}}\left(F_B(\gamma^*) \geq F_C(\gamma^*)\right),$$

$$f_2(\mathcal{D}) \coloneqq \lim_{\gamma^* \to \gamma} p_{\mathcal{D}}\left(F_A(\gamma^*) \geq F_C(\gamma^*)\right)$$

are well-defined for all $\mathcal{D} \in \mathfrak{D}$.

Let F_Z satisfy $\text{ND}(F_C) \subseteq F_Z \subseteq F_C$. Suppose F_B contains a copy of F_A via ϕ such that $\phi \cdot \Big(F_Z \setminus \left(F_B \setminus \phi \cdot F_A\right)\Big) = F_Z \setminus \left(F_B \setminus \phi \cdot F_A\right)$. Then $f_2(\mathfrak{D}) \leq_{\text{most: } \mathfrak{D}} f_1(\mathfrak{D})$.

Proof. Suppose $\mathcal{D} \in \mathfrak{D}$ is such that $f_2(\mathcal{D}) > f_1(\mathcal{D})$.

$$f_2 (\phi \cdot \mathcal{D}) = f_2 \left(\phi^{-1} \cdot \mathcal{D}\right) \tag{D.67}$$

$$\coloneqq \lim_{\gamma^* \to \gamma} p_{\phi^{-1} \cdot \mathcal{D}} \left(F_A(\gamma^*) \geq F_C(\gamma^*)\right) \tag{D.68}$$

$$\leq \lim_{\gamma^* \to \gamma} p_{\mathcal{D}} \left(F_B(\gamma^*) \geq F_C(\gamma^*)\right) \tag{D.69}$$

$$< \lim_{\gamma^* \to \gamma} p_{\mathcal{D}} \left(F_A(\gamma^*) \geq F_C(\gamma^*)\right) \tag{D.70}$$

$$\leq \lim_{\gamma^* \to \gamma} p_{\phi \cdot \mathcal{D}} \left(F_B(\gamma^*) \geq F_C(\gamma^*)\right) \tag{D.71}$$

$$\eqqcolon f_1 (\phi \cdot \mathcal{D}) . \tag{D.72}$$

By the assumption that $\mathfrak{D}$ is closed under permutation and f_2 is well-defined for all $\mathcal{D} \in \mathfrak{D}$, $f_2(\phi \cdot \mathcal{D})$ is well-defined. Equation (D.67) follows since $\phi = \phi^{-1}$ because ϕ is an involution. For all $\gamma^* \in I$, let $A \coloneqq F_A(\gamma^*), B \coloneqq F_B(\gamma^*), C \coloneqq F_C(\gamma^*), Z \coloneqq F_Z(\gamma^*)$ (by definition D.18, $\operatorname{ND}(C) \subseteq Z \subseteq C$). Since $\phi \cdot A \subseteq B$ by assumption, and since $\operatorname{ND}(A) \subseteq A$, B also contains a copy of $\operatorname{ND}(A)$ via ϕ. Furthermore, $\phi \cdot \Big(Z \setminus \big(B \setminus \phi \cdot A\big)\Big) = Z \setminus \big(B \setminus \phi \cdot A\big)$ (by assumption), and so apply lemma D.32 to conclude that $p_{\phi^{-1} \cdot \mathcal{D}} \left(F_A(\gamma^*) \geq F_C(\gamma^*)\right) \leq p_{\mathcal{D}} \left(F_B(\gamma^*) \geq F_C(\gamma^*)\right)$. Therefore, the limit inequality eq. (D.69) holds. Equation (D.70) follows because we assumed that $f_1(\mathcal{D}) < f_2(\mathcal{D})$. Equation (D.71) holds by reasoning similar to that given for eq. (D.69).

Therefore, $f_2(\mathcal{D}) > f_1(\mathcal{D})$ implies that $f_2 (\phi \cdot \mathcal{D}) < f_1 (\phi \cdot \mathcal{D})$, and so apply lemma D.21 to conclude that $f_2(\mathcal{D}) \leq_{\text{most: } \mathfrak{D}} f_1(\mathcal{D})$. □

D.4.1.3 $\mathcal{F}_{\text{nd}}$ results

Proposition D.35 (How to transfer optimal policy sets across discount rates)**.** *Suppose reward function R has optimal policy set $\Pi^* (R, \gamma)$ at discount rate $\gamma \in (0, 1)$. For any $\gamma^* \in (0, 1)$, we can construct a reward function R' such that $\Pi^* \left(R', \gamma^*\right) = \Pi^* (R, \gamma)$. Furthermore, $V^*_{R'} (\cdot, \gamma^*) = V^*_R (\cdot, \gamma)$.*

Proof. Let R be any reward function. Suppose $\gamma^* \in (0, 1)$ and construct $R'(s) \coloneqq V^*_R (s, \gamma) - \gamma^* \max_{a \in \mathcal{A}} \mathbb{E}_{s' \sim T(s,a)} \left[V^*_R (s', \gamma)\right]$.

Let $\pi \in \Pi$ be any policy. By the definition of optimal policies, $\pi \in \Pi^*\left(R', \gamma^*\right)$ iff for all s:

$$R'(s) + \gamma^* \mathop{\mathbb{E}}_{s' \sim T\left(s, \pi(s)\right)}\left[V^*_{R'}\left(s', \gamma^*\right)\right] = R'(s) + \gamma^* \max_{a \in \mathcal{A}} \mathop{\mathbb{E}}_{s' \sim T(s,a)}\left[V^*_{R'}\left(s', \gamma^*\right)\right] \tag{D.73}$$

$$R'(s) + \gamma^* \mathop{\mathbb{E}}_{s' \sim T\left(s, \pi(s)\right)}\left[V^*_{R}\left(s', \gamma\right)\right] = R'(s) + \gamma^* \max_{a \in \mathcal{A}} \mathop{\mathbb{E}}_{s' \sim T(s,a)}\left[V^*_{R}\left(s', \gamma\right)\right] \tag{D.74}$$

$$\gamma^* \mathop{\mathbb{E}}_{s' \sim T\left(s, \pi(s)\right)}\left[V^*_{R}\left(s', \gamma\right)\right] = \gamma^* \max_{a \in \mathcal{A}} \mathop{\mathbb{E}}_{s' \sim T(s,a)}\left[V^*_{R}\left(s', \gamma\right)\right] \tag{D.75}$$

$$\mathop{\mathbb{E}}_{s' \sim T\left(s, \pi(s)\right)}\left[V^*_{R}\left(s', \gamma\right)\right] = \max_{a \in \mathcal{A}} \mathop{\mathbb{E}}_{s' \sim T(s,a)}\left[V^*_{R}\left(s', \gamma\right)\right]. \tag{D.76}$$

By the Bellman equations, $R'(s) = V^*_{R'}\left(s, \gamma^*\right) - \gamma^* \max_{a \in \mathcal{A}} \mathbb{E}_{s' \sim T(s,a)}\left[V^*_{R'}\left(s', \gamma^*\right)\right]$. By the definition of R', $V^*_{R'}\left(\cdot, \gamma^*\right) = V^*_R\left(\cdot, \gamma\right)$ must be the unique solution to the Bellman equations for R' at γ^*. Therefore, eq. (D.74) holds. Equation (D.75) follows by plugging in $R' \coloneqq V^*_R\left(s, \gamma\right) - \gamma^* \max_{a \in \mathcal{A}} \mathbb{E}_{s' \sim T(s,a)}\left[V^*_R\left(s', \gamma\right)\right]$ to eq. (D.74) and doing algebraic manipulation. Equation (D.76) follows because $\gamma^* > 0$.

Equation (D.76) shows that $\pi \in \Pi^*\left(R', \gamma^*\right)$ iff

$$\forall s : \mathop{\mathbb{E}}_{s' \sim T(s, \pi(s))}\left[V^*_{R}\left(s', \gamma\right)\right] = \max_{a \in \mathcal{A}} \mathop{\mathbb{E}}_{s' \sim T(s,a)}\left[V^*_{R}\left(s', \gamma\right)\right].$$

That is, $\pi \in \Pi^*\left(R', \gamma^*\right)$ iff $\pi \in \Pi^*\left(R, \gamma\right)$. □

Definition D.36 (Evaluating sets of visit distribution functions at γ)**.** For $\gamma \in (0,1)$, define $\mathcal{F}(s,\gamma) \coloneqq \left\{\mathbf{f}(\gamma) \mid \mathbf{f} \in \mathcal{F}(s)\right\}$ and $\mathcal{F}_{\text{nd}}(s,\gamma) \coloneqq \left\{\mathbf{f}(\gamma) \mid \mathbf{f} \in \mathcal{F}_{\text{nd}}(s)\right\}$. If $F \subseteq \mathcal{F}(s)$, then $F(\gamma) \coloneqq \left\{\mathbf{f}(\gamma) \mid \mathbf{f} \in F\right\}$.

Lemma D.37 (Non-domination across γ values for mixtures of $\mathbf{f}$)**.** *Let* $\Delta_d \in \Delta\left(\mathbb{R}^{|\mathcal{S}|}\right)$ *be any state distribution and let* $F \coloneqq \left\{\mathbb{E}_{s_d \sim \Delta_d}\left[\mathbf{f}^{\pi, s_d}\right] \mid \pi \in \Pi\right\}$. $\mathbf{f} \in \text{ND}\left(F\right)$ *iff* $\forall \gamma^* \in (0,1)$: $\mathbf{f}(\gamma^*) \in \text{ND}\left(F(\gamma^*)\right)$.

Proof. Let $\mathbf{f}^\pi \in \text{ND}\left(F\right)$ be strictly optimal for reward function R at discount rate

$\gamma \in (0,1)$:

$$\mathbf{f}^{\pi}(\gamma)^{\top}\mathbf{r} > \max_{\mathbf{f}^{\pi'} \in F \setminus \{\mathbf{f}^{\pi}\}} \mathbf{f}^{\pi'}(\gamma)^{\top}\mathbf{r}. \tag{D.77}$$

Let $\gamma^* \in (0,1)$. By proposition D.35, there exists R' such that $\Pi^*\left(R', \gamma^*\right) = \Pi^*\left(R, \gamma\right)$. Since the optimal policy sets are equal, lemma D.6 implies that

$$\mathbf{f}^{\pi}(\gamma^*)^{\top}\mathbf{r}' > \max_{\mathbf{f}^{\pi'} \in F \setminus \{\mathbf{f}^{\pi}\}} \mathbf{f}^{\pi'}(\gamma^*)^{\top}\mathbf{r}'. \tag{D.78}$$

Therefore, $\mathbf{f}^{\pi}(\gamma^*) \in \mathrm{ND}\left(F(\gamma^*)\right)$.

The reverse direction follows by the definition of $\mathrm{ND}\left(F\right)$. $\square$

Lemma D.38 ($\forall \gamma \in (0,1) : \mathbf{d} \in \mathcal{F}_{\mathrm{nd}}(s,\gamma)$ iff $\mathbf{d} \in \mathrm{ND}\left(\mathcal{F}(s,\gamma)\right)$)**.**

Proof. By definition D.36, $\mathcal{F}_{\mathrm{nd}}(s,\gamma) := \left\{\mathbf{f}(\gamma) \mid \mathbf{f} \in \mathrm{ND}\left(\mathcal{F}(s)\right)\right\}$. By applying lemma D.37 with $\Delta_d := \mathbf{e}_s$, $\mathbf{f} \in \mathrm{ND}\left(\mathcal{F}(s)\right)$ iff $\forall \gamma \in (0,1) : \mathbf{f}(\gamma) \in \mathrm{ND}\left(\mathcal{F}(s,\gamma)\right)$. $\square$

Lemma D.39 ($\forall \gamma \in [0,1) : V^*_R\left(s,\gamma\right) = \max_{\mathbf{f} \in \mathcal{F}_{\mathrm{nd}}(s)} \mathbf{f}(\gamma)^{\top}\mathbf{r}$)**.**

Proof. $\mathrm{ND}\left(\mathcal{F}(s,\gamma)\right) = \mathcal{F}_{\mathrm{nd}}(s,\gamma)$ by lemma D.38, so apply corollary D.16 with $X := \mathcal{F}(s,\gamma)$. $\square$

D.4.2 Some actions have greater probability of being optimal

Lemma D.40 (Optimal policy shift bound)**.** *For fixed R, $\Pi^*\left(R,\gamma\right)$ can take on at most $\left(2\left|\mathcal{S}\right|+1\right)\sum_s \binom{\left|\mathcal{F}(s)\right|}{2}$ distinct values over $\gamma \in (0,1)$.*

Proof. By lemma D.6, $\Pi^*\left(R,\gamma\right)$ changes value iff there is a change in optimality status for some visit distribution function at some state. Lippman [48] showed that two visit distribution functions can trade off optimality status at most $2\left|\mathcal{S}\right|+1$ times. At each state s, there are $\binom{\left|\mathcal{F}(s)\right|}{2}$ such pairs. $\square$

Proposition D.41 (Optimality probability's limits exist)**.** *Let $F \subseteq \mathcal{F}(s)$. $\mathbb{P}_{\mathcal{D}_{any}}(F, 0) = \lim_{\gamma\to 0} \mathbb{P}_{\mathcal{D}_{any}}(F, \gamma)$ and $\mathbb{P}_{\mathcal{D}_{any}}(F, 1) = \lim_{\gamma\to 1} \mathbb{P}_{\mathcal{D}_{any}}(F, \gamma)$.*

Proof. First consider the limit as $\gamma \to 1$. Let $\mathcal{D}_{\text{any}}$ have probability measure F_{any}, and define $\delta(\gamma) \coloneqq F_{\text{any}}\left(\left\{R \in \mathbb{R}^{\mathcal{S}} \mid \exists \gamma^* \in [\gamma, 1) : \Pi^*(R, \gamma^*) \neq \Pi^*(R, 1)\right\}\right)$. Since F_{any} is a probability measure, $\delta(\gamma)$ is bounded $[0, 1]$, and $\delta(\gamma)$ is monotone decreasing. Therefore, $\lim_{\gamma\to 1} \delta(\gamma)$ exists.

If $\lim_{\gamma\to 1} \delta(\gamma) > 0$, then there exist reward functions whose optimal policy sets $\Pi^*(R, \gamma)$ never converge (in the discrete topology on sets) to $\Pi^*(R, 1)$, contradicting lemma D.40. So $\lim_{\gamma\to 1} \delta(\gamma) = 0$.

By the definition of optimality probability (definition 5.9) and of $\delta(\gamma)$, $|\,\mathbb{P}_{\mathcal{D}_{\text{any}}}(F, \gamma) - \mathbb{P}_{\mathcal{D}_{\text{any}}}(F, 1)\,| \leq \delta(\gamma)$. Since $\lim_{\gamma\to 1} \delta(\gamma) = 0$, $\lim_{\gamma\to 1} \mathbb{P}_{\mathcal{D}_{\text{any}}}(F, \gamma) = \mathbb{P}_{\mathcal{D}_{\text{any}}}(F, 1)$.

A similar proof shows that $\lim_{\gamma\to 0} \mathbb{P}_{\mathcal{D}_{\text{any}}}(F, \gamma) = \mathbb{P}_{\mathcal{D}_{\text{any}}}(F, 0)$. □

Lemma D.42 (Optimality probability identity)**.** *Let $\gamma \in (0, 1)$ and let $F \subseteq \mathcal{F}(s)$.*

$$\mathbb{P}_{\mathcal{D}_{any}}(F, \gamma) = p_{\mathcal{D}'}\left(F(\gamma) \geq \mathcal{F}(s, \gamma)\right) = p_{\mathcal{D}'}\left(F(\gamma) \geq \mathcal{F}_{\text{nd}}(s, \gamma)\right). \tag{D.79}$$

Proof. Let $\gamma \in (0, 1)$.

$$\mathbb{P}_{\mathcal{D}_{\text{any}}}(F, \gamma) \coloneqq \mathop{\mathbb{P}}_{R\sim\mathcal{D}_{\text{any}}}\left(\exists \mathbf{f}^{\pi} \in F : \pi \in \Pi^*(R, \gamma)\right) \tag{D.80}$$

$$= \mathop{\mathbb{E}}_{\mathbf{r}\sim\mathcal{D}_{\text{any}}}\left[\mathbb{1}_{\max_{\mathbf{f}\in F} \mathbf{f}(\gamma)^\top \mathbf{r} = \max_{\mathbf{f}'\in\mathcal{F}(s)} \mathbf{f}'(\gamma)^\top \mathbf{r}}\right] \tag{D.81}$$

$$= \mathop{\mathbb{E}}_{\mathbf{r}\sim\mathcal{D}_{\text{any}}}\left[\mathbb{1}_{\max_{\mathbf{f}\in F} \mathbf{f}(\gamma)^\top \mathbf{r} = \max_{\mathbf{f}'\in\mathcal{F}_{\text{nd}}(s)} \mathbf{f}'(\gamma)^\top \mathbf{r}}\right] \tag{D.82}$$

$$\eqqcolon p_{\mathcal{D}'}\left(F(\gamma) \geq \mathcal{F}_{\text{nd}}(s, \gamma)\right). \tag{D.83}$$

Equation (D.81) follows because lemma D.6 shows that π is optimal iff it induces an optimal visit distribution $\mathbf{f}$ at every state. Equation (D.82) follows because $\forall \mathbf{r} \in \mathbb{R}^{|\mathcal{S}|}$:

$$\max_{\mathbf{f}'\in\mathcal{F}(s)} \mathbf{f}'(\gamma)^\top \mathbf{r} = \max_{\mathbf{f}'\in\mathcal{F}_{\text{nd}}(s)} \mathbf{f}'(\gamma)^\top \mathbf{r}$$

by lemma D.39. □

D.4.3 Basic properties of Power

Lemma D.43 (Power identities)**.** *Let* $\gamma \in (0,1)$.

$$\textsc{Power}_{\mathcal{D}_{bound}}(s,\gamma) = \mathop{\mathbb{E}}_{\mathbf{r}\sim\mathcal{D}_{bound}}\left[\max_{\mathbf{f}\in\mathcal{F}_{\text{nd}}(s)} \frac{1-\gamma}{\gamma}\left(\mathbf{f}(\gamma)-\mathbf{e}_s\right)^\top \mathbf{r}\right] \tag{D.84}$$

$$= \frac{1-\gamma}{\gamma} \mathop{\mathbb{E}}_{\mathbf{r}\sim\mathcal{D}_{bound}}\left[V^*_R(s,\gamma) - R(s)\right] \tag{D.85}$$

$$= \frac{1-\gamma}{\gamma}\left(V^*_{\mathcal{D}_{bound}}(s,\gamma) - \mathop{\mathbb{E}}_{R\sim\mathcal{D}_{bound}}\left[R(s)\right]\right) \tag{D.86}$$

$$= \mathop{\mathbb{E}}_{R\sim\mathcal{D}_{bound}}\left[\max_{\pi\in\Pi} \mathop{\mathbb{E}}_{s'\sim T\left(s,\pi(s)\right)}\left[(1-\gamma)\, V^\pi_R\left(s',\gamma\right)\right]\right]. \tag{D.87}$$

Proof.

$$\textsc{Power}_{\mathcal{D}_{\text{bound}}}(s,\gamma) := \mathop{\mathbb{E}}_{\mathbf{r}\sim\mathcal{D}_{\text{bound}}}\left[\max_{\mathbf{f}\in\mathcal{F}(s)} \frac{1-\gamma}{\gamma}\left(\mathbf{f}(\gamma)-\mathbf{e}_s\right)^\top \mathbf{r}\right] \tag{D.88}$$

$$= \mathop{\mathbb{E}}_{\mathbf{r}\sim\mathcal{D}_{\text{bound}}}\left[\max_{\mathbf{f}\in\mathcal{F}_{\text{nd}}(s)} \frac{1-\gamma}{\gamma}\left(\mathbf{f}(\gamma)-\mathbf{e}_s\right)^\top \mathbf{r}\right] \tag{D.89}$$

$$= \mathop{\mathbb{E}}_{\mathbf{r}\sim\mathcal{D}_{\text{bound}}}\left[\max_{\mathbf{f}\in\mathcal{F}(s)} \frac{1-\gamma}{\gamma}\left(\mathbf{f}(\gamma)-\mathbf{e}_s\right)^\top \mathbf{r}\right] \tag{D.90}$$

$$= \frac{1-\gamma}{\gamma} \mathop{\mathbb{E}}_{\mathbf{r}\sim\mathcal{D}_{\text{bound}}}\left[V^*_R(s,\gamma) - R(s)\right] \tag{D.91}$$

$$= \frac{1-\gamma}{\gamma}\left(V^*_{\mathcal{D}_{\text{bound}}}(s,\gamma) - \mathop{\mathbb{E}}_{R\sim\mathcal{D}_{\text{bound}}}\left[R(s)\right]\right) \tag{D.92}$$

$$= \mathop{\mathbb{E}}_{\mathbf{r}\sim\mathcal{D}_{\text{bound}}}\left[\max_{\pi\in\Pi} \mathop{\mathbb{E}}_{s'\sim T\left(s,\pi(s)\right)}\left[(1-\gamma)\, \mathbf{f}^{\pi,s'}(\gamma)^\top \mathbf{r}\right]\right] \tag{D.93}$$

$$= \mathop{\mathbb{E}}_{R\sim\mathcal{D}_{\text{bound}}}\left[\max_{\pi\in\Pi} \mathop{\mathbb{E}}_{s'\sim T\left(s,\pi(s)\right)}\left[(1-\gamma)\, V^\pi_R\left(s',\gamma\right)\right]\right]. \tag{D.94}$$

Equation (D.89) follows from lemma D.39. Equation (D.91) follows from the dual formulation of optimal value functions. Equation (D.92) holds by the definition of $V^*_{\mathcal{D}_{\text{bound}}}(s, \gamma)$ (definition 5.11). Equation (D.93) holds because $\mathbf{f}^{\pi,s}(\gamma) = \mathbf{e}_s + \gamma\, \mathbb{E}_{s' \sim T\left(s,\pi(s)\right)}\left[\mathbf{f}^{\pi,s'}(\gamma)\right]$ by the definition of a visit distribution function (definition 5.3). ☐

Definition D.44 (Discount-normalized value function)**.** Let π be a policy, R a reward function, and s a state. For $\gamma \in [0,1]$, $V^\pi_{R,\,\text{norm}}(s,\gamma) \coloneqq \lim_{\gamma^* \to \gamma}(1-\gamma^*)V^\pi_R(s,\gamma^*)$.

Lemma D.45 (Normalized value functions have uniformly bounded derivative)**.** *There exists $K \geq 0$ such that for all reward functions $\mathbf{r} \in \mathbb{R}^{|\mathcal{S}|}$, $\sup_{s\in\mathcal{S},\pi\in\Pi,\gamma\in[0,1]}\left|\frac{d}{d\gamma}V^\pi_{R,\,norm}(s,\gamma)\right| \leq K\left\|\mathbf{r}\right\|_1$.*

Proof. Let π be any policy, s a state, and R a reward function. Since

$$V^\pi_{R,\,\text{norm}}(s,\gamma) = \lim_{\gamma^*\to\gamma}(1-\gamma^*)\mathbf{f}^{\pi,s}(\gamma^*)^\top\mathbf{r},$$

$\frac{d}{d\gamma}V^\pi_{R,\,\text{norm}}(s,\gamma)$ is controlled by the behavior of $\lim_{\gamma^*\to\gamma}(1-\gamma^*)\mathbf{f}^{\pi,s}(\gamma^*)$. We show that this function's gradient is bounded in infinity norm.

By lemma D.9, $\mathbf{f}^{\pi,s}(\gamma)$ is a multivariate rational function on γ. Therefore, for any state s', $\mathbf{f}^{\pi,s}(\gamma)^\top\mathbf{e}_{s'} = \frac{P(\gamma)}{Q(\gamma)}$ in reduced form. By proposition D.8, $0 \leq \mathbf{f}^{\pi,s}(\gamma)^\top\mathbf{e}_{s'} \leq \frac{1}{1-\gamma}$. Thus, Q may only have a root of multiplicity 1 at $\gamma = 1$, and $Q(\gamma) \neq 0$ for $\gamma \in [0,1)$. Let $f_{s'}(\gamma) \coloneqq (1-\gamma)\mathbf{f}^{\pi,s}(\gamma)^\top\mathbf{e}_{s'}$.

If $Q(1) \neq 0$, then the derivative $f'_{s'}(\gamma)$ is bounded on $\gamma \in [0,1)$ because the polynomial $(1-\gamma)P(\gamma)$ cannot diverge on a bounded domain.

If $Q(1) = 0$, then factor out the root as $Q(\gamma) = (1-\gamma)Q^*(\gamma)$.

$$f'_{s'}(\gamma) = \frac{d}{d\gamma}\left(\frac{(1-\gamma)P(\gamma)}{Q(\gamma)}\right) \tag{D.95}$$
$$= \frac{d}{d\gamma}\left(\frac{P(\gamma)}{Q^*(\gamma)}\right) \tag{D.96}$$
$$= \frac{P'(\gamma)Q^*(\gamma) - (Q^*)'(\gamma)P(\gamma)}{(Q^*(\gamma))^2}. \tag{D.97}$$

Since $Q^*(\gamma)$ is a polynomial with no roots on $\gamma \in [0,1]$, $f'_{s'}(\gamma)$ is bounded on $\gamma \in [0,1)$.

Therefore, whether or not $Q(\gamma)$ has a root at $\gamma = 1$, $f'_{s'}(\gamma)$ is bounded on $\gamma \in [0,1)$. Furthermore, $\sup_{\gamma\in[0,1)} \left\|\nabla(1-\gamma)\mathbf{f}^{\pi,s}(\gamma)\right\|_\infty = \sup_{\gamma\in[0,1)} \max_{s'\in\mathcal{S}} \left|f'_{s'}(\gamma)\right|$ is finite since there are only finitely many states.

There are finitely many $\pi \in \Pi$, and finitely many states s, and so there exists some K' such that $\sup_{\substack{s\in\mathcal{S},\\ \pi\in\Pi,\gamma\in[0,1)}} \left\|\nabla(1-\gamma)\mathbf{f}^{\pi,s}(\gamma)\right\|_\infty \leq K'$. Then $\left\|\nabla(1-\gamma)\mathbf{f}^{\pi,s}(\gamma)\right\|_1 \leq |\mathcal{S}|\, K' =: K$.

$$\sup_{\substack{s\in\mathcal{S},\\ \pi\in\Pi,\gamma\in[0,1)}} \left|\frac{d}{d\gamma} V^\pi_{R,\text{norm}}(s,\gamma)\right| := \sup_{\substack{s\in\mathcal{S},\\ \pi\in\Pi,\gamma\in[0,1)}} \left|\frac{d}{d\gamma} \lim_{\gamma^*\to\gamma}(1-\gamma^*)V^\pi_R\left(s,\gamma^*\right)\right| \tag{D.98}$$

$$= \sup_{\substack{s\in\mathcal{S},\\ \pi\in\Pi,\gamma\in[0,1)}} \left|\frac{d}{d\gamma}(1-\gamma)V^\pi_R(s,\gamma)\right| \tag{D.99}$$

$$= \sup_{\substack{s\in\mathcal{S},\\ \pi\in\Pi,\gamma\in[0,1)}} \left|\nabla(1-\gamma)\mathbf{f}^{\pi,s}(\gamma)^\top\mathbf{r}\right| \tag{D.100}$$

$$\leq \sup_{\substack{s\in\mathcal{S},\\ \pi\in\Pi,\gamma\in[0,1)}} \left\|\nabla(1-\gamma)\mathbf{f}^{\pi,s}(\gamma)\right\|_1 \|\mathbf{r}\|_1 \tag{D.101}$$

$$\leq K\|\mathbf{r}\|_1. \tag{D.102}$$

Equation (D.99) holds because $V^\pi_R(s,\gamma)$ is continuous on $\gamma \in [0,1)$ by corollary D.10. Equation (D.101) holds by the Cauchy-Schwarz inequality.

Since $\left|\frac{d}{d\gamma}V^\pi_{R,\text{norm}}(s,\gamma)\right|$ is bounded for all $\gamma \in [0,1)$, eq. (D.102) also holds for $\gamma \to 1$. □

Lemma 5.13 (Continuity of Power)**.** $\text{Power}_{\mathcal{D}_{bound}}(s,\gamma)$ *is Lipschitz continuous on* $\gamma \in [0,1]$.

Proof. Let b, c be such that $\text{supp}(\mathcal{D}_{\text{bound}}) \subseteq [b,c]^{|\mathcal{S}|}$. For any $\mathbf{r} \in \text{supp}(\mathcal{D}_{\text{bound}})$ and $\pi \in \Pi$, $V^\pi_{R,\,\text{norm}}(s,\gamma)$ has Lipschitz constant $K\|\mathbf{r}\|_1 \leq K|\mathcal{S}|\,\|\mathbf{r}\|_\infty \leq K|\mathcal{S}|\max(|c|,|b|)$ on $\gamma \in (0,1)$ by lemma D.45.

For $\gamma \in (0,1)$, $\text{Power}_{\mathcal{D}_{\text{bound}}}(s,\gamma) = \mathbb{E}_{R\sim\mathcal{D}_{\text{bound}}}\left[\max_{\pi\in\Pi}\mathbb{E}_{s'\sim T\left(s,\pi(s)\right)}\left[(1-\gamma)V^\pi_R\left(s',\gamma\right)\right]\right]$

by eq. (D.94). The expectation of the maximum of a set of functions which share a Lipschitz constant, also shares the Lipschitz constant. This shows that $\text{POWER}_{\mathcal{D}_{\text{bound}}}(s,\gamma)$ is Lipschitz continuous on $\gamma \in (0,1)$. Thus, its limits are well-defined as $\gamma \to 0$ and $\gamma \to 1$. So it is Lipschitz continuous on the closed unit interval. □

Proposition 5.14 (Maximal POWER)**.** $\text{POWER}_{\mathcal{D}_{bound}}(s,\gamma) \leq \mathbb{E}_{R\sim\mathcal{D}_{bound}}\left[\max_{s\in\mathcal{S}} R(s)\right]$, *with equality if s can deterministically reach all states in one step and all states are 1-cycles.*

Proof. Let $\gamma \in (0,1)$.

$$\text{POWER}_{\mathcal{D}_{\text{bound}}}(s,\gamma) = \mathop{\mathbb{E}}_{R\sim\mathcal{D}_{\text{bound}}}\left[\max_{\pi\in\Pi} \mathop{\mathbb{E}}_{s'\sim T(s,\pi(s))}\left[(1-\gamma)V_R^*\left(s',\gamma\right)\right]\right] \tag{D.103}$$

$$\leq \mathop{\mathbb{E}}_{R\sim\mathcal{D}_{\text{bound}}}\left[\max_{\pi\in\Pi} \mathop{\mathbb{E}}_{s'\sim T(s,\pi(s))}\left[(1-\gamma)\frac{\max_{s''\in\mathcal{S}} R(s'')}{1-\gamma}\right]\right] \tag{D.104}$$

$$= \mathop{\mathbb{E}}_{R\sim\mathcal{D}_{\text{bound}}}\left[\max_{s''\in\mathcal{S}} R(s'')\right]. \tag{D.105}$$

Equation (D.103) follows from lemma D.43. Equation (D.104) follows because $V_R^*\left(s',\gamma\right) \leq \frac{\max_{s''\in\mathcal{S}} R(s'')}{1-\gamma}$, as no policy can do better than achieving maximal reward at each time step. Taking limits, the inequality holds for all $\gamma \in [0,1]$.

Suppose that s can deterministically reach all states in one step and all states are 1-cycles. Then eq. (D.104) is an equality for all $\gamma \in (0,1)$, since for each R, the agent can select an action which deterministically transitions to a state with maximal reward. Thus the equality holds for all $\gamma \in [0,1]$. □

Lemma D.46 (Lower bound on current POWER based on future POWER)**.**

$$\text{POWER}_{\mathcal{D}_{bound}}(s,\gamma) \geq (1-\gamma)\min_a \mathop{\mathbb{E}}_{\substack{s'\sim T(s,a),\\ R\sim\mathcal{D}_{bound}}}\left[R(s')\right] + \gamma \max_a \mathop{\mathbb{E}}_{s'\sim T(s,a)}\left[\text{POWER}_{\mathcal{D}_{bound}}\left(s',\gamma\right)\right]. \tag{D.106}$$

Proof. Let $\gamma \in (0,1)$ and let $a^* \in \arg\max_a \mathbb{E}_{s' \sim T(s,a)}\left[\textsc{Power}_{\mathcal{D}_{\text{bound}}}\left(s',\gamma\right)\right]$.

$$\begin{aligned}
&\textsc{Power}_{\mathcal{D}_{\text{bound}}}(s,\gamma) && \text{(D.107)}\\
&= (1-\gamma) \mathop{\mathbb{E}}_{R\sim\mathcal{D}_{\text{bound}}}\left[\max_a \mathop{\mathbb{E}}_{s'\sim T(s,a)}\left[V_R^*\left(s',\gamma\right)\right]\right] && \text{(D.108)}\\
&\geq (1-\gamma) \max_a \mathop{\mathbb{E}}_{s'\sim T(s,a)}\left[\mathop{\mathbb{E}}_{R\sim\mathcal{D}_{\text{bound}}}\left[V_R^*\left(s',\gamma\right)\right]\right] && \text{(D.109)}\\
&= (1-\gamma) \max_a \mathop{\mathbb{E}}_{s'\sim T(s,a)}\left[V_{\mathcal{D}_{\text{bound}}}^*\left(s',\gamma\right)\right] && \text{(D.110)}\\
&= (1-\gamma) \max_a \mathop{\mathbb{E}}_{s'\sim T(s,a)}\left[\mathop{\mathbb{E}}_{R\sim\mathcal{D}_{\text{bound}}}\left[R(s')\right] + \frac{\gamma}{1-\gamma}\textsc{Power}_{\mathcal{D}_{\text{bound}}}\left(s',\gamma\right)\right] && \text{(D.111)}\\
&\geq (1-\gamma) \mathop{\mathbb{E}}_{s'\sim T(s,a^*)}\left[\mathop{\mathbb{E}}_{R\sim\mathcal{D}_{\text{bound}}}\left[R(s')\right] + \frac{\gamma}{1-\gamma}\textsc{Power}_{\mathcal{D}_{\text{bound}}}\left(s',\gamma\right)\right] && \text{(D.112)}\\
&\geq (1-\gamma) \min_a \mathop{\mathbb{E}}_{\substack{s'\sim T(s,a),\\ R\sim\mathcal{D}_{\text{bound}}}}\left[R(s')\right] + \gamma \mathop{\mathbb{E}}_{s'\sim T(s,a^*)}\left[\textsc{Power}_{\mathcal{D}_{\text{bound}}}\left(s',\gamma\right)\right]. && \text{(D.113)}
\end{aligned}$$

Equation (D.108) holds by lemma D.43. Equation (D.109) follows because

$$\mathop{\mathbb{E}}_{x\sim X}\left[\max_a f(a,x)\right] \geq \max_a \mathop{\mathbb{E}}_{x\sim X}\left[f(a,x)\right]$$

by Jensen's inequality, and eq. (D.111) follows by lemma D.43.

The inequality also holds when we take the limits $\gamma \to 0$ or $\gamma \to 1$. □

Proposition 5.15 (Power is smooth across reversible dynamics)**.** *Let $\mathcal{D}_{bound}$ be bounded $[b,c]$. Suppose s and s' can both reach each other in one step with probability* 1.

$$\left|\textsc{Power}_{\mathcal{D}_{bound}}(s,\gamma) - \textsc{Power}_{\mathcal{D}_{bound}}\left(s',\gamma\right)\right| \leq (c-b)(1-\gamma). \tag{5.7}$$

Proof. Let $\gamma \in [0,1]$. First consider the case where $\textsc{Power}_{\mathcal{D}_{\text{bound}}}(s,\gamma) \geq \textsc{Power}_{\mathcal{D}_{\text{bound}}}\left(s',\gamma\right)$.

$$\textsc{Power}_{\mathcal{D}_{\text{bound}}}\left(s',\gamma\right) \geq (1-\gamma)\min_a \mathop{\mathbb{E}}_{\substack{s_x\sim T(s',a),\\ R\sim\mathcal{D}_{\text{bound}}}}\left[R(s_x)\right] + \gamma \max_a \mathop{\mathbb{E}}_{s_x\sim T(s',a)}\left[\textsc{Power}_{\mathcal{D}_{\text{bound}}}(s_x,\gamma)\right] \tag{D.114}$$

$$\geq (1-\gamma)b + \gamma \text{POWER}_{\mathcal{D}_{\text{bound}}}(s, \gamma). \tag{D.115}$$

Equation (D.114) follows by lemma D.46. Equation (D.115) follows because reward is lower-bounded by b and because s' can reach s in one step with probability 1.

$$\left| \text{POWER}_{\mathcal{D}_{\text{bound}}}(s, \gamma) - \text{POWER}_{\mathcal{D}_{\text{bound}}}(s', \gamma) \right| \tag{D.116}$$

$$= \text{POWER}_{\mathcal{D}_{\text{bound}}}(s, \gamma) - \text{POWER}_{\mathcal{D}_{\text{bound}}}(s', \gamma) \tag{D.117}$$

$$\leq \text{POWER}_{\mathcal{D}_{\text{bound}}}(s, \gamma) - \big((1-\gamma)b + \gamma \text{POWER}_{\mathcal{D}_{\text{bound}}}(s, \gamma)\big) \tag{D.118}$$

$$= (1-\gamma)\left(\text{POWER}_{\mathcal{D}_{\text{bound}}}(s, \gamma) - b\right) \tag{D.119}$$

$$\leq (1-\gamma)\left(\mathop{\mathbb{E}}_{R \sim \mathcal{D}_{\text{bound}}}\left[\max_{s'' \in \mathcal{S}} R(s'')\right] - b\right) \tag{D.120}$$

$$\leq (1-\gamma)(c-b). \tag{D.121}$$

Equation (D.117) follows because

$$\text{POWER}_{\mathcal{D}_{\text{bound}}}(s, \gamma) \geq \text{POWER}_{\mathcal{D}_{\text{bound}}}(s', \gamma).$$

Equation (D.118) follows by eq. (D.115). Equation (D.120) follows by proposition 5.14. Equation (D.121) follows because reward under $\mathcal{D}_{\text{bound}}$ is upper-bounded by c.

The case where $\text{POWER}_{\mathcal{D}_{\text{bound}}}(s, \gamma) \leq \text{POWER}_{\mathcal{D}_{\text{bound}}}(s', \gamma)$ is similar, leveraging the fact that s can also reach s' in one step with probability 1. $\square$

D.4.4 Seeking POWER is often more probable under optimality

D.4.4.1 Keeping options open tends to be POWER-seeking and tends to be optimal

Definition D.47 (Normalized visit distribution function)**.** Let $\mathbf{f} : [0,1) \to \mathbb{R}^{|\mathcal{S}|}$ be a vector function. For $\gamma \in [0,1]$, $\text{NORM}(\mathbf{f}, \gamma) := \lim_{\gamma^* \to \gamma}(1-\gamma^*)\mathbf{f}(\gamma^*)$ (this limit need not exist for arbitrary $\mathbf{f}$). If F is a set of such $\mathbf{f}$, then $\text{NORM}(F, \gamma) := \left\{\text{NORM}(\mathbf{f}, \gamma) \mid \mathbf{f} \in F\right\}$.

Remark. $\text{RSD}(s) = \textsc{Norm}\left(\mathcal{F}(s), 1\right)$.

Lemma D.48 (Normalized visit distribution functions are continuous)**.** *Let $\Delta_s \in \Delta(\mathcal{S})$ be a state probability distribution, let $\pi \in \Pi$, and let $\mathbf{f}^* := \mathbb{E}_{s\sim\Delta_s}\left[\mathbf{f}^{\pi,s}\right]$. $\textsc{Norm}\left(\mathbf{f}^*, \gamma\right)$ is continuous on $\gamma \in [0,1]$.*

Proof.

$$\textsc{Norm}\left(\mathbf{f}^*, \gamma\right) := \lim_{\gamma^*\to\gamma}(1-\gamma^*)\mathop{\mathbb{E}}_{s\sim\Delta_s}\left[\mathbf{f}^{\pi,s}(\gamma^*)\right] \tag{D.122}$$

$$= \mathop{\mathbb{E}}_{s\sim\Delta_s}\left[\lim_{\gamma^*\to\gamma}(1-\gamma^*)\mathbf{f}^{\pi,s}(\gamma^*)\right] \tag{D.123}$$

$$=: \mathop{\mathbb{E}}_{s\sim\Delta_s}\left[\textsc{Norm}\left(\mathbf{f}^{\pi,s}, \gamma\right)\right]. \tag{D.124}$$

Equation (D.123) follows because the expectation is over a finite set. Each $\mathbf{f}^{\pi,s} \in \mathcal{F}(s)$ is continuous on $\gamma \in [0,1)$ by lemma D.9, and $\lim_{\gamma^*\to 1}(1-\gamma^*)\mathbf{f}^{\pi,s}(\gamma^*)$ exists because RSDs are well-defined [68]. Therefore, each $\textsc{Norm}\left(\mathbf{f}^{\pi,s}, \gamma\right)$ is continuous on $\gamma \in [0,1]$. Lastly, eq. (D.124)'s expectation over finitely many continuous functions is itself continuous. □

Lemma D.49 (Non-domination of normalized visit distribution functions)**.** *Let $\Delta_s \in \Delta(\mathcal{S})$ be a state probability distribution and let $F := \left\{\mathbb{E}_{s\sim\Delta_s}\left[\mathbf{f}^{\pi,s}\right] \mid \pi \in \Pi\right\}$. For all $\gamma \in [0,1]$, $\text{ND}\left(\textsc{Norm}\left(F, \gamma\right)\right) \subseteq \textsc{Norm}\left(\text{ND}\left(F\right), \gamma\right)$, with equality when $\gamma \in (0,1)$.*

Proof. Suppose $\gamma \in (0,1)$.

$$\text{ND}\left(\textsc{Norm}\left(F, \gamma\right)\right) = \text{ND}\left((1-\gamma)F(\gamma)\right) \tag{D.125}$$

$$= (1-\gamma)\text{ND}\left(F(\gamma)\right) \tag{D.126}$$

$$= (1-\gamma)\left(\text{ND}\left(F\right)(\gamma)\right) \tag{D.127}$$

$$= \textsc{Norm}\left(\text{ND}\left(F\right), \gamma\right). \tag{D.128}$$

Equation (D.125) and eq. (D.128) follow by the continuity of $\textsc{Norm}\left(\mathbf{f}, \gamma\right)$ (lemma D.48). Equation (D.126) follows by lemma D.20 item 1. Equation (D.127) follows by lemma D.37.

Let $\gamma = 1$. Let $\mathbf{d} \in \text{ND}\left(\textsc{Norm}\left(F, 1\right)\right)$ be strictly optimal for $\mathbf{r}^* \in \mathbb{R}^{|\mathcal{S}|}$. Then let $F_{\mathbf{d}} \subseteq F$

be the subset of $\mathbf{f} \in F$ such that $\textsc{Norm}(\mathbf{f}, 1) = \mathbf{d}$.

$$\max_{\mathbf{f} \in F_{\mathbf{d}}} \textsc{Norm}(\mathbf{f}, 1)^\top \mathbf{r}^* > \max_{\mathbf{f}' \in F \setminus F_{\mathbf{d}}} \textsc{Norm}\left(\mathbf{f}', 1\right)^\top \mathbf{r}^*. \tag{D.129}$$

Since $\textsc{Norm}(\mathbf{f}, 1)$ is continuous at $\gamma = 1$ (lemma D.48), $\mathbf{x}^\top \mathbf{r}^*$ is continuous on $\mathbf{x} \in \mathbb{R}^{|\mathcal{S}|}$, and F is finite, eq. (D.129) holds for some $\gamma^* \in (0, 1)$ sufficiently close to $\gamma = 1$. By lemma D.15, at least one $\mathbf{f} \in F_{\mathbf{d}}$ is an element of $\text{ND}\left(F(\gamma^*)\right)$. Then by lemma D.37, $\mathbf{f} \in \text{ND}(F)$. We conclude that $\text{ND}\left(\textsc{Norm}(F, 1)\right) \subseteq \textsc{Norm}\left(\text{ND}(F), 1\right)$.

The case for $\gamma = 0$ proceeds similarly. $\square$

Lemma D.50 ($\textsc{Power}$ limit identity)**.** *Let* $\gamma \in [0, 1]$.

$$\textsc{Power}_{\mathcal{D}_{bound}}(s, \gamma) = \mathbb{E}_{\mathbf{r} \sim \mathcal{D}_{bound}}\left[\max_{\mathbf{f} \in \mathcal{F}_{\text{nd}}(s)} \lim_{\gamma^* \to \gamma} \frac{1 - \gamma^*}{\gamma^*}\left(\mathbf{f}(\gamma^*) - \mathbf{e}_s\right)^\top \mathbf{r}\right]. \tag{D.130}$$

Proof. Let $\gamma \in [0, 1]$.

$$\begin{aligned}
\textsc{Power}_{\mathcal{D}_{\text{bound}}}(s, \gamma) &= \lim_{\gamma^* \to \gamma} \textsc{Power}_{\mathcal{D}_{\text{bound}}}\left(s, \gamma^*\right) && \text{(D.131)}\\
&= \lim_{\gamma^* \to \gamma} \mathbb{E}_{\mathbf{r} \sim \mathcal{D}_{\text{bound}}}\left[\max_{\mathbf{f} \in \mathcal{F}_{\text{nd}}(s)} \frac{1 - \gamma^*}{\gamma^*}\left(\mathbf{f}(\gamma^*) - \mathbf{e}_s\right)^\top \mathbf{r}\right] && \text{(D.132)}\\
&= \mathbb{E}_{\mathbf{r} \sim \mathcal{D}_{\text{bound}}}\left[\lim_{\gamma^* \to \gamma} \max_{\mathbf{f} \in \mathcal{F}_{\text{nd}}(s)} \frac{1 - \gamma^*}{\gamma^*}\left(\mathbf{f}(\gamma^*) - \mathbf{e}_s\right)^\top \mathbf{r}\right] && \text{(D.133)}\\
&= \mathbb{E}_{\mathbf{r} \sim \mathcal{D}_{\text{bound}}}\left[\max_{\mathbf{f} \in \mathcal{F}_{\text{nd}}(s)} \lim_{\gamma^* \to \gamma} \frac{1 - \gamma^*}{\gamma^*}\left(\mathbf{f}(\gamma^*) - \mathbf{e}_s\right)^\top \mathbf{r}\right]. && \text{(D.134)}
\end{aligned}$$

Equation (D.131) follows since $\textsc{Power}_{\mathcal{D}_{\text{bound}}}(s, \gamma)$ is continuous on $\gamma \in [0, 1]$ by lemma 5.13. Equation (D.132) follows by lemma D.43.

For $\gamma^* \in (0, 1)$, let $f_{\gamma^*}(\mathbf{r}) := \max_{\mathbf{f} \in \mathcal{F}_{\text{nd}}(s)} \frac{1-\gamma^*}{\gamma^*}\left(\mathbf{f}(\gamma^*) - \mathbf{e}_s\right)^\top \mathbf{r}$. For any sequence $\gamma_n \to \gamma$, $\left(f_{\gamma_n}\right)_{n=1}^\infty$ is a sequence of functions which are piecewise linear on $\mathbf{r} \in \mathbb{R}^{|\mathcal{S}|}$, which means they are continuous and therefore measurable. Since lemma D.9 shows that each $\mathbf{f} \in \mathcal{F}_{\text{nd}}(s)$ is multivariate rational on γ^* (and therefore continuous on γ^*), $\left\{f_{\gamma_n}\right\}_{n=1}^\infty$ converges pointwise to limit function f_γ. Furthermore, $\left|V^*_R(s,\gamma_n)-R(s)\right| \leq \frac{\gamma}{1-\gamma_n}\|R\|_\infty$, and so $\left|f_{\gamma_n}(\mathbf{r})\right| = \left|\frac{1-\gamma_n}{\gamma_n}(V^*_R(s,\gamma_n)-R(s))\right| \leq g(\mathbf{r}) \leq \|\mathbf{r}\|_\infty =: g(\mathbf{r})$, which is measurable. Therefore, apply Lebesgue's dominated convergence theorem to conclude that eq. (D.133) holds. Equation (D.134) holds because max is a continuous function. □

Lemma D.51 (Lemma for Power superiority)**.** *Let $\Delta_1, \Delta_2 \in \Delta(\mathcal{S})$ be state probability distributions. For $i = 1,2$, let $F_{\Delta_i} := \left\{\gamma^{-1}\,\mathbb{E}_{s_i\sim\Delta_i}\left[\mathbf{f}^{\pi,s_i}-\mathbf{e}_{s_i}\right] \mid \pi \in \Pi\right\}$. Suppose F_{Δ_2} contains a copy of $\operatorname{ND}\left(F_{\Delta_1}\right)$ via ϕ. Then $\forall\gamma\in[0,1]: \mathbb{E}_{s_1\sim\Delta_1}\left[\textsc{Power}_{\mathcal{D}_{bound}}(s_1,\gamma)\right] \leq_{\text{most: } \mathfrak{D}_{bound}} \mathbb{E}_{s_2\sim\Delta_2}\left[\textsc{Power}_{\mathcal{D}_{bound}}(s_2,\gamma)\right]$.*

If $\operatorname{ND}\left(F_{\Delta_2}\right)\setminus\phi\cdot\operatorname{ND}\left(F_{\Delta_1}\right)$ is non-empty, then for all $\gamma\in(0,1)$, the inequality is strict for all $\mathcal{D}_{X\text{-IID}} \in \mathfrak{D}_{\text{C/B/IID}}$ and $\mathbb{E}_{s_1\sim\Delta_1}\left[\textsc{Power}_{\mathcal{D}_{bound}}(s_1,\gamma)\right] \not\geq_{\text{most: } \mathfrak{D}_{bound}} \mathbb{E}_{s_2\sim\Delta_2}\left[\textsc{Power}_{\mathcal{D}_{bound}}(s_2,\gamma)\right]$.

These results also hold when replacing F_{Δ_i} with $F^*_{\Delta_i} := \left\{\mathbb{E}_{s_i\sim\Delta_i}\left[\mathbf{f}^{\pi,s_i}\right] \mid \pi\in\Pi\right\}$ for $i=1,2$.

Proof.

$$
\begin{aligned}
\phi\cdot\operatorname{ND}\left(\textsc{Norm}\left(F_{\Delta_1},\gamma\right)\right) &\subseteq \phi\cdot\textsc{Norm}\left(\operatorname{ND}\left(F_{\Delta_1}\right),\gamma\right) && \text{(D.135)}\\
&:= \left\{\mathbf{P}_\phi \lim_{\gamma^*\to\gamma}(1-\gamma^*)\mathbf{f}(\gamma^*) \mid \mathbf{f}\in\operatorname{ND}\left(F_{\Delta_1}\right)\right\} && \text{(D.136)}\\
&= \left\{\lim_{\gamma^*\to\gamma}(1-\gamma^*)\mathbf{P}_\phi\mathbf{f}(\gamma^*) \mid \mathbf{f}\in\operatorname{ND}\left(F_{\Delta_1}\right)\right\} && \text{(D.137)}\\
&= \left\{\lim_{\gamma^*\to\gamma}(1-\gamma^*)\mathbf{f}(\gamma^*) \mid \mathbf{f}\in F'_{\text{sub}}\right\} && \text{(D.138)}\\
&\subseteq \left\{\lim_{\gamma^*\to\gamma}(1-\gamma^*)\mathbf{f}(\gamma^*) \mid \mathbf{f}\in F_{\Delta_2}\right\} && \text{(D.139)}\\
&=: \textsc{Norm}\left(F_{\Delta_2},\gamma\right). && \text{(D.140)}
\end{aligned}
$$

Equation (D.135) follows by lemma D.49. Equation (D.137) follows because $\mathbf{P}_\phi$ is a continuous linear operator. Equation (D.139) follows by assumption.

$$
\mathbb{E}_{s_1\sim\Delta_1}\left[\textsc{Power}_{\mathcal{D}_{\text{bound}}}(s_1,\gamma)\right] := \mathbb{E}_{\substack{s_1\sim\Delta_1,\\ \mathbf{r}\sim\mathcal{D}_{\text{bound}}}}\left[\max_{\pi\in\Pi}\lim_{\gamma^*\to\gamma}\frac{1-\gamma^*}{\gamma^*}\left(\mathbf{f}^{\pi,s_1}(\gamma^*)-\mathbf{e}_{s_1}\right)^\top\mathbf{r}\right] \tag{D.141}
$$

$$= \mathbb{E}_{\mathbf{r}\sim\mathcal{D}_{\text{bound}}}\left[\max_{\pi\in\Pi}\lim_{\gamma^*\to\gamma}\frac{1-\gamma^*}{\gamma^*}\mathbb{E}_{s_1\sim\Delta_1}\left[\mathbf{f}^{\pi,s_1}(\gamma^*)-\mathbf{e}_{s_1}\right]^\top\mathbf{r}\right] \tag{D.142}$$

$$= \mathbb{E}_{\mathbf{r}\sim\mathcal{D}_{\text{bound}}}\left[\max_{\mathbf{d}\in\textsc{Norm}\left(F_{\Delta_1},\gamma\right)}\mathbf{d}^\top\mathbf{r}\right] \tag{D.143}$$

$$= \mathbb{E}_{\mathbf{r}\sim\mathcal{D}_{\text{bound}}}\left[\max_{\mathbf{d}\in\text{ND}\left(\textsc{Norm}\left(F_{\Delta_1},\gamma\right)\right)}\mathbf{d}^\top\mathbf{r}\right] \tag{D.144}$$

$$\leq_{\text{most: }\mathfrak{D}_{\text{bound}}} \mathbb{E}_{\mathbf{r}\sim\mathcal{D}_{\text{bound}}}\left[\max_{\mathbf{d}\in\textsc{Norm}\left(F_{\Delta_2},\gamma\right)}\mathbf{d}^\top\mathbf{r}\right] \tag{D.145}$$

$$= \mathbb{E}_{\mathbf{r}\sim\mathcal{D}_{\text{bound}}}\left[\max_{\pi\in\Pi}\lim_{\gamma^*\to\gamma}\frac{1-\gamma^*}{\gamma^*}\mathbb{E}_{s_2\sim\Delta_2}\left[\mathbf{f}^{\pi,s_2}(\gamma^*)-\mathbf{e}_{s_2}\right]^\top\mathbf{r}\right] \tag{D.146}$$

$$= \mathbb{E}_{\substack{s_2\sim\Delta_2,\\ \mathbf{r}\sim\mathcal{D}_{\text{bound}}}}\left[\max_{\pi\in\Pi}\lim_{\gamma^*\to\gamma}\frac{1-\gamma^*}{\gamma^*}\left(\mathbf{f}^{\pi,s_2}(\gamma^*)-\mathbf{e}_{s_2}\right)^\top\mathbf{r}\right] \tag{D.147}$$

$$=: \mathbb{E}_{s_2\sim\Delta_2}\left[\textsc{Power}_{\mathcal{D}_{\text{bound}}}\left(s_2,\gamma\right)\right]. \tag{D.148}$$

Equation (D.141) and eq. (D.148) follow by lemma D.50. Equation (D.142) and eq. (D.147) follow because each R has a stationary deterministic optimal policy $\pi\in\Pi^*\left(R,\gamma\right)\subseteq\Pi$ which simultaneously achieves optimal value at all states. Equation (D.144) follows by corollary D.16.

Apply lemma D.29 with $A:=\textsc{Norm}\left(F_{\Delta_1},\gamma\right), B:=\textsc{Norm}\left(F_{\Delta_2},\gamma\right)$, g the identity function, and involution ϕ (satisfying $\phi\cdot\text{ND}\left(A\right)\subseteq B$ by eq. (D.140)) in order to conclude that eq. (D.145) holds.

Suppose that $\text{ND}\left(F_{\Delta_2}\right)\setminus\phi\cdot\text{ND}\left(F_{\Delta_1}\right)$ is non-empty; let $F'_{\text{sub}}:=\phi\cdot\text{ND}\left(F_{\Delta_1}\right)$. Lemma D.37 shows that for all $\gamma\in(0,1)$, $\text{ND}\left(F_{\Delta_2}(\gamma)\right)\setminus F'_{\text{sub}}(\gamma)$ is non-empty. Lemma D.20 item 1 then implies that $\text{ND}\left(B\right)\setminus\phi\cdot A=\frac{1-\gamma}{\gamma}\left(\text{ND}\left(F_{\Delta_2}(\gamma)\right)-\mathbf{e}_s\right)\setminus\left(\frac{1-\gamma}{\gamma}F'_{\text{sub}}(\gamma)\right)$ is non-empty. Then lemma D.29 implies that for all $\gamma\in(0,1)$, eq. (D.145) is strict for all $\mathcal{D}_{X\text{-IID}}\in\mathfrak{D}_{\text{C/B/IID}}$ and $\mathbb{E}_{s_1\sim\Delta_1}\left[\textsc{Power}_{\mathcal{D}_{\text{bound}}}\left(s_1,\gamma\right)\right]\not\geq_{\text{most: }\mathfrak{D}_{\text{bound}}}\mathbb{E}_{s_2\sim\Delta_2}\left[\textsc{Power}_{\mathcal{D}_{\text{bound}}}\left(s_2,\gamma\right)\right]$.

We show that this result's preconditions holding for $F^*_{\Delta_i}$ implies the F_{Δ_i} preconditions. Suppose $F^*_{\Delta_i} := \left\{\mathbb{E}_{s_i \sim \Delta_i}\left[\mathbf{f}^{\pi, s_i}\right] \mid \pi \in \Pi\right\}$ for $i = 1, 2$ are such that $F^*_{\text{sub}} := \phi \cdot \operatorname{ND}\left(F^*_{\Delta_1}\right) \subseteq F^*_{\Delta_2}$. In the following, the Δ_i are represented as vectors in $\mathbb{R}^{|\mathcal{S}|}$, and γ is a variable.

$$
\begin{aligned}
\phi \cdot \left\{\gamma \mathbf{f} \mid \mathbf{f} \in \operatorname{ND}\left(F_{\Delta_1}\right)\right\} &= \phi \cdot \left(\operatorname{ND}\left(F^*_{\Delta_1} - \Delta_1\right)\right) && \text{(D.149)}\\
&= \phi \cdot \left(\operatorname{ND}\left(F^*_{\Delta_1}\right) - \Delta_1\right) && \text{(D.150)}\\
&= \left\{\mathbf{P}_\phi \mathbf{f} - \mathbf{P}_\phi \Delta_1 \mid \mathbf{f} \in \operatorname{ND}\left(F^*_{\Delta_1}\right)\right\} && \text{(D.151)}\\
&\subseteq \left\{\mathbf{f} - \Delta_2 \mid \mathbf{f} \in F^*_{\Delta_2}\right\} && \text{(D.152)}\\
&= \left\{\gamma \mathbf{f} \mid \mathbf{f} \in F_{\Delta_2}\right\}. && \text{(D.153)}
\end{aligned}
$$

Equation (D.150) follows from lemma D.20 item 2. Since we assumed that $\phi \cdot \operatorname{ND}\left(F^*_{\Delta_1}\right) \subseteq F^*_{\Delta_2}$, $\phi \cdot \{\Delta_1\} = \phi \cdot \left(\operatorname{ND}\left(F^*_{\Delta_1}\right)(0)\right) \subseteq F^*_{\Delta_2}(0) = \{\Delta_2\}$. This implies that $\mathbf{P}_\phi \Delta_1 = \Delta_2$ and so eq. (D.152) follows.

Equation (D.153) shows that $\phi \cdot \left\{\gamma \mathbf{f} \mid \mathbf{f} \in \operatorname{ND}\left(F_{\Delta_1}\right)\right\} \subseteq \left\{\gamma \mathbf{f} \mid \mathbf{f} \in F_{\Delta_2}\right\}$. But we then have $\phi \cdot \left\{\gamma \mathbf{f} \mid \mathbf{f} \in \operatorname{ND}\left(F_{\Delta_1}\right)\right\} := \left\{\gamma \mathbf{P}_\phi \mathbf{f} \mid \mathbf{f} \in \operatorname{ND}\left(F_{\Delta_1}\right)\right\} = \left\{\gamma \mathbf{f} \mid \mathbf{f} \in \phi \cdot \operatorname{ND}\left(F_{\Delta_1}\right)\right\} \subseteq \left\{\gamma \mathbf{f} \mid \mathbf{f} \in F_{\Delta_2}\right\}$. Thus, $\phi \cdot \operatorname{ND}\left(F_{\Delta_1}\right) \subseteq F_{\Delta_2}$.

Suppose $\operatorname{ND}\left(F^*_{\Delta_2}\right) \setminus \phi \cdot \operatorname{ND}\left(F^*_{\Delta_1}\right)$ is non-empty, which implies that

$$
\begin{aligned}
\phi \cdot \left\{\gamma \mathbf{f} \mid \mathbf{f} \in \operatorname{ND}\left(F_{\Delta_1}\right)\right\} &= \left\{\mathbf{P}_\phi \mathbf{f} - \mathbf{P}_\phi \Delta_1 \mid \mathbf{f} \in \operatorname{ND}\left(F^*_{\Delta_1}\right)\right\} && \text{(D.154)}\\
&= \left\{\mathbf{f} - \mathbf{P}_\phi \Delta_1 \mid \mathbf{f} \in \phi \cdot \operatorname{ND}\left(F^*_{\Delta_1}\right)\right\} && \text{(D.155)}\\
&\subsetneq \left\{\mathbf{f} - \Delta_2 \mid \mathbf{f} \in \operatorname{ND}\left(F^*_{\Delta_2}\right)\right\} && \text{(D.156)}\\
&= \left\{\gamma \mathbf{f} \mid \mathbf{f} \in \operatorname{ND}\left(F_{\Delta_2}\right)\right\}. && \text{(D.157)}
\end{aligned}
$$

Then $\operatorname{ND}\left(F_{\Delta_2}\right) \setminus \phi \cdot \operatorname{ND}\left(F_{\Delta_1}\right)$ must be non-empty. Therefore, if the preconditions of this result are met for $F^*_{\Delta_i}$, they are met for F_{Δ_i}. □

Proposition 5.22 (States with "more options" have more Power). *If $\mathcal{F}(s)$ contains a* *copy of $\mathcal{F}_{\text{nd}}(s')$ via ϕ, then $\forall \gamma \in [0,1] : \textsc{Power}_{\mathcal{D}_{bound}}(s,\gamma) \geq_{\text{most}} \textsc{Power}_{\mathcal{D}_{bound}}(s',\gamma)$. If $\mathcal{F}_{\text{nd}}(s) \setminus \phi \cdot \mathcal{F}_{\text{nd}}(s')$ is non-empty, then for all $\gamma \in (0,1)$, the converse $\leq_{\text{most}}$ statement does not hold.*

Proof. Let $F_{\text{sub}} \coloneqq \phi \cdot \mathcal{F}_{\text{nd}}(s') \subseteq \mathcal{F}(s)$. Let $\Delta_1 \coloneqq \mathbf{e}_{s'}, \Delta_2 \coloneqq \mathbf{e}_s$, and define $F^*_{\Delta_i} \coloneqq \left\{\mathbb{E}_{s_i \sim \Delta_i}\left[\mathbf{f}^{\pi,s_i}\right] \mid \pi \in \Pi\right\}$ for $i = 1,2$. Then $\mathcal{F}_{\text{nd}}(s') = \text{ND}\left(F^*_{\Delta_1}\right)$ is similar to $F_{\text{sub}} = F^*_{\text{sub}} \subseteq F^*_{\Delta_2} = \mathcal{F}(s)$ via involution ϕ. Apply lemma D.51 to conclude that

$$\forall \gamma \in [0,1] : \textsc{Power}_{\mathcal{D}_{\text{bound}}}\left(s',\gamma\right) \leq_{\text{most: } \mathfrak{D}_{\text{bound}}} \textsc{Power}_{\mathcal{D}_{\text{bound}}}\left(s,\gamma\right).$$

Furthermore, $\mathcal{F}_{\text{nd}}(s) = \text{ND}\left(F^*_{\Delta_2}\right)$, and $F_{\text{sub}} = F^*_{\text{sub}}$, and so if $\mathcal{F}_{\text{nd}}(s) \setminus \phi \cdot \mathcal{F}_{\text{nd}}(s') \coloneqq \mathcal{F}_{\text{nd}}(s) \setminus F_{\text{sub}} = \text{ND}\left(F^*_{\Delta_2}\right) \setminus F^*_{\text{sub}}$ is non-empty, then lemma D.51 shows that for all $\gamma \in (0,1)$, the inequality is strict for all $\mathcal{D}_{X\text{-IID}} \in \mathfrak{D}_{\text{C/B/IID}}$ and $\textsc{Power}_{\mathcal{D}_{\text{bound}}}\left(s',\gamma\right) \not\geq_{\text{most: } \mathfrak{D}_{\text{bound}}} \textsc{Power}_{\mathcal{D}_{\text{bound}}}\left(s,\gamma\right)$. □

Lemma D.52 (Non-dominated visit distribution functions never agree with other visit distribution functions at that state)**.** *Let $\mathbf{f} \in \mathcal{F}_{\text{nd}}(s), \mathbf{f}' \in \mathcal{F}(s) \setminus \{\mathbf{f}\}$. $\forall \gamma \in (0,1) : \mathbf{f}(\gamma) \neq \mathbf{f}'(\gamma)$.*

Proof. Let $\gamma \in (0,1)$. Since $\mathbf{f} \in \mathcal{F}_{\text{nd}}(s)$, there exists a $\gamma^* \in (0,1)$ at which $\mathbf{f}$ is strictly optimal for some reward function. Then by proposition D.35, we can produce another reward function for which $\mathbf{f}$ is strictly optimal at discount rate γ; in particular, proposition D.35 guarantees that the policies which induce $\mathbf{f}'$ are not optimal at γ. So $\mathbf{f}(\gamma) \neq \mathbf{f}'(\gamma)$. □

Corollary D.53 (Cardinality of non-dominated visit distributions)**.** *Let $F \subseteq \mathcal{F}(s)$. $\forall \gamma \in (0,1) : \left|F \cap \mathcal{F}_{\text{nd}}(s)\right| = \left|F(\gamma) \cap \mathcal{F}_{\text{nd}}(s,\gamma)\right|$.*

Proof. Let $\gamma \in (0,1)$. By applying lemma D.37 with $\Delta_d \coloneqq \mathbf{e}_s$, $\mathbf{f} \in \mathcal{F}_{\text{nd}}(s) = \text{ND}\left(\mathcal{F}(s)\right)$ iff $\mathbf{f}(\gamma) \in \text{ND}\left(\mathcal{F}(s,\gamma)\right)$. By lemma D.38, $\text{ND}\left(\mathcal{F}(s,\gamma)\right) = \mathcal{F}_{\text{nd}}(s,\gamma)$. So all $\mathbf{f} \in F \cap \mathcal{F}_{\text{nd}}(s)$ induce $\mathbf{f}(\gamma) \in F(\gamma) \cap \mathcal{F}_{\text{nd}}(s,\gamma)$, and $\left|F \cap \mathcal{F}_{\text{nd}}(s)\right| \geq \left|F(\gamma) \cap \mathcal{F}_{\text{nd}}(s,\gamma)\right|$.

Lemma D.52 implies that for all $\mathbf{f}, \mathbf{f}' \in \mathcal{F}_{\text{nd}}(s)$, $\mathbf{f} = \mathbf{f}'$ iff $\mathbf{f}(\gamma) = \mathbf{f}'(\gamma)$. Therefore, $\left|F \cap \mathcal{F}_{\text{nd}}(s)\right| \leq \left|F(\gamma) \cap \mathcal{F}_{\text{nd}}(s,\gamma)\right|$. So $\left|F \cap \mathcal{F}_{\text{nd}}(s)\right| = \left|F(\gamma) \cap \mathcal{F}_{\text{nd}}(s,\gamma)\right|$. □

Lemma D.54 (Optimality probability and state bottlenecks)**.** *Let* $X \coloneqq \textsc{Reach}\left(s', a'\right) \cup \textsc{Reach}\left(s', a\right)$*. Suppose that* s *can reach* X*, but only by taking actions equivalent to* a' *or* a *at state* s'*.* $F_{nd,a'} \coloneqq \mathcal{F}_{\text{nd}}(s \mid \pi(s') = a'), F_a \coloneqq \mathcal{F}(s \mid \pi(s') = a)$*. Suppose* F_a *contains a copy of* $F_{nd,a'}$ *via* ϕ *which fixes all states not belonging to* X*. Then* $\forall \gamma \in [0,1] : \mathbb{P}_{\mathcal{D}_{any}}\left(F_{nd,a'}, \gamma\right) \leq_{\text{most: } \mathfrak{D}_{any}} \mathbb{P}_{\mathcal{D}_{any}}\left(F_a, \gamma\right)$*.*

If $\mathcal{F}_{\text{nd}}(s) \cap \left(F_a \setminus \phi \cdot F_{nd,a'}\right)$ *is non-empty, then for all* $\gamma \in (0,1)$*, the inequality is strict for all* $\mathcal{D}_{X\text{-IID}} \in \mathfrak{D}_{\text{C/B/IID}}$*, and* $\mathbb{P}_{\mathcal{D}_{any}}\left(F_{nd,a'}, \gamma\right) \not\geq_{\text{most: } \mathfrak{D}_{any}} \mathbb{P}_{\mathcal{D}_{any}}\left(F_a, \gamma\right)$*.*

Proof. Let $F_{\text{sub}} \coloneqq \phi \cdot F_{\text{nd},a'}$. Let $F^* \coloneqq \bigcup_{\substack{a'' \in \mathcal{A}: \\ \left(a'' \not\equiv_{s'} a\right) \wedge \left(a'' \not\equiv_{s'} a'\right)}} \mathcal{F}(s \mid \pi(s') = a'') \cup F_{\text{nd},a'} \cup F_{\text{sub}}$.

$$
\begin{aligned}
\phi \cdot F^* &\coloneqq \phi \cdot \left(\bigcup_{\substack{a'' \in \mathcal{A}: \\ \left(a'' \not\equiv_{s'} a\right) \wedge \left(a'' \not\equiv_{s'} a'\right)}} \mathcal{F}(s \mid \pi(s') = a'') \cup F_{\text{nd},a'} \cup F_{\text{sub}} \right) && \text{(D.158)} \\
&= \bigcup_{\substack{a'' \in \mathcal{A}: \\ \left(a'' \not\equiv_{s'} a\right) \wedge \left(a'' \not\equiv_{s'} a'\right)}} \phi \cdot \mathcal{F}(s \mid \pi(s') = a'') \cup \left(\phi \cdot F_{\text{nd},a'}\right) \cup \left(\phi \cdot F_{\text{sub}}\right) && \text{(D.159)} \\
&= \bigcup_{\substack{a'' \in \mathcal{A}: \\ \left(a'' \not\equiv_{s'} a\right) \wedge \left(a'' \not\equiv_{s'} a'\right)}} \phi \cdot \mathcal{F}(s \mid \pi(s') = a'') \cup F_{\text{sub}} \cup F_{\text{nd},a'} && \text{(D.160)} \\
&= \bigcup_{\substack{a'' \in \mathcal{A}: \\ \left(a'' \not\equiv_{s'} a\right) \wedge \left(a'' \not\equiv_{s'} a'\right)}} \mathcal{F}(s \mid \pi(s') = a'') \cup F_{\text{sub}} \cup F_{\text{nd},a'} && \text{(D.161)} \\
&\eqqcolon F^*. && \text{(D.162)}
\end{aligned}
$$

Equation (D.160) follows because the involution ϕ ensures that $\phi \cdot F_{\text{sub}} = F_{\text{nd},a'}$. By assumption, ϕ fixes all $s' \notin \textsc{Reach}\left(s', a'\right) \cup \textsc{Reach}\left(s', a\right)$. Suppose $\mathbf{f} \in \mathcal{F}(s) \setminus \left(F_{\text{nd},a'} \cup F_a\right)$. By the bottleneck assumption, $\mathbf{f}$ does not visit states in $\textsc{Reach}\left(s', a'\right) \cup \textsc{Reach}\left(s', a\right)$. Therefore, $\mathbf{P}_\phi \mathbf{f} = \mathbf{f}$, and so eq. (D.161) follows.

Let $F_Z \coloneqq \left(\mathcal{F}(s) \setminus (\mathcal{F}(s \mid \pi(s) = a') \cup F_a)\right) \cup F_{\text{nd},a'} \cup F_a$. By definition, $F_Z \subseteq \mathcal{F}(s)$. Furthermore, $\mathcal{F}_{\text{nd}}(s) = \bigcup_{a'' \in \mathcal{A}} \mathcal{F}_{\text{nd}}(s \mid \pi(s') = a'') \subseteq \left(\mathcal{F}(s) \setminus (\mathcal{F}(s \mid \pi(s) = a') \cup F_a)\right) \cup \mathcal{F}_{\text{nd}}(s \mid \pi(s) = a') \cup F_a \eqqcolon F_Z$, and so $\mathcal{F}_{\text{nd}}(s) \subseteq F_Z$. Note that $F^* = F_Z \setminus (F_a \setminus F_{\text{sub}})$.

Case: $\gamma \in (0, 1)$**.**

$$\mathbb{P}_{\mathcal{D}_{\text{any}}}\left(F_{\text{nd},a'}, \gamma\right) = p_{\mathcal{D}_{\text{any}}}\left(F_{\text{nd},a'}(\gamma) \geq \mathcal{F}(s, \gamma)\right) \tag{D.163}$$
$$\leq_{\text{most: } \mathfrak{D}_{\text{any}}} p_{\mathcal{D}_{\text{any}}}\left(F_a(\gamma) \geq \mathcal{F}(s, \gamma)\right) \tag{D.164}$$
$$= \mathbb{P}_{\mathcal{D}_{\text{any}}}\left(F_{\text{nd},a'}, \gamma\right). \tag{D.165}$$

Equation (D.163) and eq. (D.165) follow from lemma D.42. Equation (D.164) follows by applying lemma D.33 with $A := F_{\text{nd},a'}(\gamma), B' := F_{\text{sub}}(\gamma), B := F_a(\gamma), C := \mathcal{F}(s, \gamma), Z := F_Z(\gamma)$ which satisfies $\text{ND}\,(C) = \mathcal{F}_{\text{nd}}(s, \gamma) \subseteq F_Z(\gamma) \subseteq \mathcal{F}(s, \gamma) = C$, and involution ϕ which satisfies $\phi \cdot F^*(\gamma) = \phi \cdot \Big(Z \setminus \big(B \setminus B'\big)\Big) = Z \setminus \big(B \setminus B'\big) = F^*(\gamma)$.

Suppose $\mathcal{F}_{\text{nd}}(s) \cap \big(F_a \setminus F_{\text{sub}}\big)$ is non-empty.

$$0 < \left|\mathcal{F}_{\text{nd}}(s) \cap \big(F_a \setminus F_{\text{sub}}\big)\right| = \left|\mathcal{F}_{\text{nd}}(s, \gamma) \cap \big(F_a(\gamma) \setminus F_{\text{sub}}(\gamma)\big)\right| =: \left|\text{ND}\,(C) \cap \big(B \setminus B'\big)\right|,$$

(with the first equality holding by corollary D.53), and so $\text{ND}\,(C) \cap \big(B \setminus B'\big)$ is non-empty. We also have $B := F_a(\gamma) \subseteq \mathcal{F}(s, \gamma) =: C$. Then reapplying lemma D.33, eq. (D.164) is strict for all $\mathcal{D}_{X\text{-IID}} \in \mathfrak{D}_{\text{C/B/IID}}$, and $\mathbb{P}_{\mathcal{D}_{\text{any}}}\left(F_{\text{nd},a'}, \gamma\right) \not\geq_{\text{most: } \mathfrak{D}_{\text{any}}} \mathbb{P}_{\mathcal{D}_{\text{any}}}\left(F_a, \gamma\right)$.

Case: $\gamma = 1$**,** $\gamma = 0$**.**

$$\mathbb{P}_{\mathcal{D}_{\text{any}}}\left(F_{\text{nd},a'}, 1\right) = \lim_{\gamma^* \to 1} \mathbb{P}_{\mathcal{D}_{\text{any}}}\left(F_{\text{nd},a'}, \gamma^*\right) \tag{D.166}$$
$$= \lim_{\gamma^* \to 1} p_{\mathcal{D}_{\text{any}}}\left(F_{\text{nd},a'}(\gamma^*) \geq \mathcal{F}(s, \gamma^*)\right) \tag{D.167}$$
$$\leq_{\text{most: } \mathfrak{D}_{\text{any}}} \lim_{\gamma^* \to 1} p_{\mathcal{D}_{\text{any}}}\left(F_a(\gamma^*) \geq \mathcal{F}(s, \gamma^*)\right) \tag{D.168}$$
$$= \lim_{\gamma^* \to 1} \mathbb{P}_{\mathcal{D}_{\text{any}}}\left(F_a, \gamma^*\right) \tag{D.169}$$
$$= \mathbb{P}_{\mathcal{D}_{\text{any}}}\left(F_a, 1\right). \tag{D.170}$$

Equation (D.166) and eq. (D.170) hold by proposition D.41. Equation (D.167) and eq. (D.169) follow by lemma D.42. Applying lemma D.34 with $\gamma := 1, I := (0, 1), F_A := F_{\text{nd},a'}, F_B := F_a, F_C := \mathcal{F}(s)$, F_Z as defined above, and involution ϕ (for which $\phi \cdot \Big(F_Z \setminus \big(F_B \setminus \phi \cdot F_A\big)\Big) = F_Z \setminus \big(F_B \setminus \phi \cdot F_A\big)$), we conclude that eq. (D.168) follows.

The $\gamma = 0$ case proceeds similarly to $\gamma = 1$. $\square$

Lemma D.55 (Action optimality probability is a special case of visit distribution optimality probability)**.** $\mathbb{P}_{\mathcal{D}_{any}}(s, a, \gamma) = \mathbb{P}_{\mathcal{D}_{any}}\big(\mathcal{F}(s \mid \pi(s) = a), \gamma\big)$.

Proof. Let $F_a := \mathcal{F}(s \mid \pi(s) = a)$. For $\gamma \in (0, 1)$,

$$\mathbb{P}_{\mathcal{D}_{\text{any}}}(s, a, \gamma) := \mathbb{P}_{R \sim \mathcal{D}_{\text{any}}}\big(\exists \pi^* \in \Pi^*(R, \gamma) : \pi^*(s) = a\big) \tag{D.171}$$

$$= \mathbb{P}_{\mathbf{r} \sim \mathcal{D}_{\text{any}}}\left(\exists \mathbf{f}^{\pi^*, s} \in F_a : \mathbf{f}^{\pi^*, s}(\gamma)^\top \mathbf{r} = \max_{\mathbf{f} \in \mathcal{F}(s)} \mathbf{f}(\gamma)^\top \mathbf{r}\right) \tag{D.172}$$

$$= \mathbb{P}_{\mathcal{D}_{\text{any}}}(F_a, \gamma). \tag{D.173}$$

By lemma D.6, if $\exists \pi^* \in \Pi^*(R, \gamma) : \pi^*(s) = a$, then it induces some optimal $\mathbf{f}^{\pi^*, s} \in F_a$. Conversely, if $\mathbf{f}^{\pi^*, s} \in F_a$ is optimal at $\gamma \in (0, 1)$, then π^* chooses optimal actions on the support of $\mathbf{f}^{\pi^*, s}(\gamma)$. Let π' agree with π^* on that support and let π' take optimal actions at all other states. Then $\pi' \in \Pi^*(R, \gamma)$ and $\pi'(s) = a$. So eq. (D.172) follows.

Suppose $\gamma = 0$ or $\gamma = 1$. Consider any sequence $(\gamma_n)_{n=1}^{\infty}$ converging to γ, and let $\mathcal{D}_{\text{any}}$ induce probability measure F.

$$\mathbb{P}_{\mathcal{D}_{\text{any}}}(F_a, \gamma) := \lim_{\gamma^* \to \gamma} \mathbb{P}_{\mathcal{D}_{\text{any}}}\big(F_a, \gamma^*\big) \tag{D.174}$$

$$= \lim_{\gamma^* \to \gamma} \mathbb{P}_{R \sim \mathcal{D}_{\text{any}}}\Big(\exists \pi^* \in \Pi^*\big(R, \gamma^*\big) : \pi^*(s) = a\Big) \tag{D.175}$$

$$= \lim_{n \to \infty} \mathbb{P}_{R \sim \mathcal{D}_{\text{any}}}\big(\exists \pi^* \in \Pi^*(R, \gamma_n) : \pi^*(s) = a\big) \tag{D.176}$$

$$= \lim_{n \to \infty} \int_{\mathbb{R}^{\mathcal{S}}} \mathbb{1}_{\exists \pi^* \in \Pi^*(R, \gamma_n) : \pi^*(s) = a} \, \mathrm{d}F(R) \tag{D.177}$$

$$= \int_{\mathbb{R}^{\mathcal{S}}} \lim_{n \to \infty} \mathbb{1}_{\exists \pi^* \in \Pi^*(R, \gamma_n) : \pi^*(s) = a} \, \mathrm{d}F(R) \tag{D.178}$$

$$= \int_{\mathbb{R}^{\mathcal{S}}} \mathbb{1}_{\exists \pi^* \in \Pi^*(R, \gamma) : \pi^*(s) = a} \, \mathrm{d}F(R) \tag{D.179}$$

$$=: \mathbb{P}_{\mathcal{D}_{\text{any}}}(s, a, \gamma). \tag{D.180}$$

Equation (D.175) follows by eq. (D.173). for $\gamma^* \in [0, 1]$, let $f_{\gamma^*}(R) := \mathbb{1}_{\exists \pi^* \in \Pi^*(R, \gamma^*) : \pi^*(s) = a}$.

For each $R \in \mathbb{R}^{\mathcal{S}}$, lemma D.40 exists $\gamma_x \approx \gamma$ such that for all intermediate γ'_x between γ_x and γ, $\Pi^*\left(R, \gamma'_x\right) = \Pi^*\left(R, \gamma\right)$. Since $\gamma_n \to \gamma$, this means that $\left(f_{\gamma_n}\right)_{n=1}^{\infty}$ converges pointwise to f_γ. Furthermore, $\forall n \in \mathbb{N}, R \in \mathbb{R}^{\mathcal{S}} : \left|f_{\gamma_n}(R)\right| \leq 1$ by definition. Therefore, eq. (D.178) follows by Lebesgue's dominated convergence theorem. □

Proposition 5.25 (Keeping options open tends to be POWER-seeking and tends to be optimal)**.**

Suppose $F_a \coloneqq \mathcal{F}(s \mid \pi(s) = a)$ *contains a copy of* $F_{a'} \coloneqq \mathcal{F}(s \mid \pi(s) = a')$ *via* ϕ*.*

1. *If* $s \notin \text{REACH}\left(s, a'\right)$*, then* $\forall \gamma \in [0,1] : \mathbb{E}_{s_a \sim T(s,a)}\left[\text{POWER}_{\mathcal{D}_{bound}}\left(s_a, \gamma\right)\right] \geq_{\text{most: } \mathfrak{D}_{bound}} \mathbb{E}_{s_{a'} \sim T(s,a')}\left[\text{POWER}_{\mathcal{D}_{bound}}\left(s_{a'}, \gamma\right)\right]$*.*
2. *If* s *can only reach the states of* $\text{REACH}\left(s, a'\right) \cup \text{REACH}\left(s, a\right)$ *by taking actions equivalent to* a' *or* a *at state* s*, then* $\forall \gamma \in [0,1] : \mathbb{P}_{\mathcal{D}_{any}}\left(s, a, \gamma\right) \geq_{\text{most: } \mathfrak{D}_{any}} \mathbb{P}_{\mathcal{D}_{any}}\left(s, a', \gamma\right)$*.*

If $\mathcal{F}_{\text{nd}}(s) \cap \left(F_a \setminus \phi \cdot F_{a'}\right)$ *is non-empty, then* $\forall \gamma \in (0,1)$*, the converse* $\leq_{\text{most}}$ *statements do not hold.*

Proof. Note that by definition 5.3, $F_{a'}(0) = \{\mathbf{e}_s\} = F_a(0)$. Since $\phi \cdot F_{a'} \subseteq F_a$, in particular we have $\phi \cdot F_{a'}(0) = \left\{\mathbf{P}_\phi \mathbf{e}_s\right\} \subseteq \{\mathbf{e}_s\} = F_a(0)$, and so $\phi(s) = s$.

Item 1. For state probability distribution $\Delta_s \in \Delta(\mathcal{S})$, let $F^*_{\Delta_s} \coloneqq \left\{\mathbb{E}_{s' \sim \Delta_s}\left[\mathbf{f}^{\pi, s'}\right] \mid \pi \in \Pi\right\}$. Unless otherwise stated, we treat γ as a variable in this item; we apply element-wise vector addition, constant multiplication, and variable multiplication via the conventions outlined in definition D.19.

$$F_{a'} = \left\{\mathbf{e}_s + \gamma \underset{s_{a'} \sim T(s,a')}{\mathbb{E}}\left[\mathbf{f}^{\pi, s_{a'}}\right] \mid \pi \in \Pi : \pi(s) = a'\right\} \tag{D.181}$$

$$= \left\{\mathbf{e}_s + \gamma \underset{s_{a'} \sim T(s,a')}{\mathbb{E}}\left[\mathbf{f}^{\pi, s_{a'}}\right] \mid \pi \in \Pi\right\} \tag{D.182}$$

$$= \mathbf{e}_s + \gamma F^*_{T(s,a')}. \tag{D.183}$$

Equation (D.181) follows by definition 5.3, since each $\mathbf{f} \in \mathcal{F}(s)$ has an initial term of $\mathbf{e}_s$. Equation (D.182) follows because $s \notin \text{REACH}\left(s, a'\right)$, and so for all $s_{a'} \in \text{supp}(T(s, a'))$,

$\mathbf{f}^{\pi,s_{a'}}$ is unaffected by the choice of action $\pi(s)$. Note that similar reasoning implies that $F_a \subseteq \mathbf{e}_s + \gamma F^*_{T(s,a)}$ (because eq. (D.182) is a containment relation in general).

Since $F_{a'} = \mathbf{e}_s + \gamma F^*_{T(s,a')}$, if F_a contains a copy of $F_{a'}$ via ϕ, then $F^*_{T(s,a)}$ contains a copy of $F^*_{T(s,a')}$ via ϕ. Then $\phi \cdot \mathrm{ND}\left(F^*_{T(s,a')}\right) \subseteq \phi \cdot F^*_{T(s,a')} \subseteq F^*_{T(s,a)}$, and so $F^*_{T(s,a)}$ contains a copy of $\mathrm{ND}\left(F^*_{T(s,a')}\right)$. Then apply lemma D.51 with $\Delta_1 \coloneqq T(s,a')$ and $\Delta_2 \coloneqq T(s,a)$ to conclude that $\forall \gamma \in [0,1]$:

$$\mathop{\mathbb{E}}_{s_{a'} \sim T(s,a')} \left[\mathrm{P{\small OWER}}_{\mathcal{D}_{\text{bound}}}\left(s_{a'}, \gamma\right)\right] \leq_{\text{most: } \mathfrak{D}_{\text{bound}}} \mathop{\mathbb{E}}_{s_a \sim T(s,a)} \left[\mathrm{P{\small OWER}}_{\mathcal{D}_{\text{bound}}}\left(s_a, \gamma\right)\right].$$

Suppose $\mathcal{F}_{\text{nd}}(s) \cap \left(F_a \setminus \phi \cdot F_{a'}\right)$ is non-empty. To apply the second condition of lemma D.51, we want to demonstrate that $\mathrm{ND}\left(F^*_{T(s,a)}\right) \setminus \phi \cdot \mathrm{ND}\left(F^*_{T(s,a')}\right)$ is also non-empty.

First consider $\mathbf{f} \in \mathcal{F}_{\text{nd}}(s) \cap F_a$. Because $F_a \subseteq \mathbf{e}_s + \gamma F^*_{T(s,a)}$, we have that $\gamma^{-1}(\mathbf{f} - \mathbf{e}_s) \in F^*_{T(s,a)}$. Because $\mathbf{f} \in \mathcal{F}_{\text{nd}}(s)$, by definition 5.6, $\exists \mathbf{r} \in \mathbb{R}^{|\mathcal{S}|}, \gamma_x \in (0,1)$ such that

$$\mathbf{f}(\gamma_x)^\top \mathbf{r} > \max_{\mathbf{f}' \in \mathcal{F}(s) \setminus \{\mathbf{f}\}} \mathbf{f}'(\gamma_x)^\top \mathbf{r}. \tag{D.184}$$

Then since $\gamma_x \in (0,1)$,

$$\gamma_x^{-1}(\mathbf{f}(\gamma_x) - \mathbf{e}_s)^\top \mathbf{r} > \max_{\mathbf{f}' \in \mathcal{F}(s) \setminus \{\mathbf{f}\}} \gamma_x^{-1}(\mathbf{f}'(\gamma_x) - \mathbf{e}_s)^\top \mathbf{r} \tag{D.185}$$

$$= \max_{\mathbf{f}' \in \gamma_x^{-1}\left((\mathcal{F}(s) \setminus \{\mathbf{f}\}) - \mathbf{e}_s\right)} \mathbf{f}'(\gamma_x)^\top \mathbf{r} \tag{D.186}$$

$$\geq \max_{\mathbf{f}' \in \gamma_x^{-1}\left((F_a \setminus \{\mathbf{f}\}) - \mathbf{e}_s\right)} \mathbf{f}'(\gamma_x)^\top \mathbf{r} \tag{D.187}$$

$$= \max_{\mathbf{f}' \in F^*_{T(s,a)} \setminus \left\{\gamma_x^{-1}(\mathbf{f} - \mathbf{e}_s)\right\}} \mathbf{f}'(\gamma_x)^\top \mathbf{r}. \tag{D.188}$$

Equation (D.187) holds because $F_a \subseteq \mathcal{F}(s)$. By assumption, action a is optimal for $\mathbf{r}$ at state s and at discount rate γ_x. Equation (D.182) shows that $F^*_{T(s,a)}$ potentially allows the agent a non-stationary policy choice at s, but non-stationary policies cannot increase optimal value [68]. Therefore, eq. (D.188) holds.

We assumed that $\gamma^{-1}(\mathbf{f} - \mathbf{e}_s) \in \gamma^{-1}(\mathcal{F}_{\text{nd}}(s) - \mathbf{e}_s)$. Furthermore, since we just showed that $\gamma^{-1}(\mathbf{f}-\mathbf{e}_s) \in F^*_{T(s,a)}$ is strictly optimal over the other elements of $F^*_{T(s,a)}$ for reward function $\mathbf{r}$ at discount rate $\gamma_x \in (0,1)$, we conclude that it is an element of $\operatorname{ND}\left(F^*_{T(s,a)}\right)$ by definition D.18. Then we conclude that $\gamma^{-1}(\mathcal{F}_{\text{nd}}(s)-\mathbf{e}_s) \cap F^*_{T(s,a)} \subseteq \operatorname{ND}\left(F^*_{T(s,a)}\right)$.

We now show that $\operatorname{ND}\left(F^*_{T(s,a)}\right) \setminus \phi \cdot \operatorname{ND}\left(F^*_{T(s,a')}\right)$ is non-empty.

$$0 < \left|\mathcal{F}_{\text{nd}}(s) \cap \left(F_a \setminus \phi \cdot F_{a'}\right)\right| \tag{D.189}$$

$$= \left|\gamma^{-1}\left(\mathcal{F}_{\text{nd}}(s) \cap \left(F_a \setminus \phi \cdot F_{a'}\right) - \mathbf{e}_s\right)\right| \tag{D.190}$$

$$\leq \left|\gamma^{-1}\left(\mathcal{F}_{\text{nd}}(s) - \mathbf{e}_s\right) \cap \left(F^*_{T(s,a)} \setminus \phi \cdot F^*_{T(s,a')}\right)\right| \tag{D.191}$$

$$= \left|\left(\gamma^{-1}\left(\mathcal{F}_{\text{nd}}(s) - \mathbf{e}_s\right) \cap F^*_{T(s,a)}\right) \setminus \phi \cdot F^*_{T(s,a')}\right| \tag{D.192}$$

$$\leq \left|\operatorname{ND}\left(F^*_{T(s,a)}\right) \setminus \phi \cdot F^*_{T(s,a')}\right| \tag{D.193}$$

$$\leq \left|\operatorname{ND}\left(F^*_{T(s,a)}\right) \setminus \phi \cdot \operatorname{ND}\left(F^*_{T(s,a')}\right)\right|. \tag{D.194}$$

Equation (D.189) follows by the assumption that $\mathcal{F}_{\text{nd}}(s) \cap \left(F_a \setminus \phi \cdot F_{a'}\right)$ is non-empty. Let $\mathbf{f}, \mathbf{f}' \in \mathcal{F}_{\text{nd}}(s) \cap \left(F_a \setminus \phi \cdot F_{a'}\right)$ be distinct. Then we must have that for some $\gamma_x \in (0,1)$, $\mathbf{f}(\gamma_x) \neq \mathbf{f}'(\gamma_x)$. This holds iff $\gamma_x^{-1}(\mathbf{f}(\gamma_x) - \mathbf{e}_s) \neq \gamma_x^{-1}(\mathbf{f}'(\gamma_x) - \mathbf{e}_s)$, and so eq. (D.190) holds.

Equation (D.191) holds because $F_a \subseteq \mathbf{e}_s + \gamma F^*_{T(s,a)}$ and $F'_a = \mathbf{e}_s + \gamma F^*_{T(s,a')}$ by eq. (D.183). Equation (D.193) holds because we showed above that

$$\gamma^{-1}(\mathcal{F}_{\text{nd}}(s) - \mathbf{e}_s) \cap F^*_{T(s,a)} \subseteq \operatorname{ND}\left(F^*_{T(s,a)}\right).$$

Equation (D.194) holds because $\operatorname{ND}\left(F^*_{T(s,a')}\right) \subseteq F^*_{T(s,a')}$ by definition D.18.

Therefore, $\operatorname{ND}\left(F^*_{T(s,a)}\right) \setminus \phi \cdot \operatorname{ND}\left(F^*_{T(s,a')}\right)$ is non-empty, and so apply the second condition of lemma D.51 to conclude that for all $\mathcal{D}_{X\text{-IID}} \in \mathfrak{D}_{\text{C/B/IID}}$,

$$\forall \gamma \in (0,1): \mathop{\mathbb{E}}_{s_{a'} \sim T(s,a')}\left[\operatorname{POWER}_{\mathcal{D}_{X\text{-IID}}}(s_{a'}, \gamma)\right] < \mathop{\mathbb{E}}_{s_a \sim T(s,a)}\left[\operatorname{POWER}_{\mathcal{D}_{X\text{-IID}}}(s_a, \gamma)\right],$$

and that

$$\forall \gamma \in (0,1) : \mathop{\mathbb{E}}_{s_{a'} \sim T(s,a')} \left[\textsc{Power}_{\mathcal{D}_{\text{bound}}}(s_{a'}, \gamma)\right] \not\geq_{\text{most: } \mathfrak{D}_{\text{bound}}} \mathop{\mathbb{E}}_{s_a \sim T(s,a)} \left[\textsc{Power}_{\mathcal{D}_{\text{bound}}}(s_a, \gamma)\right].$$

Item 2. Let $\phi'(s_x) \coloneqq \phi(s_x)$ when $s_x \in \textsc{Reach}\left(s, a'\right) \cup \textsc{Reach}\left(s, a\right)$, and equal s_x otherwise. Since ϕ is an involution, so is ϕ'.

$$\phi' \cdot F_{a'} \coloneqq \left\{\mathbf{P}_{\phi'}\left(\mathbf{e}_s + \gamma \mathop{\mathbb{E}}_{s_{a'} \sim T(s,a')} \left[\mathbf{f}^{\pi, s_{a'}}\right]\right) \mid \pi \in \Pi, \pi(s) = a'\right\} \tag{D.195}$$

$$= \left\{\mathbf{e}_s + \gamma \mathop{\mathbb{E}}_{s_{a'} \sim T(s,a')} \left[\mathbf{P}_{\phi'}\mathbf{f}^{\pi, s_{a'}}\right] \mid \pi \in \Pi, \pi(s) = a'\right\} \tag{D.196}$$

$$= \left\{\mathbf{P}_{\phi}\mathbf{e}_s + \gamma \mathop{\mathbb{E}}_{s_{a'} \sim T(s,a')} \left[\mathbf{P}_{\phi}\mathbf{f}^{\pi, s_{a'}}\right] \mid \pi \in \Pi, \pi(s) = a'\right\} \tag{D.197}$$

$$\eqqcolon \phi \cdot F_{a'} \tag{D.198}$$

$$\subseteq F_a. \tag{D.199}$$

Equation (D.196) follows because if $s \in \textsc{Reach}\left(s, a'\right) \cup \textsc{Reach}\left(s, a\right)$, then we already showed that ϕ fixes s. Otherwise, $\phi'(s) = s$ by definition. Equation (D.197) follows by the definition of ϕ' on $\textsc{Reach}\left(s, a'\right) \cup \textsc{Reach}\left(s, a\right)$ and because $\mathbf{e}_s = \mathbf{P}_{\phi}\mathbf{e}_s$. Next, we assumed that $\phi \cdot F_{a'} \subseteq F_a$, and so eq. (D.199) holds.

Therefore, F_a contains a copy of $F_{a'}$ via ϕ' fixing all $s_x \notin \textsc{Reach}\left(s, a'\right) \cup \textsc{Reach}\left(s, a\right)$. Therefore, F_a contains a copy of $F_{\text{nd},a'} \coloneqq \mathcal{F}_{\text{nd}}(s) \cap F_{a'}$ via the same ϕ'. Then apply lemma D.54 with $s' \coloneqq s$ to conclude that $\forall \gamma \in [0,1] : \mathbb{P}_{\mathcal{D}_{\text{any}}}\left(F_{a'}, \gamma\right) \leq_{\text{most: } \mathfrak{D}_{\text{any}}} \mathbb{P}_{\mathcal{D}_{\text{any}}}\left(F_a, \gamma\right)$. By lemma D.55, $\mathbb{P}_{\mathcal{D}_{\text{any}}}\left(s, a', \gamma\right) = \mathbb{P}_{\mathcal{D}_{\text{any}}}\left(F_{a'}, \gamma\right)$ and $\mathbb{P}_{\mathcal{D}_{\text{any}}}\left(s, a, \gamma\right) = \mathbb{P}_{\mathcal{D}_{\text{any}}}\left(F_a, \gamma\right)$. Therefore, $\forall \gamma \in [0,1] : \mathbb{P}_{\mathcal{D}_{\text{any}}}\left(s, a', \gamma\right) \leq_{\text{most: } \mathfrak{D}_{\text{any}}} \mathbb{P}_{\mathcal{D}_{\text{any}}}\left(s, a, \gamma\right)$.

If $\mathcal{F}_{\text{nd}}(s) \cap \left(F_a \setminus \phi \cdot F_{a'}\right)$ is non-empty, then apply the second condition of lemma D.54 to conclude that for all $\gamma \in (0,1)$, the inequality is strict for all $\mathcal{D}_{X\text{-IID}} \in \mathfrak{D}_{\text{C/B/IID}}$, and $\mathbb{P}_{\mathcal{D}_{\text{any}}}\left(s, a', \gamma\right) \not\geq_{\text{most: } \mathfrak{D}_{\text{any}}} \mathbb{P}_{\mathcal{D}_{\text{any}}}\left(s, a, \gamma\right)$. $\square$

D.4.4.2 When $\gamma = 1$, optimal policies tend to navigate towards "larger" sets of cycles

Lemma D.56 (POWER identity when $\gamma = 1$).

$$\text{POWER}_{\mathcal{D}_{bound}}(s, 1) = \mathbb{E}_{\mathbf{r}\sim\mathcal{D}_{bound}}\left[\max_{\mathbf{d}\in\text{RSD}(s)} \mathbf{d}^\top \mathbf{r}\right] = \mathbb{E}_{\mathbf{r}\sim\mathcal{D}_{bound}}\left[\max_{\mathbf{d}\in\text{RSD}_{\text{nd}}(s)} \mathbf{d}^\top \mathbf{r}\right]. \quad \text{(D.200)}$$

Proof.

$$\text{POWER}_{\mathcal{D}_{\text{bound}}}(s, 1) = \mathbb{E}_{\mathbf{r}\sim\mathcal{D}_{\text{bound}}}\left[\max_{\mathbf{f}^{\pi,s}\in\mathcal{F}(s)} \lim_{\gamma\to 1} \frac{1-\gamma}{\gamma}\left(\mathbf{f}^{\pi,s}(\gamma) - \mathbf{e}_s\right)^\top \mathbf{r}\right] \quad \text{(D.201)}$$

$$= \mathbb{E}_{\mathbf{r}\sim\mathcal{D}_{\text{bound}}}\left[\max_{\mathbf{d}\in\text{RSD}(s)} \mathbf{d}^\top \mathbf{r}\right] \quad \text{(D.202)}$$

$$= \mathbb{E}_{\mathbf{r}\sim\mathcal{D}_{\text{bound}}}\left[\max_{\mathbf{d}\in\text{RSD}_{\text{nd}}(s)} \mathbf{d}^\top \mathbf{r}\right]. \quad \text{(D.203)}$$

Equation (D.201) follows by lemma D.50. Equation (D.202) follows by the definition of RSD (s) (definition 5.26). Equation (D.203) follows because for all $\mathbf{r} \in \mathbb{R}^{|\mathcal{S}|}$, corollary D.16 shows that $\max_{\mathbf{d}\in\text{RSD}(s)} \mathbf{d}^\top \mathbf{r} = \max_{\mathbf{d}\in\text{ND}(\text{RSD}(s))} \mathbf{d}^\top \mathbf{r} =: \max_{\mathbf{d}\in\text{RSD}_{\text{nd}}(s)} \mathbf{d}^\top \mathbf{r}$. □

Proposition 5.28 (When $\gamma = 1$, RSDs control POWER)**.** *If* RSD (s) *contains a copy of* $\text{RSD}_{\text{nd}}(s')$ *via* ϕ*, then* $\text{POWER}_{\mathcal{D}_{bound}}(s, 1) \geq_{\text{most}} \text{POWER}_{\mathcal{D}_{bound}}(s', 1)$*. If* $\text{RSD}_{\text{nd}}(s) \setminus \phi \cdot \text{RSD}_{\text{nd}}(s')$ *is non-empty, then the converse* $\leq_{\text{most}}$ *statement does not hold.*

Proof. Suppose $\text{RSD}_{\text{nd}}(s')$ is similar to $D \subseteq \text{RSD}(s)$ via involution ϕ.

$$\text{POWER}_{\mathcal{D}_{\text{bound}}}(s', 1) = \mathbb{E}_{\mathbf{r}\sim\mathcal{D}_{\text{bound}}}\left[\max_{\mathbf{d}\in\text{RSD}_{\text{nd}}(s')} \mathbf{d}^\top \mathbf{r}\right] \quad \text{(D.204)}$$

$$\leq_{\text{most: } \mathfrak{D}_{\text{bound}}} \mathbb{E}_{\mathbf{r}\sim\mathcal{D}_{\text{bound}}}\left[\max_{\mathbf{d}\in\text{RSD}_{\text{nd}}(s)} \mathbf{d}^\top \mathbf{r}\right] \quad \text{(D.205)}$$

$$= \text{POWER}_{\mathcal{D}_{\text{bound}}}(s, 1) \quad \text{(D.206)}$$

Equation (D.204) and eq. (D.206) follow from lemma D.56. By applying lemma D.29 with $A \coloneqq \mathrm{RSD}\left(s'\right), B' \coloneqq D, B \coloneqq \mathrm{RSD}\left(s\right)$ and g the identity function, eq. (D.205) follows.

Suppose $\mathrm{RSD}_{\mathrm{nd}}\left(s\right) \setminus D$ is non-empty. By the same result, eq. (D.205) is a strict inequality for all $\mathcal{D}_{X\text{-IID}} \in \mathfrak{D}_{\mathrm{C/B/IID}}$, and we conclude that $\textsc{Power}_{\mathcal{D}_{\mathrm{bound}}}\left(s', 1\right) \not\geq_{\mathrm{most:}\, \mathfrak{D}_{\mathrm{bound}}} \textsc{Power}_{\mathcal{D}_{\mathrm{bound}}}\left(s, 1\right)$. ☐

Theorem 5.29 (Average-optimal policies tend to end up in "larger" sets of RSDs)**.** *Let $D, D' \subseteq \mathrm{RSD}\left(s\right)$. Suppose that D contains a copy of D' via ϕ, and that the sets $D \cup D'$ and $\mathrm{RSD}_{\mathrm{nd}}\left(s\right) \setminus \left(D' \cup D\right)$ have pairwise orthogonal vector elements (*i.e. *pairwise disjoint vector support). Then $\mathbb{P}_{\mathcal{D}_{any}}\left(D, \text{average}\right) \geq_{\mathrm{most}} \mathbb{P}_{\mathcal{D}_{any}}\left(D', \text{average}\right)$. If $\mathrm{RSD}_{\mathrm{nd}}\left(s\right) \cap \left(D \setminus \phi \cdot D'\right)$ is non-empty, the converse $\leq_{\mathrm{most}}$ statement does not hold.*

Proof. Let $D_{\mathrm{sub}} \coloneqq \phi \cdot D'$, where $D_{\mathrm{sub}} \subseteq D$ by assumption. Let

$$X \coloneqq \{s_i \in \mathcal{S} \mid \max_{\mathbf{d} \in D' \cup D} \mathbf{d}^\top \mathbf{e}_{s_i} > 0\}.$$

Define

$$\phi'(s_i) \coloneqq \begin{cases} \phi(s_i) & \text{if } s_i \in X \\ s_i & \text{else.} \end{cases} \tag{D.207}$$

Since ϕ is an involution, ϕ' is also an involution. Furthermore, by the definition of X, $\phi' \cdot D' = D_{\mathrm{sub}}$ and $\phi' \cdot D_{\mathrm{sub}} = D'$ (because we assumed that both equalities hold for ϕ).

Let $D^* \coloneqq D' \cup D_{\mathrm{sub}} \cup \left(\mathrm{RSD}_{\mathrm{nd}}\left(s\right) \setminus (D' \cup D)\right)$.

$$\begin{aligned} \phi' \cdot D^* &\coloneqq \phi' \cdot \Big(D' \cup D_{\mathrm{sub}} \cup \left(\mathrm{RSD}_{\mathrm{nd}}\left(s\right) \setminus (D' \cup D)\right)\Big) && \text{(D.208)} \\ &= \left(\phi' \cdot D'\right) \cup \left(\phi' \cdot D_{\mathrm{sub}}\right) \cup \phi' \cdot \left(\mathrm{RSD}_{\mathrm{nd}}\left(s\right) \setminus (D' \cup D)\right) && \text{(D.209)} \\ &= D_{\mathrm{sub}} \cup D' \cup \left(\mathrm{RSD}_{\mathrm{nd}}\left(s\right) \setminus (D' \cup D)\right) && \text{(D.210)} \\ &=: D^*. && \text{(D.211)} \end{aligned}$$

In eq. (D.210), we know that $\phi' \cdot D' = D_{\mathrm{sub}}$ and $\phi' \cdot D_{\mathrm{sub}} = D'$. We just need to show that $\phi' \cdot \left(\mathrm{RSD}_{\mathrm{nd}}\left(s\right) \setminus (D' \cup D)\right) = \mathrm{RSD}_{\mathrm{nd}}\left(s\right) \setminus (D' \cup D)$.

Suppose $\exists s_i \in X, \mathbf{d}' \in \mathrm{RSD}_{\mathrm{nd}}(s) \setminus (D' \cup D) : \mathbf{d}'^\top \mathbf{e}_{s_i} > 0$. By the definition of X, $\exists \mathbf{d} \in D' \cup D : \mathbf{d}^\top \mathbf{e}_{s_i} > 0$. Then

$$\begin{aligned}
\mathbf{d}^\top \mathbf{d}' &= \sum_{j=1}^{|\mathcal{S}|} \mathbf{d}^\top (\mathbf{d}' \odot \mathbf{e}_{s_j}) && \text{(D.212)}\\
&\geq \mathbf{d}^\top (\mathbf{d}' \odot \mathbf{e}_{s_i}) && \text{(D.213)}\\
&= \mathbf{d}^\top \left((\mathbf{d}'^\top \mathbf{e}_{s_i}) \mathbf{e}_{s_i} \right) && \text{(D.214)}\\
&= (\mathbf{d}'^\top \mathbf{e}_{s_i}) \cdot (\mathbf{d}^\top \mathbf{e}_{s_i}) && \text{(D.215)}\\
&> 0. && \text{(D.216)}
\end{aligned}$$

Equation (D.212) follows from the definitions of the dot and Hadamard products. Equation (D.213) follows because $\mathbf{d}$ and $\mathbf{d}'$ have non-negative entries. Equation (D.216) follows because $\mathbf{d}^\top \mathbf{e}_{s_i}$ and $\mathbf{d}'^\top \mathbf{e}_{s_i}$ are both positive. But eq. (D.216) shows that $\mathbf{d}^\top \mathbf{d}' > 0$, contradicting our assumption that $\mathbf{d}$ and $\mathbf{d}'$ are orthogonal.

Therefore, such an s_i cannot exist, and $X' := \left\{ s'_i \in \mathcal{S} \mid \max_{\mathbf{d}' \in \mathrm{RSD}_{\mathrm{nd}}(s) \setminus (D' \cup D)} \mathbf{d}'^\top \mathbf{e}_{s_i} > 0 \right\} \subseteq (\mathcal{S} \setminus X)$. By eq. (D.207), $\forall s'_i \in X' : \phi'(s'_i) = s'_i$. Thus, $\phi' \cdot \left(\mathrm{RSD}_{\mathrm{nd}}(s) \setminus (D' \cup D) \right) = \mathrm{RSD}_{\mathrm{nd}}(s) \setminus (D' \cup D)$, and eq. (D.210) follows. We conclude that $\phi' \cdot D^* = D^*$.

Consider $Z := \left(\mathrm{RSD}_{\mathrm{nd}}(s) \setminus (D' \cup D) \right) \cup D \cup D'$. First, $Z \subseteq \mathrm{RSD}(s)$ by definition. Second, $\mathrm{RSD}_{\mathrm{nd}}(s) = \mathrm{RSD}_{\mathrm{nd}}(s) \setminus (D' \cup D) \cup (\mathrm{RSD}_{\mathrm{nd}}(s) \cap D') \cup (\mathrm{RSD}_{\mathrm{nd}}(s) \cap D) \subseteq Z$. Note that $D^* = Z \setminus (D \setminus D_{\mathrm{sub}})$.

$$\begin{aligned}
\mathbb{P}_{\mathcal{D}_{\mathrm{any}}} \left(D', \text{average} \right) &= p_{\mathcal{D}_{\mathrm{any}}} \left(D' \geq \mathrm{RSD}(s) \right) && \text{(D.217)}\\
&\leq_{\text{most: } \mathfrak{D}_{\mathrm{any}}} p_{\mathcal{D}_{\mathrm{any}}} \left(D \geq \mathrm{RSD}(s) \right) && \text{(D.218)}\\
&= \mathbb{P}_{\mathcal{D}_{\mathrm{any}}} (D, \text{average}). && \text{(D.219)}
\end{aligned}$$

Since $\phi \cdot D' \subseteq D$ and $\mathrm{ND}\left(D'\right) \subseteq D'$, $\phi \cdot \mathrm{ND}\left(D'\right) \subseteq D$. Then eq. (D.218) holds by applying lemma D.33 with $A := D', B' := D_{\mathrm{sub}}, B := D, C := \mathrm{RSD}(s)$, and the previously defined Z which we showed satisfies $\mathrm{ND}(C) \subseteq Z \subseteq C$. Furthermore, involution ϕ' satisfies $\phi' \cdot B^* = \phi' \cdot \left(Z \setminus (B \setminus B') \right) = Z \setminus (B \setminus B') = B^*$ by eq. (D.211).

When $\mathrm{RSD}_{\mathrm{nd}}(s) \cap (D \setminus D_{\mathrm{sub}})$ is non-empty, since $B' \subseteq C$ by assumption, lemma D.33 also shows that eq. (D.218) is strict for all $\mathcal{D}_{X\text{-IID}} \in \mathfrak{D}_{\text{C/B/IID}}$, and that $\mathbb{P}_{\mathcal{D}_{\text{any}}}(D', \text{average}) \not\geq_{\text{most: } \mathfrak{D}_{\text{any}}} \mathbb{P}_{\mathcal{D}_{\text{any}}}(D, \text{average})$. ☐

Proposition D.57 (RSD properties). *Let* $\mathbf{d} \in \mathrm{RSD}(s)$. $\mathbf{d}$ *is element-wise non-negative and* $\|\mathbf{d}\|_1 = 1$.

Proof. $\mathbf{d}$ has non-negative elements because it equals the limit of $\lim_{\gamma \to 1}(1-\gamma)\mathbf{f}(\gamma)$, whose elements are non-negative by proposition D.8 item 1.

$$\|\mathbf{d}\|_1 = \left\| \lim_{\gamma \to 1} (1-\gamma)\mathbf{f}(\gamma) \right\|_1 \tag{D.220}$$
$$= \lim_{\gamma \to 1} (1-\gamma) \left\| \mathbf{f}(\gamma) \right\|_1 \tag{D.221}$$
$$= 1. \tag{D.222}$$

Equation (D.220) follows because the definition of RSDs (definition 5.26) ensures that $\exists \mathbf{f} \in \mathcal{F}(s) : \lim_{\gamma \to 1}(1-\gamma)\mathbf{f}(\gamma) = \mathbf{d}$. Equation (D.221) follows because $\|\cdot\|_1$ is a continuous function. Equation (D.222) follows because $\left\|\mathbf{f}(\gamma)\right\|_1 = \frac{1}{1-\gamma}$ by proposition D.8 item 2. ☐

Lemma D.58 (When reachable with probability 1, 1-cycles induce non-dominated RSDs). *If* $\mathbf{e}_{s'} \in \mathrm{RSD}(s)$, *then* $\mathbf{e}_{s'} \in \mathrm{RSD}_{\mathrm{nd}}(s)$.

Proof. If $\mathbf{d} \in \mathrm{RSD}(s)$ is distinct from $\mathbf{e}_{s'}$, then $\|\mathbf{d}\|_1 = 1$ and $\mathbf{d}$ has non-negative entries by proposition D.57. Since $\mathbf{d}$ is distinct from $\mathbf{e}_{s'}$, then its entry for index s' must be strictly less than 1: $\mathbf{d}^\top \mathbf{e}_{s'} < 1 = \mathbf{e}_{s'}^\top \mathbf{e}_{s'}$. Therefore, $\mathbf{e}_{s'} \in \mathrm{RSD}(s)$ is strictly optimal for the *reward function* $\mathbf{r} \coloneqq \mathbf{e}_{s'}$, and so $\mathbf{e}_{s'} \in \mathrm{RSD}_{\mathrm{nd}}(s)$. ☐

Corollary 5.30 (Average-optimal policies tend not to end up in any given 1-cycle). *Suppose* $\mathbf{e}_{s_x}, \mathbf{e}_{s'} \in \mathrm{RSD}(s)$ *are distinct. Then*

$$\mathbb{P}_{\mathcal{D}_{any}}\left(\mathrm{RSD}(s) \setminus \{\mathbf{e}_{s_x}\}, \text{average}\right) \geq_{\text{most}} \mathbb{P}_{\mathcal{D}_{any}}\left(\{\mathbf{e}_{s_x}\}, \text{average}\right). \tag{5.11}$$

If there is a third $\mathbf{e}_{s''} \in \mathrm{RSD}(s)$, *the converse* $\leq_{\text{most}}$ *statement does not hold.*

Proof. Suppose $\mathbf{e}_{s_x}, \mathbf{e}_{s'} \in \mathrm{RSD}(s)$ are distinct. Let $\phi \coloneqq (s_x \ \ s'), D' \coloneqq \{\mathbf{e}_{s_x}\}, D \coloneqq \mathrm{RSD}(s) \setminus \{\mathbf{e}_{s_x}\}$. $\phi \cdot D' = \{\mathbf{e}_{s'}\} \subseteq \mathrm{RSD}(s) \setminus \{\mathbf{e}_{s_x}\} =: D$ since $s_x \neq s'$. $D' \cup D = \mathrm{RSD}(s)$ and $\mathrm{RSD}_{\mathrm{nd}}(s) \setminus (D' \cup D) = \mathrm{RSD}_{\mathrm{nd}}(s) \setminus \mathrm{RSD}(s) = \emptyset$ trivially have pairwise orthogonal vector elements. Then apply theorem 5.29 to conclude that $\mathbb{P}_{\mathcal{D}_{\mathrm{any}}}\big(\{\mathbf{e}_{s_x}\}, \mathrm{average}\big) \leq_{\mathrm{most:}\ \mathfrak{D}_{\mathrm{any}}} \mathbb{P}_{\mathcal{D}_{\mathrm{any}}}\big(\mathrm{RSD}(s) \setminus \{\mathbf{e}_{s_x}\}, \mathrm{average}\big)$.

Suppose there exists another $\mathbf{e}_{s''} \in \mathrm{RSD}(s)$. By lemma D.58, $\mathbf{e}_{s''} \in \mathrm{RSD}_{\mathrm{nd}}(s)$. Furthermore, since $s'' \notin \big\{s', s_x\big\}$, $\mathbf{e}_{s''} \in \big(\mathrm{RSD}(s) \setminus \{\mathbf{e}_{s_x}\}\big) \setminus \{\mathbf{e}_{s'}\} = D \setminus \phi \cdot D'$. Therefore, $\mathbf{e}_{s''} \in \mathrm{RSD}_{\mathrm{nd}}(s) \cap \big(D \setminus \phi \cdot D'\big)$. Then apply the second condition of theorem 5.29 to conclude that $\mathbb{P}_{\mathcal{D}_{\mathrm{any}}}\big(\{\mathbf{e}_{s_x}\}, \mathrm{average}\big) \not\geq_{\mathrm{most:}\ \mathfrak{D}_{\mathrm{bound}}} \mathbb{P}_{\mathcal{D}_{\mathrm{any}}}\big(\mathrm{RSD}(s) \setminus \{\mathbf{e}_{s_x}\}, \mathrm{average}\big)$. $\square$

E

Parametrically Retargetable Decision-Makers Tend To Seek Power

E.1 Retargetability over outcome lotteries

Suppose we are interested in d outcomes. Each outcome could be the visitation of an MDP state, or a trajectory, or the receipt of a physical item. In the card example of section 6.2, $d = 3$ playing cards. The agent can induce each outcome with probability 1, so let $\mathbf{e}_o \in \mathbb{R}^3$ be the standard basis vector with probability 1 on outcome o and 0 elsewhere. Then the agent chooses among outcome lotteries $C_{\text{cards}} := \left\{\mathbf{e}_{\spadesuit\text{B}}, \mathbf{e}_{\heartsuit\text{B}}, \mathbf{e}_{\diamondsuit\text{A}}\right\}$, which we partition into $A_{\text{cards}} := \left\{\mathbf{e}_{\spadesuit\text{B}}\right\}$ and $B_{\text{cards}} := \{\mathbf{e}_{\heartsuit\text{B}}, \mathbf{e}_{\diamondsuit\text{A}}\}$.

Definition E.1 (Outcome lotteries)**.** A unit vector $\mathbf{x} \in \mathbb{R}^d$ with non-negative entries is an *outcome lottery*.[1]

[1] Our results on outcome lotteries hold for generic $\mathbf{x}' \in \mathbb{R}^d$, but we find it conceptually helpful to

Many decisions are made consequentially: based on the consequences of the decision, on what outcomes are brought about by an act. For example, in a deterministic setting like Pac-Man, a policy induces a trajectory. A reward function and discount rate tuple (R, γ) assigns a *return* to each state trajectory $\tau = s_0, s_1, \ldots$: $G(\tau) = \sum_{i=0}^{\infty} \gamma^i R(s_i)$. The relevant outcome lottery is the discounted visit distribution over future states in the Pac-Man video game, and policies are optimal or not depending on which outcome lottery is induced by the policy.

Definition E.2 (Optimality indicator function)**.** Let $X, C \subsetneq \mathbb{R}^d$ be finite, and let $\mathbf{u} \in \mathbb{R}^d$. Optimal $\left(X \mid C, \mathbf{u}\right)$ returns 1 if $\max_{\mathbf{x} \in X} \mathbf{x}^\top \mathbf{u} \geq \max_{\mathbf{c} \in C} \mathbf{c}^\top \mathbf{u}$, and 0 otherwise.

We consider decision-making procedures which take in a targeting parameter $\mathbf{u}$. For example, the column headers of table E.1a show the 6 permutations of the utility function $u(\spadesuit^{\text{B}}) := 10, u(\heartsuit^{\text{B}}) := 5, u(\diamondsuit^{\text{A}}) := 0$, representable as a vector $\mathbf{u} \in \mathbb{R}^3$.

$\mathbf{u}$ can be permuted as follows. The outcome permutation $\phi \in S_d$ inducing an $d \times d$ permutation matrix $\mathbf{P}_\phi$ in row representation: $(\mathbf{P}_\phi)_{ij} = 1$ if $i = \phi(j)$ and 0 otherwise. Table E.1a shows that for a given utility function, $\frac{2}{3}$ of its orbit agrees that B_{cards} is strictly optimal over A_{cards}.

Orbit-level incentives occur when an inequality holds for most permuted parameter choices $\mathbf{u}'$. Table E.1a demonstrates an application of Turner et al. [99]'s results: Optimal decision-making induces orbit-level incentives for choosing outcomes in B_{cards} over outcomes in A_{cards}.

Furthermore, Turner et al. [99] conjectured that "larger" B_{cards} will imply stronger orbit-level tendencies: If going right leads to 500 times as many options as going left, then right is better than left for at least 500 times as many reward functions for which the opposite is true. We prove this conjecture with theorem E.46 in appendix E.4.

However, orbit-level incentives do not require optimality. One clue is that the same results hold for anti-optimal agents, since anti-optimality/utility minimization of $\mathbf{u}$ is equivalent to maximizing $-\mathbf{u}$. Table E.1b illustrates that the same orbit guarantees hold in this case.

consider the non-negative unit vector case.

Definition E.3 (Anti-optimality indicator function)**.** Let $X, C \subsetneq \mathbb{R}^d$ be finite, and let $\mathbf{u} \in \mathbb{R}^d$. AntiOpt $\left(X \mid C, \mathbf{u}\right)$ returns 1 if $\min_{\mathbf{x}\in X} \mathbf{x}^\top \mathbf{u} \leq \min_{\mathbf{c}\in C} \mathbf{c}^\top \mathbf{u}$, and 0 otherwise.

Stepping beyond expected utility maximization/minimization, Boltzmann-rational decision-making selects outcome lotteries proportional to the exponential of their expected utility.

Definition E.4 (Boltzmann rationality [7])**.** For $X \subseteq C$ and temperature $T > 0$, let

$$\text{Boltzmann}_T \left(X \mid C, \mathbf{u}\right) \coloneqq \frac{\sum_{\mathbf{x}\in X} e^{T^{-1}\mathbf{x}^\top \mathbf{u}}}{\sum_{\mathbf{c}\in C} e^{T^{-1}\mathbf{c}^\top \mathbf{u}}}$$

be the probability that some element of X is Boltzmann-rational.

Lastly, orbit-level tendencies occur even under decision-making procedures which partially ignore expected utility and which "don't optimize too hard." Satisficing agents randomly choose an outcome lottery with expected utility exceeding some threshold. Table E.1d demonstrates that satisficing induces orbit-level tendencies.

Definition E.5 (Satisficing)**.** Let $t \in \mathbb{R}$, let $X \subseteq C \subsetneq \mathbb{R}^d$ be finite. $\text{Satisfice}_t \left(X, C \mid \mathbf{u}\right) \coloneqq \frac{\left|X\cap\left\{\mathbf{c}\in C \mid \mathbf{c}^\top \mathbf{u} \geq t\right\}\right|}{\left|\left\{\mathbf{c}\in C \mid \mathbf{c}^\top \mathbf{u} \geq t\right\}\right|}$ is the fraction of X whose value exceeds threshold t. $\text{Satisfice}_t \left(X, C \mid \mathbf{u}\right)$ evaluates to 0 the denominator equals 0.

For each table, two-thirds of the utility permutations (columns) assign strictly larger values (shaded dark gray) to an element of $B_{\text{cards}} \coloneqq \{\mathbf{e}_{\heartsuit \text{B}}, \mathbf{e}_{\diamondsuit \text{A}}\}$ than to an element of $A_{\text{cards}} \coloneqq \left\{\mathbf{e}_{\spadesuit \text{B}}\right\}$. For optimal, anti-optimal, Boltzmann-rational, and satisficing agents, proposition E.11 proves that these tendencies hold for all targeting parameter orbits.

E.1.1 A range of decision-making functions are retargetable

In MDPs, Turner et al. [99] consider *state visitation distributions* which record the total discounted time steps spent in each environment state, given that the agent follows some policy π from an initial state s. These visitation distributions are one kind of outcome lottery, with $d = |\mathcal{S}|$ the number of MDP states.

In general, we suppose the agent has an objective function $\mathbf{u} \in \mathbb{R}^d$ which maps outcomes to real numbers. In Turner et al. [99], $\mathbf{u}$ was a state-based reward function (and so the outcomes were *states*). However, we need not restrict ourselves to the MDP setting.

To state our key results, we define several technical concepts which we informally used when reasoning about $A_{\text{cards}} := \{\mathbf{e}_{\spadesuit^{\text{B}}}\}$ and $B_{\text{cards}} := \{\mathbf{e}_{\heartsuit^{\text{B}}}, \mathbf{e}_{\diamondsuit^{\text{A}}}\}$.

Definition E.6 (Similarity of vector sets)**.** For $\phi \in S_d$ and $X \subseteq \mathbb{R}^d$, $\phi \cdot X := \{\mathbf{P}_\phi \mathbf{x} \mid \mathbf{x} \in X\}$. $X' \subseteq \mathbb{R}^{|\mathcal{S}|}$ *is similar to* X when $\exists \phi : \phi \cdot X' = X$. ϕ is an *involution* if $\phi = \phi^{-1}$ (it either transposes states, or fixes them). X *contains a copy of* X' when X' is similar to a subset of X via an involution ϕ.

Definition E.7 (Containment of set copies)**.** Let n be a positive integer, and let $A, B \subseteq \mathbb{R}^d$. We say that B *contains* n *copies of* A when there exist involutions $\phi_1, \ldots, \phi_n \in S_d$ such that $\forall i : \phi_i \cdot A =: B_i \subseteq B$ and $\forall j \neq i : \phi_i \cdot B_j = B_j$.[2]

$B_{\text{cards}} := \{\mathbf{e}_{\heartsuit^{\text{B}}}, \mathbf{e}_{\diamondsuit^{\text{A}}}\}$ contains two copies of $A_{\text{cards}} := \{\mathbf{e}_{\spadesuit^{\text{B}}}\}$ via $\phi_1 := \spadesuit^{\text{B}} \leftrightarrow \heartsuit^{\text{B}}$ and $\phi_2 := \spadesuit^{\text{B}} \leftrightarrow \diamondsuit^{\text{A}}$.

Definition E.8 (Targeting parameter distribution assumptions)**.** Results with $\mathcal{D}_{\text{any}}$ hold for any probability distribution over $\mathbb{R}^d$. Let $\mathfrak{D}_{\text{any}} := \Delta(\mathbb{R}^d)$. For a function $f : \mathbb{R}^d \mapsto \mathbb{R}$, we write $f(\mathcal{D}_{\text{any}})$ as shorthand for $\mathbb{E}_{\mathbf{u} \sim \mathcal{D}_{\text{any}}}\left[f(\mathbf{u})\right]$.

The symmetry group on d elements, S_d, acts on the set of probability distributions over $\mathbb{R}^d$.

Definition E.9 (Pushforward distribution of a permutation [99])**.** Let $\phi \in S_d$. $\phi \cdot \mathcal{D}_{\text{any}}$ is the pushforward distribution induced by applying the random vector $p(\mathbf{u}) := \mathbf{P}_\phi \mathbf{u}$ to $\mathcal{D}_{\text{any}}$.

Definition E.10 (Orbit of a probability distribution [99])**.** The *orbit* of $\mathcal{D}_{\text{any}}$ under the symmetric group S_d is $S_d \cdot \mathcal{D}_{\text{any}} := \{\phi \cdot \mathcal{D}_{\text{any}} \mid \phi \in S_d\}$.

Because B_{cards} contains 2 copies of A_{cards}, there are "at least two times as many ways" for B to be optimal, than for A to be optimal. Similarly, B is "at least two times as likely"

[2] Technically, definition E.7 implies that A contains n copies of A holds for all n, via n applications of the identity permutation. For our purposes, this provides greater generality, as all of the relevant results still hold. Enforcing pairwise disjointness of the B_i would handle these issues, but would narrow our results to not apply *e.g.* when the B_i share a constant vector.

to contain an anti-rational outcome lottery for generic utility functions. As demonstrated by table E.1, the key idea is that "larger" sets (a set B containing several *copies* of set A) are more likely to be chosen under a wide range of decision-making criteria.

Proposition E.11 (Orbit incentives for different rationalities)**.** *Let $A, B \subseteq C \subsetneq \mathbb{R}^d$ be finite, such that B contains n copies of A via involutions ϕ_i such that $\phi_i \cdot C = C$.*

1. ***Rational choice [99].***

$$\text{Optimal}\left(B \mid C, \mathcal{D}_{any}\right) \geq^n_{\text{most: } \mathfrak{D}_{any}} \text{Optimal}\left(A \mid C, \mathcal{D}_{any}\right).$$

2. ***Uniformly randomly choosing an optimal lottery.*** *For $X \subseteq C$, let*

$$\text{FracOptimal}\left(X \mid C, \mathbf{u}\right) := \frac{\left|\left\{\arg\max_{\mathbf{c} \in C} \mathbf{c}^\top \mathbf{u}\right\} \cap X\right|}{\left|\left\{\arg\max_{\mathbf{c} \in C} \mathbf{c}^\top \mathbf{u}\right\}\right|}.$$

Then $\text{FracOptimal}\left(B \mid C, \mathcal{D}_{any}\right) \geq^n_{\text{most: } \mathfrak{D}_{any}} \text{FracOptimal}\left(A \mid C, \mathcal{D}_{any}\right)$.

3. ***Anti-rational choice.*** $\text{AntiOpt}\left(B \mid C, \mathcal{D}_{any}\right) \geq^n_{\text{most: } \mathfrak{D}_{any}} \text{AntiOpt}\left(A \mid C, \mathcal{D}_{any}\right)$.

4. ***Boltzmann rationality.***

$$\text{Boltzmann}_T\left(B \mid C, \mathcal{D}_{any}\right) \geq^n_{\text{most: } \mathfrak{D}_{any}} \text{Boltzmann}_T\left(A \mid C, \mathcal{D}_{any}\right).$$

5. ***Uniformly randomly drawing k outcome lotteries and choosing the best.*** *For $X \subseteq C$, $\mathbf{u} \in \mathbb{R}^d$, and $k \geq 1$, let*

$$\textit{best-of-k}(X, C \mid \mathbf{u}) := \mathop{\mathbb{E}}_{\mathbf{a}_1, \ldots, \mathbf{a}_k \sim \textit{unif}(C)}\left[\text{FracOptimal}\left(X \cap \{\mathbf{a}_1, \ldots, \mathbf{a}_k\} \mid \{\mathbf{a}_1, \ldots, \mathbf{a}_k\}, \mathbf{u}\right)\right].$$

Then $\textit{best-of-k}(B \mid C, \mathcal{D}_{any}) \geq^n_{\text{most: } \mathfrak{D}_{any}} \textit{best-of-k}(A \mid C, \mathcal{D}_{any})$.

6. ***Satisficing [86].*** $\text{Satisfice}_t\left(B \mid C, \mathcal{D}_{any}\right) \geq^n_{\text{most: } \mathfrak{D}_{any}} \text{Satisfice}_t\left(A \mid C, \mathcal{D}_{any}\right)$.

7. ***Quantilizing over outcome lotteries [93].*** *Let P be the uniform probability*

distribution over C. For $X \subseteq C$, $\mathbf{u} \in \mathbb{R}^d$, and $q \in (0,1]$, let $Q_{q,P}(X \mid C, \mathbf{u})$ (definition E.25) return the probability that an outcome lottery in X is drawn from the top q-quantile of P, sorted by expected utility under $\mathbf{u}$. Then $Q_{q,P}(B \mid C, \mathbf{u}) \geq^n_{\text{most: } \mathbb{R}^d} Q_{q,P}(A \mid C, \mathbf{u})$.

One highly retargetable class of decision-making functions are those which only account for the expected utilities of available choices.

Definition E.12 (EU-determined functions)**.** Let $\mathcal{P}\left(\mathbb{R}^d\right)$ be the power set of $\mathbb{R}^d$, and let $f : \prod_{i=1}^m \mathcal{P}\left(\mathbb{R}^d\right) \times \mathbb{R}^d \to \mathbb{R}$. f is an *EU-determined function* if there exists a family of functions $\{g^{\omega_1,\ldots,\omega_m}\}$ such that

$$f(X_1, \ldots, X_m \mid \mathbf{u}) = g^{|X_1|,\ldots,|X_m|}\left(\left[\mathbf{x}_1^\top \mathbf{u}\right]_{\mathbf{x}_1 \in X_1}, \ldots, \left[\mathbf{x}_m^\top \mathbf{u}\right]_{\mathbf{x}_m \in X_m}\right), \tag{E.1}$$

where $[r_i]$ is the multiset of its elements r_i.

For example, let $X \subseteq C \subsetneq \mathbb{R}^d$ be finite, and consider utility function $\mathbf{u} \in \mathbb{R}^d$. A Boltzmann-rational agent is more likely to select outcome lotteries with greater expected utility. Formally, $\text{Boltzmann}_T\left(X \mid C, \mathbf{u}\right) \coloneqq \sum_{\mathbf{x} \in X} \frac{e^{T \cdot \mathbf{x}^\top \mathbf{u}}}{\sum_{\mathbf{c} \in C} e^{T \cdot \mathbf{c}^\top \mathbf{u}}}$ depends only on the expected utility of outcome lotteries in X, relative to the expected utility of all outcome lotteries in C. Therefore, Boltzmann_T is a function of expected utilities. This is *why* Boltzmann_T satisfies the $\geq^n_{\text{most: } \mathfrak{D}_{\text{any}}}$ relation.

Theorem E.13 (Orbit tendencies occur for EU-determined decision-making functions)**.** *Let $A, B, C \subseteq \mathbb{R}^d$ be such that B contains n copies of A via ϕ_i such that $\phi_i \cdot C = C$. Let $h : \prod_{i=1}^2 \mathcal{P}\left(\mathbb{R}^d\right) \times \mathbb{R}^d \to \mathbb{R}$ be an EU-determined function, and let $p(X \mid \mathbf{u}) \coloneqq h(X, C \mid \mathbf{u})$. Suppose that p returns a probability of selecting an element of X from C. Then $p(B \mid \mathbf{u}) \geq^n_{\text{most: } \mathbb{R}^d} p(A \mid \mathbf{u})$.*

The key takeaway is that decision rules determined by expected utility are highly retargetable. By changing the targeting parameter hyperparameter, the decision-making procedure can be flexibly retargeted to choose elements of "larger" sets (in terms of set copies via definition E.7). Less abstractly, for many agent rationalities—ways of making decisions over outcome lotteries—it is generally the case that larger sets will more often be chosen over smaller sets.

For example, consider a Pac-Man playing agent choosing which environmental state cycle it should end up in. Turner et al. [99] show that for most reward functions, average-reward maximizing agents will tend to stay alive so that they can reach a wider range of environmental cycles. However, our results show that average-reward *minimizing* agents also exhibit this tendency, as do Boltzmann-rational agents who assign greater probability to higher-reward cycles. Any EU-based cycle selection method will—for most reward functions—tend to choose cycles which require Pac-Man to stay alive (at first).

E.2 Theoretical results

Definition 6.2 (Inequalities which hold for most orbit elements)**.** Suppose Θ is a subset of a set acted on by S_d, the symmetric group on d elements. Let $f_1, f_2 : \Theta \to \mathbb{R}$ and let $n \geq 1$. We write $f_1(\theta) \geq^n_{\text{most: }\Theta} f_2(\theta)$ when, for *all* $\theta \in \Theta$, the following cardinality inequality holds:

$$\left|\left\{\theta' \in \text{Orbit}|_\Theta\left(\theta\right) \mid f_1(\theta') > f_2(\theta')\right\}\right| \geq n\left|\left\{\theta' \in \text{Orbit}|_\Theta\left(\theta\right) \mid f_1(\theta') < f_2(\theta')\right\}\right|. \tag{6.2}$$

Lemma E.14 (Limited transitivity of $\geq_{\text{most}}$)**.** *Let $f_0, f_1, f_2, f_3 : \Theta \to \mathbb{R}$, and suppose Θ is a subset of a set acted on by S_d. Suppose that $f_1(\theta) \geq^n_{\text{most: }\Theta} f_2(\theta)$ and $\forall \theta \in \Theta$: $f_0(\theta) \geq f_1(\theta)$ and $f_2(\theta) \geq f_3(\theta)$. Then $f_0(\theta) \geq^n_{\text{most: }\Theta} f_3(\theta)$.*

Proof. Let $\theta \in \Theta$ and let $\text{Orbit}|_{\Theta, f_a > f_b}\left(\theta\right) \coloneqq \left\{\theta' \in \text{Orbit}|_\Theta\left(\theta\right) \mid f_a(\theta') > f_b(\theta')\right\}$.

$$\begin{align}
\left|\text{Orbit}|_{\Theta, f_0 > f_3}\left(\theta\right)\right| &\geq \left|\text{Orbit}|_{\Theta, f_1 > f_2}\left(\theta\right)\right| \tag{E.2}\\
&\geq n\left|\text{Orbit}|_{\Theta, f_2 > f_1}\left(\theta\right)\right| \tag{E.3}\\
&\geq n\left|\text{Orbit}|_{\Theta, f_3 > f_0}\left(\theta\right)\right|. \tag{E.4}
\end{align}$$

For all $\theta' \in \text{Orbit}|_{\Theta, f_1 > f_2}\left(\theta\right)$,

$$f_0(\theta') \geq f_1(\theta') > f_2(\theta') \geq f_3(\theta')$$

by assumption, and so

$$\mathrm{Orbit}|_{\Theta,f_1>f_2}(\theta) \subseteq \mathrm{Orbit}|_{\Theta,f_0>f_3}(\theta).$$

Therefore, eq. (E.2) follows. By assumption,

$$\left|\mathrm{Orbit}|_{\Theta,f_1>f_2}(\theta)\right| \geq n\left|\mathrm{Orbit}|_{\Theta,f_2>f_1}(\theta)\right|;$$

eq. (E.3) follows. For all $\theta' \in \mathrm{Orbit}|_{\Theta,f_2>f_1}(\theta)$, our assumptions on f_0 and f_3 ensure that

$$f_0(\theta') \leq f_1(\theta') < f_3(\theta') \leq f_2(\theta'),$$

so

$$\mathrm{Orbit}|_{\Theta,f_3>f_0}(\theta) \subseteq \mathrm{Orbit}|_{\Theta,f_2>f_1}(\theta).$$

Then eq. (E.4) follows. By eq. (E.4), $f_0(\theta) \geq^n_{\text{most: }\Theta} f_3(\theta)$. □

Lemma E.15 (Order inversion for $\geq_{\text{most}}$)**.** *Let $f_1, f_2 : \Theta \to \mathbb{R}$, and suppose Θ is a subset of a set acted on by S_d. Suppose that $f_1(\theta) \geq^n_{\text{most: }\Theta} f_2(\theta)$. Then $-f_2(\theta) \geq^n_{\text{most: }\Theta} -f_1(\theta)$.*

Proof. By definition E.10, $f_1(\theta) \geq^n_{\text{most: }\Theta} f_2(\theta)$ means that

$$\left|\left\{\theta' \in \mathrm{Orbit}|_{\Theta}(\theta) \mid f_1(\theta') > f_2(\theta')\right\}\right| \geq n\left|\left\{\theta' \in \mathrm{Orbit}|_{\Theta}(\theta) \mid f_1(\theta') < f_2(\theta')\right\}\right| \tag{E.5}$$

$$\left|\left\{\theta' \in \mathrm{Orbit}|_{\Theta}(\theta) \mid -f_2(\theta') > -f_1(\theta')\right\}\right| \geq n\left|\left\{\theta' \in \mathrm{Orbit}|_{\Theta}(\theta) \mid -f_2(\theta') < -f_1(\theta')\right\}\right|. \tag{E.6}$$

Then $-f_2(\theta) \geq^n_{\text{most: }\Theta} -f_1(\theta)$. □

Remark. Lemma E.16 generalizes Turner et al. [99]'s lemma D.2.

Lemma E.16 (Orbital fraction which agrees on (weak) inequality)**.** *Suppose $f_1, f_2 : \Theta \to \mathbb{R}$ are such that $f_1(\theta) \geq^n_{\text{most: }\Theta} f_2(\theta)$. Then for all $\theta \in \Theta$, $\frac{\left|\left\{\theta' \in (S_d \cdot \theta)\cap\Theta \mid f_1(\theta') \geq f_2(\theta')\right\}\right|}{\left|(S_d\cdot\theta)\cap\Theta\right|} \geq \dfrac{n}{n+1}$.*

Proof. All $\theta' \in (S_d \cdot \theta) \cap \Theta$ such that $f_1(\theta') = f_2(\theta')$ satisfy $f_1(\theta') \geq f_2(\theta')$. Otherwise, consider the $\theta' \in (S_d \cdot \theta) \cap \Theta$ such that $f_1(\theta') \neq f_2(\theta')$. By assumption, at least $\frac{n}{n+1}$ of these θ' satisfy $f_1(\theta') > f_2(\theta')$, in which case $f_1(\theta') \geq f_2(\theta')$. Then the desired inequality follows. □

E.2.1 General results on retargetable functions

Definition E.17 (Functions which are increasing under joint permutation)**.** Suppose that S_d acts on sets $\mathbf{E}_1, \ldots, \mathbf{E}_m$, and let $f : \prod_{i=1}^m \mathbf{E}_i \to \mathbb{R}$. $f(X_1, \ldots, X_m)$ is *increasing under joint permutation by* $P \subseteq S_d$ when $\forall \phi \in P : f(X_1, \ldots, X_m) \leq f(\phi \cdot X_1, \ldots, \phi \cdot X_m)$. If equality always holds, then $f(X_1, \ldots, X_m)$ is *invariant under joint permutation by* P.

Lemma E.18 (Expectations of joint-permutation-increasing functions are also joint-permutation-increasing)**.** *For* $\mathbf{E}$ *which is a subset of a set acted on by* S_d*, let* $f : \mathbf{E} \times \mathbb{R}^d \to \mathbb{R}$ *be a bounded function which is measurable on its second argument, and let* $P \subseteq S_d$*. Then if* $f(X \mid \mathbf{u})$ *is increasing under joint permutation by* P*, then* $f'(X \mid \mathcal{D}_{any}) := \mathbb{E}_{\mathbf{u} \sim \mathcal{D}_{any}}\left[f(X \mid \mathbf{u})\right]$ *is increasing under joint permutation by* P*. If* f *is* invariant *under joint permutation by* P*, then so is* f'*.*

Proof. Let distribution $\mathcal{D}_{\text{any}}$ have probability measure F, and let $\phi \cdot \mathcal{D}_{\text{any}}$ have probability measure F_ϕ.

$$
\begin{aligned}
f\left(X \mid \mathcal{D}_{\text{any}}\right) &:= \mathbb{E}_{\mathbf{u} \sim \mathcal{D}_{\text{any}}}\left[f(X \mid \mathbf{u})\right] && \text{(E.7)}\\
&:= \int_{\mathbb{R}^d} f(X \mid \mathbf{u}) \,\mathrm{d}F(\mathbf{u}) && \text{(E.8)}\\
&\leq \int_{\mathbb{R}^d} f(\phi \cdot X \mid \mathbf{P}_\phi \mathbf{u}) \,\mathrm{d}F(\mathbf{u}) && \text{(E.9)}\\
&= \int_{\mathbb{R}^d} f(\phi \cdot X \mid \mathbf{u}') \left|\det \mathbf{P}_\phi\right| \mathrm{d}F_\phi(\mathbf{u}') && \text{(E.10)}\\
&= \int_{\mathbb{R}^d} f(\phi \cdot X \mid \mathbf{u}') \,\mathrm{d}F_\phi(\mathbf{u}') && \text{(E.11)}\\
&=: f'\left(\phi \cdot X \mid \phi \cdot \mathcal{D}_{\text{any}}\right). && \text{(E.12)}
\end{aligned}
$$

Equation (E.9) holds by assumption on f: $f(X \mid \mathbf{u}) \leq f(\phi \cdot X \mid \mathbf{P}_\phi \mathbf{u})$. Furthermore,

$f(\phi \cdot X \mid \cdot)$ is still measurable, and so the inequality holds. Equation (E.10) follows by the definition of F_ϕ (definition 5.19) and by substituting $\mathbf{r}' := \mathbf{P}_\phi \mathbf{r}$. Equation (E.11) follows from the fact that all permutation matrices have unitary determinant. $\square$

Lemma E.19 (Closure of orbit incentives under increasing functions)**.** *Suppose that S_d acts on sets $\mathbf{E}_1, \ldots, \mathbf{E}_m$ (with $\mathbf{E}_1$ being a poset), and let $P \subseteq S_d$. Let $f_1, \ldots, f_n : \prod_{i=1}^m \mathbf{E}_i \to \mathbb{R}$ be increasing under joint permutation by P on input $(X_1, \ldots, X_m)$, and suppose the f_i are order-preserving with respect to $\preceq_{\mathbf{E}_1}$. Let $g : \prod_{j=1}^n \mathbb{R} \to \mathbb{R}$ be monotonically increasing on each argument. Then*

$$f\left(X_1, \ldots, X_m\right) := g\left(f_1\left(X_1, \ldots, X_m\right), \ldots, f_n\left(X_1, \ldots, X_m\right)\right) \tag{E.13}$$

is increasing under joint permutation by P and order-preserving with respect to set inclusion on its first argument. Furthermore, if the f_i are invariant *under joint permutation by P, then so is f.*

Proof. Let $\phi \in P$.

$$\begin{aligned} f\left(X_1, \ldots, X_m\right) &:= g\left(f_1\left(X_1, \ldots, X_m\right), \ldots, f_n\left(X_1, \ldots, X_m\right)\right) && \text{(E.14)} \\ &\leq g\left(f_1\left(\phi \cdot X_1, \ldots, \phi \cdot X_m\right), \ldots, f_n\left(\phi \cdot X_1, \ldots, \phi \cdot X_m\right)\right) && \text{(E.15)} \\ &=: f\left(\phi \cdot X_1, \ldots, \phi \cdot X_m\right). && \text{(E.16)} \end{aligned}$$

Equation (E.15) follows because we assumed that $f_i\left(X_1, \ldots, X_m\right) \leq f_i\left(\phi \cdot X_1, \ldots, \phi \cdot X_m\right)$, and because g is monotonically increasing on each argument. If the f_i are all invariant, then eq. (E.15) is an equality.

Similarly, suppose $X_1' \preceq_{\mathbf{E}_1} X_1$. The f_i are order-preserving on the first argument, and g is monotonically increasing on each argument. Then $f\left(X_1', \ldots, X_m\right) \leq f\left(X_1, \ldots, X_m\right)$. This shows that f is order-preserving on its first argument. $\square$

Remark. g could take the convex combination of its arguments, or multiply two f_i together and add them to a third f_3.

Definition 6.5 (Multiply retargetable function)**.** Let Θ be a subset of a set acted on by S_d, and let $f : \{A, B\} \times \Theta \to \mathbb{R}$.

f is a $(\Theta, A \overset{n}{\to} B)$*-retargetable function* when, for each $\theta \in \Theta$, we can choose permutations $\phi_1, \dots, \phi_n \in S_d$ which satisfy the following conditions: Consider any $\theta^A \in \text{Orbit}|_{\Theta,A>B}(\theta) := \left\{\theta^* \in \text{Orbit}|_{\Theta}(\theta) \mid f(A \mid \theta^*) > f(B \mid \theta^*)\right\}$.

1. **Retargetable via n permutations.** $\forall i = 1, \dots, n : f\left(A \mid \phi_i \cdot \theta^A\right) < f\left(B \mid \phi_i \cdot \theta^A\right)$.
2. **Parameter permutation is allowed by Θ.** $\forall i : \phi_i \cdot \theta^A \in \Theta$.
3. **Permuted parameters are distinct.** $\forall i \neq j, \theta' \in \text{Orbit}|_{\Theta,A>B}(\theta) : \phi_i \cdot \theta^A \neq \phi_j \cdot \theta'$.

Theorem 6.6 (Multiply retargetable functions have orbit-level tendencies)**.**

If f is $(\Theta, A \overset{n}{\to} B)$-retargetable, then $f(B \mid \theta) \geq^n_{\text{most: }\Theta} f(A \mid \theta)$.

Proof. Let $\theta \in \Theta$, and let $\phi_i \cdot \text{Orbit}|_{\Theta,A>B}(\theta) := \left\{\phi_i \cdot \theta^A \mid \theta^A \in \text{Orbit}|_{\Theta,A>B}(\theta)\right\}$.

$$\left|\text{Orbit}|_{\Theta,B>A}(\theta)]\right| \geq \left|\bigcup_{i=1}^{n} \phi_i \cdot \text{Orbit}|_{\Theta,A>B}(\theta)\right| \tag{E.17}$$

$$= \sum_{i=1}^{n} \left|\phi_i \cdot \text{Orbit}|_{\Theta,A>B}(\theta)\right| \tag{E.18}$$

$$= n \left|\text{Orbit}|_{\Theta,A>B}(\theta)\right|. \tag{E.19}$$

By item 1 and item 2, $\phi_i \cdot \phi_i \cdot \text{Orbit}|_{\Theta,A>B}(\theta) \subseteq \phi_i \cdot \text{Orbit}|_{\Theta,B>A}(\theta)]$ for all i. Therefore, eq. (E.17) holds. Equation (E.18) follows by the assumption that parameters are distinct, and so therefore the cosets $\phi_i \cdot \text{Orbit}|_{\Theta,A>B}(\theta)$ and $\phi_j \cdot \text{Orbit}|_{\Theta,A>B}(\theta)$ are pairwise disjoint for $i \neq j$. Equation (E.19) follows because each ϕ_i acts injectively on orbit elements.

Letting $f_A(\theta) := f(A \mid \theta)$ and $f_B(\theta) := f(B \mid \theta)$, the shown inequality satisfies definition 6.2. We conclude that $f(B \mid \theta) \geq^n_{\text{most: }\Theta} f(A \mid \theta)$. □

Definition 6.3 (Simply-retargetable function)**.** Let Θ be a set acted on by S_d, and let $f : \{A, B\} \times \Theta \to \mathbb{R}$. If there exists a permutation $\phi \in S_d$ such that, if $f(B \mid \theta^A) < f(A \mid \theta^A)$ implies that $f(A \mid \phi \cdot \theta^A) < f(B \mid \phi \cdot \theta^A)$, then f is a $(\Theta, A \overset{\textit{simple}}{\to} B)$*-retargetable function.*

Proposition 6.4 (Simply-retargetable functions have orbit-level tendencies)**.**

$$\textit{If } f \textit{ is } (\Theta, A \overset{\textit{simple}}{\to} B)\textit{-retargetable, then } f(B \mid \theta) \geq^1_{\text{most: }\Theta} f(A \mid \theta).$$

Proof. Given that f is a $(\Theta, A \overset{\text{simple}}{\to} B)$-retargetable function (definition 6.3), we want to show that f is a $(\Theta, A \overset{1}{\to} B)$-retargetable function (definition 6.5 when $n = 1$). Definition 6.5's item 1 is true by assumption. Since Θ is acted on by S_d, Θ is closed under permutation and so definition 6.5's item 2 holds. When $n = 1$, there are no $i \neq j$, and so definition 6.5's item 3 is tautologically true.

Then f is a $(\Theta, A \overset{1}{\to} B)$-retargetable function; apply lemma E.20. □

E.2.2 Helper results on retargetable functions

Lemma E.20 (Quantitative general orbit lemma)**.** *Let Θ be a subset of a set acted on by S_d, and let $f : \mathbf{E} \times \Theta \to \mathbb{R}$. Consider $A, B \in \mathbf{E}$.*

For each $\theta \in \Theta$, choose involutions $\phi_1, \ldots, \phi_n \in S_d$. Let $\theta^* \in \operatorname{Orbit}|_\Theta(\theta)$.

1. ***Retargetable under parameter permutation.*** *There exist $B_i^\star \in \mathbf{E}$ such that if $f(B \mid \theta^*) < f(A \mid \theta^*)$, then $\forall i : f\left(A \mid \theta^*\right) \leq f\left(B_i^\star \mid \phi_i \cdot \theta^*\right)$.*
2. *Θ **is closed under certain symmetries.** $f(B \mid \theta^*) < f(A \mid \theta^*) \implies \forall i : \phi_i \cdot \theta^* \in \Theta$.*
3. *f **is increasing on certain inputs.** $\forall i : f(B_i^\star \mid \theta^*) \leq f(B \mid \theta^*)$.*
4. ***Increasing under alternate symmetries.*** *For $j = 1, \ldots, n$ and $i \neq j$, if $f(A \mid \theta^*) < f(B \mid \theta^*)$, then $f\left(B_j^\star \mid \theta^*\right) \leq f\left(B_j^\star \mid \phi_i \cdot \theta^*\right)$.*

If these conditions hold for all $\theta \in \Theta$, then

$$f(B \mid \theta) \geq^n_{\text{most: }\Theta} f(A \mid \theta). \tag{E.20}$$

Proof. Let θ and θ^* be as described in the assumptions, and let $i \in \{1, \ldots, n\}$.

$$f(A \mid \phi_i \cdot \theta^*) = f(A \mid \phi_i^{-1} \cdot \theta^*) \tag{E.21}$$

$$\leq f(B_i^\star \mid \theta^*) \tag{E.22}$$
$$\leq f(B \mid \theta^*) \tag{E.23}$$
$$< f(A \mid \theta^*) \tag{E.24}$$
$$\leq f(B_i^\star \mid \phi_i \cdot \theta^*) \tag{E.25}$$
$$\leq f(B \mid \phi_i \cdot \theta^*). \tag{E.26}$$

Equation (E.21) follows because ϕ_i is an involution. Equation (E.22) and eq. (E.25) follow by item 1. Equation (E.23) and eq. (E.26) follow by item 3. Equation (E.24) holds by assumption on θ^*. Then eq. (E.26) shows that for any i, $f(A \mid \phi_i \cdot \theta^*) < f(B \mid \phi_i \cdot \theta^*)$, satisfying definition 6.5's item 1.

This result's item 2 satisfies definition 6.5's item 2. We now just need to show definition 6.5's item 3.

Disjointness. Let $\theta', \theta'' \in \text{Orbit}|_{\Theta, A>B}(\theta)$ and let $i \neq j$. Suppose $\phi_i \cdot \theta' = \phi_j \cdot \theta''$. We want to show that this leads to contradiction.

$$f(A \mid \theta'') \leq f(B_j^\star \mid \phi_j \cdot \theta'') \tag{E.27}$$
$$= f(B_j^\star \mid \phi_i^{-1} \cdot \theta') \tag{E.28}$$
$$\leq f(B_j^\star \mid \theta') \tag{E.29}$$
$$\leq f(B \mid \theta') \tag{E.30}$$
$$< f(A \mid \theta') \tag{E.31}$$
$$\leq f(B_i^\star \mid \phi_i \cdot \theta') \tag{E.32}$$
$$= f(B_i^\star \mid \phi_j^{-1} \cdot \theta'') \tag{E.33}$$
$$\leq f(B_i^\star \mid \theta'') \tag{E.34}$$
$$\leq f(B \mid \theta'') \tag{E.35}$$
$$< f(A \mid \theta''). \tag{E.36}$$

Equation (E.27) follows by our assumption of item 1. Equation (E.28) holds because we assumed that $\phi_j \cdot \theta'' = \phi_i \cdot \theta'$, and the involution ensures that $\phi_i = \phi_i^{-1}$. Equation (E.29) is guaranteed by our assumption of item 4, given that $\phi_i^{-1} \cdot \theta' = \phi_i \cdot \theta' \in \text{Orbit}|_{\Theta, B>A}(\theta)]$

by the first half of this proof. Equation (E.30) follows by our assumption of item 3. Equation (E.31) follows because we assumed that $\theta' \in \text{Orbit}|_{\Theta, A>B}(\theta)$.

Equation (E.32) through eq. (E.36) follow by the same reasoning, switching the roles of θ' and θ'', and of i and j. But then we have demonstrated that a quantity is strictly less than itself, a contradiction. So for all $\theta', \theta'' \in \text{Orbit}|_{\Theta, A>B}(\theta)$, when $i \neq j$, $\phi_i \cdot \theta' \neq \phi_j \cdot \theta''$.

Therefore, we have shown definition 6.5's item 3, and so f is a $(\Theta, A \overset{n}{\to} B)$-retargetable function. Apply theorem 6.6 in order to conclude that eq. (E.20) holds. □

Definition E.21 (Superset-of-copy containment)**.** Let $A, B \subseteq \mathbb{R}^d$. *$B$ contains n superset-copies $B_i^\star$ of A* when there exist involutions $\phi_1, \ldots, \phi_n$ such that $\phi_i \cdot A \subseteq B_i^\star \subseteq B$, and whenever $i \neq j$, $\phi_i \cdot B_j^\star = B_j^\star$.

Lemma E.22 (Looser sufficient conditions for orbit-level incentives)**.** *Suppose that Θ is a subset of a set acted on by S_d and is closed under permutation by S_d. Let $A, B \in \mathbf{E} \subseteq \mathcal{P}\left(\mathbb{R}^d\right)$. Suppose that B contains n superset-copies $B_i^\star \in \mathbf{E}$ of A via ϕ_i. Suppose that $f(X \mid \theta)$ is increasing under joint permutation by $\phi_1, \ldots, \phi_n \in S_d$ for all $X \in \mathbf{E}, \theta \in \Theta$, and suppose that $\forall i : \phi_i \cdot A \in \mathbf{E}$. Suppose that f is monotonically increasing on its first argument. Then $f(B \mid \theta) \geq^n_{\text{most: }\Theta} f(A \mid \theta)$.*

Proof. We check the conditions of lemma E.20. Let $\theta \in \Theta$, and let $\theta^* \in (S_d \cdot \theta) \cap \Theta$ be an orbit element.

Item 1. Holds since $f(A \mid \theta^*) \leq f(\phi_i \cdot A \mid \phi_i \cdot \theta^*) \leq f(B_i^\star \mid \phi_i \cdot \theta^*)$, with the first inequality by assumption of joint increasing under permutation, and the second following from monotonicity (as $\phi_i \cdot A \subseteq B_i^\star$ by superset copy definition E.21).

Item 2. We have $\forall \theta^* \in (S_d \cdot \theta^*) \cap \Theta : f(B \mid \theta^*) < f(A \mid \theta^*) \implies \forall i = 1, ..., n : \phi_i \cdot \theta^* \in \Theta$ since Θ is closed under permutation.

Item 3. Holds because we assumed that f is monotonic on its first argument.

Item 4. Holds because f is increasing under joint permutation on *all* of its inputs $X, \theta^{'}$, and definition E.21 shows that $\phi_i \cdot B_j^\star = B_j^\star$ when $i \neq j$. Combining these two steps of reasoning, for *all* $\theta' \in \Theta$, it is true that $f\left(B_j^\star \mid \theta'\right) \leq f\left(\phi_i \cdot B_j^\star \mid \phi_i \cdot \theta'\right) \leq f\left(B_j^\star \mid \phi_i \cdot \theta'\right)$.

Then apply lemma E.20. ☐

Lemma E.23 (Hiding an argument which is invariant under certain permutations)**.** *Let $\mathbf{E}_1$, $\mathbf{E}_2$, Θ be subsets of sets which are acted on by S_d. Let $A \in \mathbf{E}_1$, $C \in \mathbf{E}_2$. Suppose there exist $\phi_1, \ldots, \phi_n \in S_d$ such that $\phi_i \cdot C = C$. Suppose $h : \mathbf{E}_1 \times \mathbf{E}_2 \times \Theta \to \mathbb{R}$ satisfies $\forall i : h(A, C \mid \theta) \leq h(\phi_i \cdot A, \phi_i \cdot C \mid \phi_i \cdot \theta)$. For any $X \in \mathbf{E}_1$, let $f(X \mid \theta) \coloneqq h(X, C \mid \theta)$. Then $f(A \mid \theta)$ is increasing under joint permutation by ϕ_i.*

Furthermore, if h is invariant *under joint permutation by ϕ_i, then so is f.*

Proof.

$$
\begin{aligned}
f(X \mid \theta) &\coloneqq h(X, C \mid \theta) && \text{(E.37)}\\
&\leq h(\phi_i \cdot X, \phi_i \cdot C \mid \phi_i \cdot \theta) && \text{(E.38)}\\
&= h(\phi_i \cdot X, C \mid \phi_i \cdot \theta) && \text{(E.39)}\\
&\eqqcolon f(\phi_i \cdot X \mid \phi_i \cdot \theta). && \text{(E.40)}
\end{aligned}
$$

Equation (E.38) holds by assumption. Equation (E.39) follows because we assumed $\phi_i \cdot C = C$. Then f is increasing under joint permutation by the ϕ_i.

If h is *invariant*, then eq. (E.38) is an equality, and so $\forall i : f(X \mid \theta) = f(\phi_i \cdot X \mid \phi_i \cdot \theta)$. ☐

E.2.2.1 EU-determined functions

Lemma E.24 and lemma E.18 together extend Turner et al. [99]'s lemma D.22 beyond functions of $\max_{\mathbf{x} \in X_i}$, to any functions of cardinalities and of expected utilities of set elements.

Definition E.12 (EU-determined functions)**.** Let $\mathcal{P}\left(\mathbb{R}^d\right)$ be the power set of $\mathbb{R}^d$, and let $f : \prod_{i=1}^m \mathcal{P}\left(\mathbb{R}^d\right) \times \mathbb{R}^d \to \mathbb{R}$. f is an *EU-determined function* if there exists a family of functions $\{g^{\omega_1,\ldots,\omega_m}\}$ such that

$$f(X_1,\ldots,X_m \mid \mathbf{u}) = g^{|X_1|,\ldots,|X_m|}\left(\left[\mathbf{x}_1^\top \mathbf{u}\right]_{\mathbf{x}_1 \in X_1},\ldots,\left[\mathbf{x}_m^\top \mathbf{u}\right]_{\mathbf{x}_m \in X_m}\right), \tag{E.1}$$

where $[r_i]$ is the multiset of its elements r_i.

Lemma E.24 (EU-determined functions are invariant under joint permutation)**.** *Suppose that* $f : \prod_{i=1}^m \mathcal{P}\left(\mathbb{R}^d\right) \times \mathbb{R}^d \to \mathbb{R}$ *is an EU-determined function. Then for any* $\phi \in S_d$ *and* $X_1,\ldots,X_m,\mathbf{u}$*, we have* $f(X_1,\ldots,X_m \mid \mathbf{u}) = f(\phi \cdot X_1,\ldots,\phi \cdot X_m \mid \phi \cdot \mathbf{u})$*.*

Proof.

$$f(X_1,\ldots,X_m \mid \mathbf{u}) \tag{E.41}$$
$$= g^{|X_1|,\ldots,|X_m|}\left(\left[\mathbf{x}_1^\top \mathbf{u}\right]_{\mathbf{x}_1 \in X_1},\ldots,\left[\mathbf{x}_m^\top \mathbf{u}\right]_{\mathbf{x}_m \in X_m}\right) \tag{E.42}$$
$$= g^{|\phi\cdot X_1|,\ldots,|\phi\cdot X_m|}\left(\left[\mathbf{x}_1^\top \mathbf{u}\right]_{\mathbf{x}_1 \in X_1},\ldots,\left[\mathbf{x}_m^\top \mathbf{u}\right]_{\mathbf{x}_m \in X_m}\right) \tag{E.43}$$
$$= g^{|\phi\cdot X_1|,\ldots,|\phi\cdot X_m|}\left(\left[(\mathbf{P}_\phi\mathbf{x}_1)^\top (\mathbf{P}_\phi\mathbf{u})\right]_{\mathbf{x}_1 \in X_1},\ldots,\left[(\mathbf{P}_\phi\mathbf{x}_m)^\top (\mathbf{P}_\phi\mathbf{u})\right]_{\mathbf{x}_m \in X_m}\right) \tag{E.44}$$
$$= f(\phi \cdot X_1,\ldots,\phi \cdot X_m \mid \phi \cdot \mathbf{u}). \tag{E.45}$$

Equation (E.43) holds because permutations ϕ act injectively on $\mathbb{R}^d$. Equation (E.44) follows because $\mathbf{I} = \mathbf{P}_\phi^{-1}\mathbf{P}_\phi = \mathbf{P}_\phi^\top \mathbf{P}_\phi$ by the orthogonality of permutation matrices, and $\mathbf{x}^\top \mathbf{P}_\phi^\top = (\mathbf{P}_\phi \mathbf{x})^\top$, so $\mathbf{x}^\top \mathbf{u} = \mathbf{x}^\top \mathbf{P}_\phi^\top \mathbf{P}_\phi \mathbf{u} = (\mathbf{P}_\phi\mathbf{x})^\top (\mathbf{P}_\phi \mathbf{u})$. □

Theorem E.13 (Orbit tendencies occur for EU-determined decision-making functions)**.** *Let* $A, B, C \subseteq \mathbb{R}^d$ *be such that* B *contains* n *copies of* A *via* ϕ_i *such that* $\phi_i \cdot C = C$*. Let* $h : \prod_{i=1}^2 \mathcal{P}\left(\mathbb{R}^d\right) \times \mathbb{R}^d \to \mathbb{R}$ *be an EU-determined function, and let* $p(X \mid \mathbf{u}) := h(X, C \mid \mathbf{u})$*. Suppose that* p *returns a probability of selecting an element of* X *from* C*. Then* $p(B \mid \mathbf{u}) \geq^n_{\text{most: } \mathbb{R}^d} p(A \mid \mathbf{u})$*.*

Proof. By assumption, there exists a family of functions $\left\{g^{i,|C|}\right\}$ such that for all $X \subseteq \mathbb{R}^d$, $h(X, C \mid \mathbf{u}) = g^{|X|,|C|}\left(\left[\mathbf{x}^\top \mathbf{u}\right]_{\mathbf{x} \in X}, \left[\mathbf{c}^\top \mathbf{u}\right]_{\mathbf{c} \in C}\right)$. Therefore, lemma E.24 shows that

$h(A, C \mid \mathbf{u})$ is invariant under joint permutation by the ϕ_i. Letting $\Theta := \mathbb{R}^d$, apply lemma E.23 to conclude that $f(X \mid \mathbf{u})$ is invariant under joint permutation by the ϕ_i.

Since f returns a probability of selecting an element of X, f obeys the monotonicity probability axiom: If $X' \subseteq X$, then $f(X' \mid \mathbf{u}) \leq f(X \mid \mathbf{u})$. Then $f(B \mid \mathbf{u}) \geq^n_{\text{most: } \mathbb{R}^d} f(A \mid \mathbf{u})$ by lemma E.22. $\square$

E.2.3 Particular results on retargetable functions

Definition E.25 (Quantilization, closed form)**.** Let the expected utility q-quantile threshold be

$$M_{q,P}(C \mid \mathbf{u}) := \inf \left\{ M \in \mathbb{R} \mid \mathbb{P}_{\mathbf{x} \sim P} \left(\mathbf{x}^\top \mathbf{u} > M \right) \leq q \right\}. \tag{E.46}$$

Let $C_{>M_{q,P}(C|\mathbf{u})} := \left\{ \mathbf{c} \in C \mid \mathbf{c}^\top \mathbf{u} > M_{q,P}(C \mid \mathbf{u}) \right\}$. $C_{=M_{q,P}(C|\mathbf{u})}$ is defined similarly. Let $\mathbb{1}_{L(x)}$ be the predicate function returning 1 if $L(x)$ is true and 0 otherwise. Then for $X \subseteq C$,

$$Q_{q,P}(X \mid C, \mathbf{u}) := \sum_{\mathbf{x} \in X} \frac{P(\mathbf{x})}{q} \left(\mathbb{1}_{\mathbf{x} \in C_{>M_{q,P}(C|\mathbf{u})}} + \frac{\mathbb{1}_{\mathbf{x} \in C_{=M_{q,P}(C|\mathbf{u})}}}{P\left(C_{=M_{q,P}(C|\mathbf{u})}\right)} \left(q - P\left(C_{>M_{q,P}(C|\mathbf{u})} \right) \right) \right), \tag{E.47}$$

where the summand is defined to be 0 if $P(\mathbf{x}) = 0$ and $\mathbf{x} \in C_{=M_{q,P}(C|\mathbf{u})}$.

Remark. Unlike Taylor [93]'s or Carey [20]'s definitions, definition E.25 is written in closed form and requires no arbitrary tie-breaking. Instead, in the case of an expected utility tie on the quantile threshold, eq. (E.47) allots probability to outcomes proportional to their probability under the base distribution P.

Thanks to theorem E.13, we straightforwardly prove most items of proposition E.11 by just rewriting each decision-making function as an EU-determined function. Most of the proof's length comes from showing that the functions are measurable on $\mathbf{u}$, which means that the results also apply for distributions over utility functions $\mathcal{D}_{\text{any}} \in \mathfrak{D}_{\text{any}}$.

Proposition E.11 (Orbit incentives for different rationalities)**.** *Let $A, B \subseteq C \subsetneq \mathbb{R}^d$ be finite, such that B contains n copies of A via involutions ϕ_i such that $\phi_i \cdot C = C$.*

1. ***Rational choice [99].***

$$\text{Optimal}\left(B \mid C, \mathcal{D}_{any}\right) \geq^n_{\text{most: } \mathfrak{D}_{any}} \text{Optimal}\left(A \mid C, \mathcal{D}_{any}\right).$$

2. ***Uniformly randomly choosing an optimal lottery.*** *For* $X \subseteq C$*, let*

$$\text{FracOptimal}\left(X \mid C, \mathbf{u}\right) := \frac{\left|\left\{\arg\max_{\mathbf{c} \in C} \mathbf{c}^\top \mathbf{u}\right\} \cap X\right|}{\left|\left\{\arg\max_{\mathbf{c} \in C} \mathbf{c}^\top \mathbf{u}\right\}\right|}.$$

   Then $\text{FracOptimal}\left(B \mid C, \mathcal{D}_{any}\right) \geq^n_{\text{most: } \mathfrak{D}_{any}} \text{FracOptimal}\left(A \mid C, \mathcal{D}_{any}\right)$*.*

3. ***Anti-rational choice.*** $\text{AntiOpt}\left(B \mid C, \mathcal{D}_{any}\right) \geq^n_{\text{most: } \mathfrak{D}_{any}} \text{AntiOpt}\left(A \mid C, \mathcal{D}_{any}\right)$*.*

4. ***Boltzmann rationality.***

$$\text{Boltzmann}_T\left(B \mid C, \mathcal{D}_{any}\right) \geq^n_{\text{most: } \mathfrak{D}_{any}} \text{Boltzmann}_T\left(A \mid C, \mathcal{D}_{any}\right).$$

5. ***Uniformly randomly drawing*** k ***outcome lotteries and choosing the best.*** *For* $X \subseteq C$*,* $\mathbf{u} \in \mathbb{R}^d$*, and* $k \geq 1$*, let*

$$\textit{best-of-k}(X, C \mid \mathbf{u}) := \mathop{\mathbb{E}}_{\mathbf{a}_1, \ldots, \mathbf{a}_k \sim \textit{unif}(C)}\left[\text{FracOptimal}\left(X \cap \{\mathbf{a}_1, \ldots, \mathbf{a}_k\} \mid \{\mathbf{a}_1, \ldots, \mathbf{a}_k\}, \mathbf{u}\right)\right].$$

   Then $\textit{best-of-k}(B \mid C, \mathcal{D}_{any}) \geq^n_{\text{most: } \mathfrak{D}_{any}} \textit{best-of-k}(A \mid C, \mathcal{D}_{any})$*.*

6. ***Satisficing [86].*** $\text{Satisfice}_t\left(B \mid C, \mathcal{D}_{any}\right) \geq^n_{\text{most: } \mathfrak{D}_{any}} \text{Satisfice}_t\left(A \mid C, \mathcal{D}_{any}\right)$*.*

7. ***Quantilizing over outcome lotteries [93].*** *Let* P *be the uniform probability distribution over* C*. For* $X \subseteq C$*,* $\mathbf{u} \in \mathbb{R}^d$*, and* $q \in (0, 1]$*, let* $Q_{q,P}(X \mid C, \mathbf{u})$ *(definition E.25) return the probability that an outcome lottery in* X *is drawn from the top q-quantile of* P*, sorted by expected utility under* $\mathbf{u}$*. Then* $Q_{q,P}(B \mid C, \mathbf{u}) \geq^n_{\text{most: } \mathbb{R}^d} Q_{q,P}(A \mid C, \mathbf{u})$*.*

Proof. **Item 1.** Consider

$$h(X, C \mid \mathbf{u}) := \mathbb{1}_{\exists \mathbf{x} \in X : \forall \mathbf{c} \in C : \mathbf{x}^\top \mathbf{u} \geq \mathbf{c}^\top \mathbf{u}} \tag{E.48}$$

$$= \min\left(1, \sum_{\mathbf{x} \in X} \prod_{\mathbf{c} \in C} \mathbb{1}_{(\mathbf{x}-\mathbf{c})^\top \mathbf{u} \geq 0}\right). \tag{E.49}$$

Since halfspaces are measurable, each indicator function is measurable on $\mathbf{u}$. The finite sum of the finite product of measurable functions is also measurable. Since min is continuous (and therefore measurable), $h(X, C \mid \mathbf{u})$ is measurable on $\mathbf{u}$.

Furthermore, h is an EU-determined function:

$$h(X, C \mid \mathbf{u}) = g\left(\overbrace{\left[\mathbf{x}^\top \mathbf{u}\right]_{\mathbf{x} \in X}}^{V_X}, \overbrace{\left[\mathbf{c}^\top \mathbf{u}\right]_{\mathbf{c} \in C}}^{V_C}\right) \tag{E.50}$$

$$:= \mathbb{1}_{\exists v_x \in V_X : \forall v_c \in V_C : v_x \geq v_c}. \tag{E.51}$$

Then by lemma E.24, h is invariant to joint permutation by the ϕ_i. Since $\phi_i \cdot C = C$, lemma E.23 shows that $h'(X \mid \mathbf{u}) := h(X, C \mid \mathbf{u})$ is also invariant under joint permutation by the ϕ_i. Since h is a measurable function of $\mathbf{u}$, so is h'. Then since h' is bounded, lemma E.18 shows that $f(X \mid \mathcal{D}_{\text{any}}) := \mathbb{E}_{\mathbf{u} \sim \mathcal{D}_{\text{any}}}\left[h'(X \mid \mathbf{u})\right]$ is invariant under joint permutation by ϕ_i.

Furthermore, if $X' \subseteq X$, $f(X' \mid \mathcal{D}_{\text{any}}) \leq f(X \mid \mathcal{D}_{\text{any}})$ by the monotonicity of probability. Then by lemma E.22,

$$f(B \mid \mathcal{D}_{\text{any}}) := \text{Optimal}\left(B \mid C, \mathcal{D}_{\text{any}}\right) \geq^n_{\text{most: } \mathfrak{D}_{\text{any}}} \text{Optimal}\left(A \mid C, \mathcal{D}_{\text{any}}\right) =: f(A \mid \mathcal{D}_{\text{any}}).$$

Item 2. Because X, C are finite sets, the denominator of $\text{FracOptimal}\left(X \mid C, \mathbf{u}\right)$ is never zero, and so the function is well-defined. $\text{FracOptimal}\left(X \mid C, \mathbf{u}\right)$ is an EU-determined function:

$$\mathrm{FracOptimal}\left(X \mid C, \mathbf{u}\right) = g\left(\overbrace{\left[\mathbf{x}^\top \mathbf{u}\right]_{\mathbf{x}\in X}}^{V_X}, \overbrace{\left[\mathbf{c}^\top \mathbf{u}\right]_{\mathbf{c}\in C}}^{V_C}\right) \tag{E.52}$$

$$:= \frac{\left|\left[v \in V_X \mid v = \max_{v'\in V_C} v'\right]\right|}{\left|\left[\arg\max_{v'\in V_C} v'\right]\right|}, \tag{E.53}$$

with the $[\cdot]$ denoting a multiset which allows and counts duplicates. Then by lemma E.24, $\mathrm{FracOptimal}\left(X \mid C, \mathbf{u}\right)$ is invariant to joint permutation by the ϕ_i.

We now show that $\mathrm{FracOptimal}\left(X \mid C, \mathbf{u}\right)$ is a measurable function of $\mathbf{u}$.

$$\mathrm{FracOptimal}\left(X \mid C, \mathbf{u}\right) := \frac{\left|\left\{\arg\max_{\mathbf{c}'\in C} \mathbf{c}'^\top \mathbf{u}\right\} \cap X\right|}{\left|\left\{\arg\max_{\mathbf{c}'\in C} \mathbf{c}'^\top \mathbf{u}\right\}\right|} \tag{E.54}$$

$$= \frac{\sum_{\mathbf{x}\in X} \mathbb{1}_{\mathbf{x}\in\arg\max_{\mathbf{c}'\in C} \mathbf{c}'^\top \mathbf{u}}}{\sum_{\mathbf{c}\in C} \mathbb{1}_{\mathbf{c}\in\arg\max_{\mathbf{c}'\in C} \mathbf{c}'^\top \mathbf{u}}} \tag{E.55}$$

$$= \frac{\sum_{\mathbf{x}\in X} \prod_{\mathbf{c}'\in C} \mathbb{1}_{(\mathbf{x}-\mathbf{c}')^\top \mathbf{u}\geq 0}}{\sum_{\mathbf{c}\in C} \prod_{\mathbf{c}'\in C} \mathbb{1}_{(\mathbf{c}-\mathbf{c}')^\top \mathbf{u}\geq 0}}. \tag{E.56}$$

Equation (E.56) holds because $\mathbf{x}$ belongs to the $\arg\max$ iff $\forall \mathbf{c} \in C : \mathbf{x}^\top \mathbf{u} \geq \mathbf{c}^\top \mathbf{u}$. Furthermore, this condition is met iff $\mathbf{u}$ belongs to the intersection of finitely many closed halfspaces; therefore, $\left\{\mathbf{u} \in \mathbb{R}^d \mid \prod_{\mathbf{c}\in C} \mathbb{1}_{(\mathbf{x}-\mathbf{c})^\top \mathbf{u}\geq 0} = 1\right\}$ is measurable. Then the sums in both the numerator and denominator are both measurable functions of $\mathbf{u}$, and the denominator cannot vanish. Therefore, $\mathrm{FracOptimal}\left(X \mid C, \mathbf{u}\right)$ is a measurable function of $\mathbf{u}$.

Let $g(X \mid \mathbf{u}) := \mathrm{FracOptimal}\left(X \mid C, \mathbf{u}\right)$. Since $\phi_i \cdot C = C$, lemma E.23 shows that $g(X \mid \mathbf{u})$ is also invariant to joint permutation by ϕ_i. Since g is measurable and bounded $[0, 1]$, apply lemma E.18 to conclude that $f(X \mid \mathcal{D}_{\text{any}}) := \mathbb{E}_{\mathbf{u}\sim\mathcal{D}_{\text{any}}}\left[g(X \mid C, \mathbf{u})\right]$ is also invariant to joint permutation by ϕ_i.

Furthermore, if $X' \subseteq X \subseteq C$, then $f(X' \mid \mathcal{D}_{\text{any}}) \leq f(X \mid \mathcal{D}_{\text{any}})$. So apply lemma E.22

to conclude that FracOptimal $\left(B \mid C, \mathcal{D}_{\text{any}}\right) =: f(B \mid \mathcal{D}_{\text{any}}) \geq^n_{\text{most: } \mathfrak{D}_{\text{any}}} f(A \mid \mathcal{D}_{\text{any}}) :=$ FracOptimal $\left(A \mid C, \mathcal{D}_{\text{any}}\right)$.

Item 3. Apply the reasoning in item 1 with inner function $h(X \mid C, \mathbf{u}) := \mathbb{1}_{\exists \mathbf{x} \in X : \forall \mathbf{c} \in C : \mathbf{x}^\top \mathbf{u} \leq \mathbf{c}^\top \mathbf{u}}$.

Item 4. Let $X \subseteq C$. Boltzmann$_T\left(X \mid C, \mathbf{u}\right)$ is the expectation of an EU function:

$$\text{Boltzmann}_T\left(X \mid C, \mathbf{u}\right) = g_T\left(\overbrace{\left[\mathbf{x}^\top \mathbf{u}\right]_{\mathbf{x} \in X}}^{V_X}, \overbrace{\left[\mathbf{c}^\top \mathbf{u}\right]_{\mathbf{c} \in C}}^{V_C}\right) \tag{E.57}$$

$$:= \frac{\sum_{v \in V_X} e^{v/T}}{\sum_{v \in V_C} e^{v/T}}. \tag{E.58}$$

Therefore, by lemma E.24, Boltzmann$_T\left(X \mid C, \mathbf{u}\right)$ is invariant to joint permutation by the ϕ_i.

Inspecting eq. (E.58), we see that g is continuous on $\mathbf{u}$ (and therefore measurable), and bounded $[0,1]$ since $X \subseteq C$ and the exponential function is positive. Therefore, by lemma E.18, the expectation version is also invariant to joint permutation for all permutations $\phi \in S_d$: Boltzmann$_T\left(X \mid C, \mathcal{D}_{\text{any}}\right) =$ Boltzmann$_T\left(\phi \cdot X \mid \phi \cdot C, \phi \cdot \mathcal{D}_{\text{any}}\right)$.

Since $\phi_i \cdot C = C$, lemma E.23 shows that $f(X \mid \mathcal{D}_{\text{any}}) :=$ Boltzmann$_T\left(X \mid C, \mathcal{D}_{\text{any}}\right)$ is also invariant under joint permutation by the ϕ_i. Furthermore, if $X' \subseteq X$, then $f(X' \mid \mathcal{D}_{\text{any}}) \leq f(X \mid \mathcal{D}_{\text{any}})$. Then apply lemma E.22 to conclude that Boltzmann$_T\left(B \mid C, \mathcal{D}_{\text{any}}\right) =: f(B \mid \mathcal{D}_{\text{any}}) \geq^n_{\text{most: } \mathfrak{D}_{\text{any}}} f(A \mid \mathcal{D}_{\text{any}}) :=$ Boltzmann$_T\left(A \mid C, \mathcal{D}_{\text{any}}\right)$.

Item 5. Let involution $\phi \in S_d$ fix C (*i.e.* $\phi \cdot C = C$).

$$\text{best-of-}k(X \mid C, \mathbf{u}) \tag{E.59}$$

$$:= \mathop{\mathbb{E}}_{\mathbf{a}_1, \ldots, \mathbf{a}_k \sim \text{unif}(C)}\left[\text{FracOptimal}\left(X \cap \{\mathbf{a}_1, \ldots, \mathbf{a}_k\} \mid \{\mathbf{a}_1, \ldots, \mathbf{a}_k\}, \mathbf{u}\right)\right] \tag{E.60}$$

$$= \mathop{\mathbb{E}}_{\mathbf{a}_1, \ldots, \mathbf{a}_k \sim \text{unif}(C)}\left[\text{FracOptimal}\left((\phi \cdot X) \cap \{\phi \cdot \mathbf{a}_1, \ldots, \phi \cdot \mathbf{a}_k\} \,|\, \{\phi \cdot \mathbf{a}_1, \ldots, \phi \cdot \mathbf{a}_k\}, \phi \cdot \mathbf{u}\right)\right] \tag{E.61}$$

$$= \mathop{\mathbb{E}}_{\phi\cdot\mathbf{a}_1,\ldots,\phi\cdot\mathbf{a}_k\sim\text{unif}(\phi\cdot C)}\left[\text{FracOptimal}\left((\phi\cdot X)\cap\{\phi\cdot\mathbf{a}_1,\ldots,\phi\cdot\mathbf{a}_k\}\,|\,\{\phi\cdot\mathbf{a}_1,\ldots,\phi\cdot\mathbf{a}_k\},\phi\cdot\mathbf{u}\right)\right] \tag{E.62}$$

$$=: \text{best-of-}k(\phi\cdot X\mid\phi\cdot C,\phi\cdot\mathbf{u}). \tag{E.63}$$

By the proof of item 2,

$$\text{FracOptimal}\left(X\cap\{\mathbf{a}_1,\ldots,\mathbf{a}_k\}\mid\{\mathbf{a}_1,\ldots,\mathbf{a}_k\},\mathbf{u}\right) = \text{FracOptimal}\left((\phi\cdot X)\cap\{\phi\cdot\mathbf{a}_1,\ldots,\phi\cdot\mathbf{a}_k\}\mid\{\phi\cdot\mathbf{a}_1,\ldots,\phi\cdot\mathbf{a}_k\},\phi\cdot\mathbf{u}\right);$$

thus, eq. (E.61) holds. Since $\phi\cdot C=C$ and since the distribution is uniform, eq. (E.62) holds. Therefore, best-of-$k(X\mid C,\mathbf{u})$ is invariant to joint permutation by the ϕ_i, which are involutions fixing C.

We now show that best-of-$k(X\mid C,\mathbf{u})$ is measurable on $\mathbf{u}$.

$$\text{best-of-}k(X\mid C,\mathbf{u}) \tag{E.64}$$

$$:= \mathop{\mathbb{E}}_{\mathbf{a}_1,\ldots,\mathbf{a}_k\sim\text{unif}(C)}\left[\text{FracOptimal}\left(X\cap\{\mathbf{a}_1,\ldots,\mathbf{a}_k\}\mid\{\mathbf{a}_1,\ldots,\mathbf{a}_k\},\mathbf{u}\right)\right] \tag{E.65}$$

$$= \frac{1}{|C|^k}\sum_{(\mathbf{a}_1,\ldots,\mathbf{a}_k)\in C^k}\text{FracOptimal}\left(X\cap\{\mathbf{a}_1,\ldots,\mathbf{a}_k\}\mid\{\mathbf{a}_1,\ldots,\mathbf{a}_k\},\mathbf{u}\right). \tag{E.66}$$

Equation (E.66) holds because FracOptimal $\left(X\mid C,\mathbf{u}\right)$ is measurable on $\mathbf{u}$ by item 2, and measurable functions are closed under finite addition and scalar multiplication. Then best-of-$k(X\mid C,\mathbf{u})$ is measurable on $\mathbf{u}$.

Let $g(X\mid\mathbf{u}):=$ best-of-$k(X\mid C,\mathbf{u})$. Since $\phi_i\cdot C=C$, lemma E.23 shows that $g(X\mid\mathbf{u})$ is also invariant to joint permutation by ϕ_i. Since g is measurable and bounded $[0,1]$, apply lemma E.18 to conclude that $f(X\mid\mathcal{D}_{\text{any}}):=\mathbb{E}_{\mathbf{u}\sim\mathcal{D}_{\text{any}}}\left[g(X\mid C,\mathbf{u})\right]$ is also invariant to joint permutation by ϕ_i.

Furthermore, if $X'\subseteq X\subseteq C$, then $f(X'\mid\mathcal{D}_{\text{any}})\leq f(X\mid\mathcal{D}_{\text{any}})$. So apply lemma E.22 to conclude that best-of-$k(B\mid C,\mathcal{D}_{\text{any}})=: f(B\mid\mathcal{D}_{\text{any}})\geq^n_{\text{most: }\mathfrak{D}_{\text{any}}} f(A\mid\mathcal{D}_{\text{any}}):=$ best-of-$k(A\mid C,\mathcal{D}_{\text{any}})$.

Item 6. $\text{Satisfice}_t\left(X \mid C, \mathbf{u}\right)$ is an EU-determined function:

$$\text{Satisfice}_t\left(X \mid C, \mathbf{u}\right) = g_t\left(\overbrace{\left[\mathbf{x}^\top \mathbf{u}\right]_{\mathbf{x}\in X}}^{V_X}, \overbrace{\left[\mathbf{c}^\top \mathbf{u}\right]_{\mathbf{c}\in C}}^{V_C}\right) \tag{E.67}$$

$$:= \frac{\sum_{v\in V_X} \mathbb{1}_{v\geq t}}{\sum_{v\in V_C} \mathbb{1}_{v\geq t}}, \tag{E.68}$$

with the function evaluating to 0 if the denominator is 0.
Then applying lemma E.24, $\text{Satisfice}_t\left(X \mid C, \mathbf{u}\right)$ is invariant under joint permutation by the ϕ_i.

We now show that $\text{Satisfice}_t\left(X \mid C, \mathbf{u}\right)$ is measurable on $\mathbf{u}$.

$$\text{Satisfice}_t\left(X \mid C, \mathbf{u}\right) = \begin{cases} \frac{\sum_{\mathbf{x}\in X} \mathbb{1}_{\mathbf{x}\in\left\{\mathbf{x}'\in\mathbb{R}^d \mid \mathbf{x}'^\top \mathbf{u}\geq t\right\}}}{\sum_{\mathbf{c}\in C} \mathbb{1}_{\mathbf{c}\in\left\{\mathbf{x}'\in\mathbb{R}^d \mid \mathbf{x}'^\top \mathbf{u}\geq t\right\}}} & \exists \mathbf{c}\in C : \mathbf{c}^\top \mathbf{u} \geq t, \\ 0 & \text{else.} \end{cases} \tag{E.69}$$

Consider the two cases.

$$\exists \mathbf{c}\in C : \mathbf{c}^\top \mathbf{u} \geq t \iff \mathbf{u} \in \bigcup_{\mathbf{c}\in C} \left\{\mathbf{u}' \in \mathbb{R}^d \mid \mathbf{c}^\top \mathbf{u} \geq t\right\}.$$

The right-hand set is the union of finitely many halfspaces (which are measurable), and so the right-hand set is also measurable. Then the casing is a measurable function of $\mathbf{u}$. Clearly the zero function is measurable. Now we turn to the first case.

In the first case, eq. (E.69)'s indicator functions test each $\mathbf{x}, \mathbf{c}$ for membership in a closed halfspace with respect to $\mathbf{u}$. Halfspaces are measurable sets. Therefore, the indicator function is a measurable function of $\mathbf{u}$, and so are the finite sums. Since the denominator does not vanish within the case, the first case as a whole is a measurable function of $\mathbf{u}$. Therefore, $\text{Satisfice}_t\left(X \mid C, \mathbf{u}\right)$ is measurable on $\mathbf{u}$.

Since $\text{Satisfice}_t\left(X \mid C, \mathbf{u}\right)$ is measurable and bounded $[0, 1]$ (as $X \subseteq C$), apply lemma E.18 to conclude that $\text{Satisfice}_t\left(X \mid C, \mathcal{D}_{\text{any}}\right) = \text{Satisfice}_t\left(\phi \cdot X \mid \phi \cdot C, \phi \cdot \mathcal{D}_{\text{any}}\right)$. Next, let $f(X \mid \mathcal{D}_{\text{any}}) := \text{Satisfice}_t\left(X \mid C, \mathcal{D}_{\text{any}}\right)$. Since we just showed that $\text{Satisfice}_t\left(X \mid C, \mathcal{D}_{\text{any}}\right)$

is invariant to joint permutation by the involutions ϕ_i and since $\phi_i \cdot C = C$, $f(X \mid \mathcal{D}_{\text{any}})$ is also invariant to joint permutation by ϕ_i.

Furthermore, if $X' \subseteq X$, we have $f(X' \mid \mathcal{D}_{\text{any}}) \leq f(X \mid \mathcal{D}_{\text{any}})$. Then applying lemma E.22, $\text{Satisfice}_t\left(B \mid C, \mathbf{u}\right) =: f(B \mid \mathcal{D}_{\text{any}}) \geq^n_{\text{most: } \mathfrak{D}_{\text{any}}} f(A \mid \mathcal{D}_{\text{any}}) \coloneqq \text{Satisfice}_t\left(A \mid C, \mathbf{u}\right)$.

Item 7. Suppose P is uniform over C and consider any of the involutions ϕ_i.

$$M_{q,P}(C \mid \mathbf{u}) \coloneqq \inf\left\{M \in \mathbb{R} \mid \underset{\mathbf{x}\sim P}{\mathbb{P}}\left(\mathbf{x}^\top \mathbf{u} > M\right) \leq q\right\} \tag{E.70}$$

$$= \inf\left\{M \in \mathbb{R} \mid \underset{\mathbf{x}\sim P}{\mathbb{P}}\left((\mathbf{P}_{\phi_i}\mathbf{x})^\top (\mathbf{P}_{\phi_i}\mathbf{u}) > M\right) \leq q\right\} \tag{E.71}$$

$$= \inf\left\{M \in \mathbb{R} \mid \underset{\mathbf{x}\sim \phi_i \cdot P}{\mathbb{P}}\left(\mathbf{x}^\top (\mathbf{P}_{\phi_i}\mathbf{u}) > M\right) \leq q\right\} \tag{E.72}$$

$$= \inf\left\{M \in \mathbb{R} \mid \underset{\mathbf{x}\sim P}{\mathbb{P}}\left(\mathbf{x}^\top (\mathbf{P}_{\phi_i}\mathbf{u}) > M\right) \leq q\right\} \tag{E.73}$$

$$=: M_{q,P}(\phi_i \cdot C \mid \phi_i \cdot \mathbf{u}). \tag{E.74}$$

Equation (E.71) follows by the orthogonality of permutation matrices. Equation (E.73) follows because if $\mathbf{x} \in \text{supp}(P) = C$, then $\phi_i \cdot \mathbf{x} \in C = \text{supp}(P)$, and furthermore $P(\mathbf{x}) = P(\mathbf{P}_{\phi_i}\mathbf{x})$ by uniformity.

Now we show the invariance of $C_{>M_{q,P}(C|\mathbf{u})}$ under joint permutation by ϕ_i:

$$C_{>M_{q,P}(C|\mathbf{u})} \coloneqq \left\{\mathbf{c} \in C \mid \mathbf{c}^\top \mathbf{u} > M_{q,P}(C \mid \mathbf{u})\right\} \tag{E.75}$$

$$= \left\{\mathbf{c} \in C \mid (\mathbf{P}_{\phi_i}\mathbf{c})^\top (\mathbf{P}_{\phi_i}\mathbf{u}) > M_{q,P}(\phi_i \cdot C \mid \phi_i \cdot \mathbf{u})\right\} \tag{E.76}$$

$$= \left\{\mathbf{c} \in \phi_i \cdot C \mid \mathbf{c}^\top (\mathbf{P}_{\phi_i}\mathbf{u}) > M_{q,P}(\phi_i \cdot C \mid \phi_i \cdot \mathbf{u})\right\} \tag{E.77}$$

$$=: C_{>M_{q,P}(\phi_i \cdot C|\phi_i \cdot \mathbf{u})}. \tag{E.78}$$

Equation (E.76) follows by the orthogonality of permutation matrices and because $M_{q,P}(C \mid \mathbf{u}) = M_{q,P}(\phi_i \cdot C \mid \phi_i \cdot \mathbf{u})$ by eq. (E.74). A similar proof shows that $C_{=M_{q,P}(C|\mathbf{u})} = C_{=M_{q,P}(\phi_i \cdot C|\phi_i \cdot \mathbf{u})}$.

Recall that

$$Q_{q,P}(X \mid C, \mathbf{u}) \coloneqq \sum_{\mathbf{x}\in X} \frac{P(\mathbf{x})}{q} \left(\mathbb{1}_{\mathbf{x}\in C_{>M_{q,P}(C\mid\mathbf{u})}} + \frac{\mathbb{1}_{\mathbf{x}\in C_{=M_{q,P}(C\mid\mathbf{u})}}}{P\left(C_{=M_{q,P}(C\mid\mathbf{u})}\right)} \left(q - P\left(C_{>M_{q,P}(C\mid\mathbf{u})}\right)\right)\right). \tag{E.79}$$

$Q_{q,P}(X \mid C, \mathbf{u}) = Q_{q,P}(\phi_i \cdot X \mid \phi_i \cdot C, \phi_i \cdot \mathbf{u})$, since Q is the sum of products of ϕ_i-invariant quantities.

$P(\mathbf{x})$ is non-negative because P is a probability distribution, and q is assumed positive. The indicator functions $\mathbb{1}$ are non-negative. By the definition of $M_{q,P}$, $P\left(C_{>M_{q,P}(C\mid\mathbf{u})}\right) \leq q$. Therefore, eq. (E.79) is the sum of non-negative terms. Thus, if $X' \subseteq X$, then $Q_{q,P}(X' \mid C, \mathbf{u}) \leq Q_{q,P}(X \mid C, \mathbf{u})$.

Let $f(X \mid \mathbf{u}) \coloneqq Q_{q,P}(X \mid C, \mathbf{u})$. Since $\phi_i \cdot C = C$ and since $Q_{q,P}(X \mid C, \mathbf{u}) = Q_{q,P}(\phi_i \cdot X \mid \phi_i \cdot C, \phi_i \cdot \mathbf{u})$, lemma E.23 shows that $f(X \mid \mathbf{u})$ is also jointly invariant to permutation by ϕ_i. Lastly, if $X' \subseteq X$, we have $f(X' \mid \mathcal{D}_{\text{any}}) \leq f(X \mid \mathcal{D}_{\text{any}})$.

Apply lemma E.22 to conclude that $Q_{q,P}(B \mid C, \mathbf{u}) =: f(B \mid \mathbf{u}) \geq^n_{\text{most: } \mathbb{R}^d} f(A \mid \mathbf{u}) \coloneqq Q_{q,P}(A \mid C, \mathbf{u})$. □

Conjecture E.26 (Orbit tendencies occur for more quantilizer base distributions)**.** Proposition E.11's item 7 holds for any base distribution P over C such that $\min_{\mathbf{b}\in B} P(\mathbf{b}) \geq \max_{\mathbf{a}\in A} P(\mathbf{a})$. Furthermore, $Q_{q,P}(X \mid C, \mathbf{u})$ is measurable on $\mathbf{u}$ and so $\geq^n_{\text{most: } \mathbb{R}^d}$ can be generalized to $\geq^n_{\text{most: } \mathfrak{D}_{\text{any}}}$.

E.3 Detailed analyses of MR scenarios

E.3.1 Action selection

Consider a bandit problem with five arms $a_1, \ldots, a_5$ partitioned $A \coloneqq \{a_1\}, B \coloneqq \{a_2, \ldots, a_5\}$, which each action has a definite utility $\mathbf{u}_i$. There are $T = 100$ trials. Suppose the training procedure train uses the ϵ-greedy strategy to learn value estimates for each arm. At the end of training, train outputs a greedy policy with respect to its value estimates. Consider any action-value initialization, and the learning rate is set $\alpha \coloneqq 1$. To learn an optimal policy, at worst, the agent just has to try each action once.

Lemma E.27 (Lower bound on success probability of the train bandit)**.** *Let* $\mathbf{u} \in \mathbb{R}^5$ *assign strictly maximal utility to* a_i*, and suppose train (described above) runs for* $T \geq 5$ *trials. Then* $f_{train}(\{a_i\} \mid \mathbf{u}) \geq 1 - (1 - \frac{\epsilon}{4})^T$.

Proof. Since the trained policy can be stochastic,

$$f_{\text{train}}(\{a_i\} \mid \mathbf{u}) \geq \mathbb{P}\left(a_i \text{ is assigned probability 1 by the learned greedy policy}\right).$$

Since a_i has strictly maximal utility which is deterministic, and since the learning rate $\alpha \coloneqq 1$, if action a_i is ever drawn, it is assigned probability 1 by the learned policy. The probability that a_i is never explored is at most $(1 - \frac{\epsilon}{4})^T$, because at worst, a_i is an "explore" action (and not an "exploit" action) at every time step, in which case it is ignored with probability $1 - \frac{\epsilon}{4}$. □

Proposition E.28 (The train bandit is 4-retargetable)**.** f_{train} *is* $(\mathbb{R}^5, A \xrightarrow{4} B)$*-retargetable.*

Proof. Let $\phi_i \coloneqq a_1 \leftrightarrow a_i$ for $i = 2, \ldots, 5$ and let $\Theta \coloneqq \mathbb{R}^5$. We want to show that whenever $\mathbf{u} \in \mathbb{R}^5$ induces $f_{\text{train}}(A \mid \mathbf{u}) > f_{\text{train}}(B \mid \mathbf{u})$, retargeting $\mathbf{u}$ will get train to instead learn to pull a B-action: $f_{\text{train}}(A \mid \phi_i \cdot \mathbf{u}) < f_{\text{train}}(B \mid \phi_i \cdot \mathbf{u})$.

Suppose we have such a $\mathbf{u}$. If $\mathbf{u}$ is constant, a symmetry argument shows that each action has equal probability of being selected, in which case $f_{\text{train}}(A \mid \mathbf{u}) = \frac{1}{5} < \frac{4}{5} = f_{\text{train}}(B \mid \mathbf{u})$—a contradiction. Therefore, $\mathbf{u}$ is not constant. Similar symmetry arguments show that A's action a_1 has strictly maximal utility ($\mathbf{u}_1 > \max_{i=2,\ldots,5} \mathbf{u}_i$).

But for $T = 100$, lemma E.27 shows that $f_{\text{train}}(A \mid \mathbf{u}) = f_{\text{train}}(\{a_1\} \mid \mathbf{u}) \approx 1$ and $f_{\text{train}}(\{a_{i \neq 1}\} \mid \mathbf{u}) \approx 0 \implies f_{\text{train}}(B \mid \mathbf{u}) = \sum_{i \neq 1} f_{\text{train}}(\{a_i\} \mid \mathbf{u}) \approx 0$. The converse statement holds when considering $\phi_i \cdot \mathbf{u}$ instead of $\mathbf{u}$. Therefore, train satisfies definition 6.5's item 1 (retargetability). These $\phi_i \cdot \mathbf{u} \in \Theta \coloneqq \mathbb{R}^5$ because $\mathbb{R}^5$ is closed under permutation by S_5, satisfying item 2.

Consider another $\mathbf{u}' \in \mathbb{R}^5$ such that $f_{\text{train}}(A \mid \mathbf{u}') > f_{\text{train}}(B \mid \mathbf{u}')$, and consider $i \neq j$. By the above symmetry arguments, $\mathbf{u}'$ must also assign a_1 maximal utility. By lemma E.27,

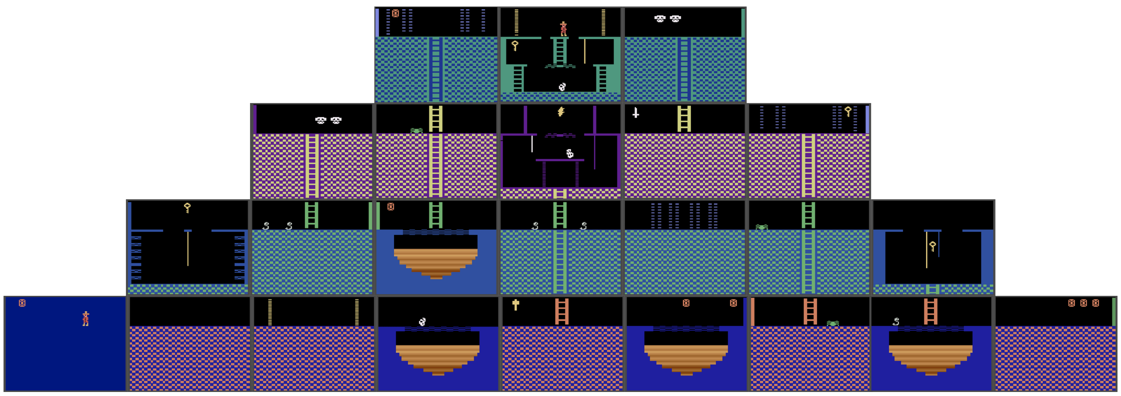


Figure E.1: Map of the first level of Montezuma's Revenge.

$f_{\text{train}}(\{a_i\} \mid \phi_i \cdot \mathbf{u}) \approx 1$ and $f_{\text{train}}(\{a_j\} \mid \phi_i \cdot \mathbf{u}) \approx 0$ since $i \neq j$, and vice versa when considering $\phi_j \cdot \mathbf{u}$ instead of $\phi_i \cdot \mathbf{u}$. Then since $\phi_i \cdot \mathbf{u}$ and $\phi_j \cdot \mathbf{u}$ induce distinct probability distributions over learned actions, they cannot be the same utility function. This satisfies item 3. $\square$

Corollary E.29 (The train bandit has orbit-level tendencies)**.** *$f_{train}(B \mid \mathbf{u}) \geq^4_{most:\ \mathbb{R}^5} f_{train}(A \mid \mathbf{u})$.*

Proof. Combine proposition E.28 and theorem 6.6. $\square$

E.3.2 Observation reward maximization

Let T be a reasonably long rollout length, so that $O_{T\text{-reach}}$ is large—many different step-T observations can be induced.

Proposition E.30 (Final reward maximization has strong orbit-level incentives in MR)**.** *Let $n := \lfloor \frac{|O_{leave}|}{|O_{stay}|} \rfloor$. $f_{max}(O_{leave} \mid R) \geq^n_{\text{most: } \mathbb{R}^{\mathcal{O}}} f_{max}(O_{stay} \mid R)$.*

Proof. Consider the vector space representation of observations, $\mathbb{R}^{|\mathcal{O}|}$. Define $A \coloneqq \{\mathbf{e}_o \mid o \in O_{\text{stay}}\}$, $B \coloneqq \{\mathbf{e}_o \mid o \in O_{\text{leave}}\}$, and $C \coloneqq O_{T\text{-reach}} = A \cup B$ the union of $O_{\text{stay}}, O_{\text{leave}}$.

Since $|O_{\text{leave}}| \geq \left|O_{\text{stay}}\right|$ by assumption that T is reasonably large, consider the involution $\phi_1 \in S_{|\mathcal{O}|}$ which embeds O_{stay} into O_{leave}, while fixing all other observations. If possible, produce another involution ϕ_2 which also embeds O_{stay} into O_{leave}, which fixes all other observations, and which "doesn't interfere with ϕ_1" (*i.e.* $\phi_2 \cdot (\phi_1 \cdot A) = \phi_1 \cdot A$). We can produce $n \coloneqq \lfloor \frac{|O_{\text{leave}}|}{|O_{\text{stay}}|} \rfloor$ such involutions. Therefore, B contains n copies (definition E.7) of A via involutions $\phi_1, \ldots, \phi_n$. Furthermore, $\phi_i \cdot (A \cup B) = A \cup B$, since each ϕ_i swaps A with $B' \subseteq B$, and fixes all $\mathbf{b} \in B \setminus B'$ by assumption. Thus, $\phi \cdot C = C$.

By proposition E.11's item 2, $\text{FracOptimal}\big(B \mid C, R\big) \geq^n_{\text{most: } \mathbb{R}^{\mathcal{O}}} \text{FracOptimal}\big(A \mid C, R\big)$. Since $f_{\max}$ uniformly randomly chooses a maximal-reward observation to induce, $\forall X \subseteq C : f_{\max}(X \mid R) = \text{FracOptimal}\big(X \mid C, R\big)$. Therefore, $f_{\max}(O_{\text{leave}} \mid R) \geq^n_{\text{most: } \mathbb{R}^{\mathcal{O}}} f_{\max}(O_{\text{stay}} \mid R)$. $\square$

We want to reason about the probability that decide leaves the initial room by time T in its rollout trajectories.

$$p_{\text{decide}}(\text{leave} \mid \theta) \coloneqq \mathbb{P}_{\substack{\pi \sim \text{decide}(\theta), \\ \tau \sim \pi \mid s_0}} \left(\tau \text{ has left the first room by step } T\right), \tag{E.80}$$

$$p_{\text{decide}}(\text{stay} \mid \theta) \coloneqq \mathbb{P}_{\substack{\pi \sim \text{decide}(\theta), \\ \tau \sim \pi \mid s_0}} \left(\tau \text{ has not left the first room by step } T\right). \tag{E.81}$$

We want to show that reward maximizers tend to leave the room: $p_{\max}(\text{leave} \mid R) \geq^n_{\text{most: } \Theta} p_{\max}(\text{stay} \mid R)$. However, we must be careful: In general, $f_{\max}(O_{\text{leave}} \mid R) \neq p_{\max}(\text{leave} \mid R)$ and $f_{\max}(O_{\text{stay}} \mid R) \neq p_{\max}(\text{stay} \mid R)$. For example, suppose that $o_T \in O_{\text{leave}}$. By the definition of O_{leave}, o_T can only be observed if the agent has left the room by time step T, and so the trajectory τ must have left the first room. The converse argument does not hold: The agent could leave the first room, re-enter, and then wait until time T. Although one of the doors would have been opened (fig. 6.1), the agent can also open the door without leaving the room, and then realize the same step-T observation. Therefore, this observation doesn't belong to O_{leave}.

Lemma E.31 (Room-status inequalities for MR)**.**

$$p_{\text{decide}}(\mathit{stay} \mid \theta) \leq f_{\text{decide}}(O_{\mathit{stay}} \mid \theta), \tag{E.82}$$

$$\textit{and } f_{\text{decide}}(O_{\mathit{leave}} \mid \theta) \leq p_{\text{decide}}(\mathit{leave} \mid \theta). \tag{E.83}$$

Proof. For any decide,

$$\begin{aligned}
&p_{\text{decide}}(\text{stay} \mid \theta) && \text{(E.84)}\\
&= \underset{\substack{\pi\sim\text{decide}(\theta),\\ \tau\sim\pi|s_0}}{\mathbb{P}} (\tau \text{ stays through step } T) && \text{(E.85)}\\
&= \sum_{o\in\mathcal{O}} \underset{\substack{\pi\sim\text{decide}(\theta),\\ \tau\sim\pi|s_0}}{\mathbb{P}} (o \text{ at step } T \text{ of } \tau) \underset{\substack{\pi\sim\text{decide}(\theta),\\ \tau\sim\pi|s_0}}{\mathbb{P}} \big(\tau \text{ stays} \mid o \text{ at step } T\big) && \text{(E.86)}\\
&= \sum_{o\in O_{T\text{-reach}}} \underset{\substack{\pi\sim\text{decide}(\theta),\\ \tau\sim\pi|s_0}}{\mathbb{P}} (o \text{ at step } T) \underset{\substack{\pi\sim\text{decide}(\theta),\\ \tau\sim\pi|s_0}}{\mathbb{P}} \big(\tau \text{ stays} \mid o \text{ at step } T\big) && \text{(E.87)}\\
&= \sum_{o\in O_{\text{stay}}} \underset{\substack{\pi\sim\text{decide}(\theta),\\ \tau\sim\pi|s_0}}{\mathbb{P}} (o \text{ at step } T) \underset{\substack{\pi\sim\text{decide}(\theta),\\ \tau\sim\pi|s_0}}{\mathbb{P}} \big(\tau \text{ stays} \mid o \text{ at step } T\big) && \text{(E.88)}\\
&\leq \sum_{o\in O_{\text{stay}}} \underset{\substack{\pi\sim\text{decide}(\theta),\\ \tau\sim\pi|s_0}}{\mathbb{P}} (o \text{ at step } T) && \text{(E.89)}\\
&= \underset{\substack{\pi\sim\text{decide}(\theta),\\ \tau\sim\pi|s_0}}{\mathbb{P}} \big(o_T \in O_{\text{stay}}\big) && \text{(E.90)}\\
&=: f_{\text{decide}}(O_{\text{stay}} \mid \theta). && \text{(E.91)}
\end{aligned}$$

Equation (E.87) holds because the definition of $O_{T\text{-reach}}$ ensures that if $o \notin O_{T\text{-reach}}$, then $\mathbb{P}_{\substack{\pi\sim\text{decide}(\theta),\\ \tau\sim\pi|s_0}} \big(o \mid \theta\big) = 0$. Because $o \in O_{T\text{-reach}} \setminus O_{\text{stay}}$ implies that τ left and so

$$\underset{\substack{\pi\sim\text{decide}(\theta),\\ \tau\sim\pi|s_0}}{\mathbb{P}} \big(\tau \text{ stays} \mid o \text{ at step } T\big) = 0,$$

eq. (E.88) follows. Then we have shown eq. (E.82).

For eq. (E.83),

$$f_{\text{decide}}(O_{\text{leave}} \mid \theta) \tag{E.92}$$

$$:= \mathbb{P}_{\substack{\pi\sim\text{decide}(\theta),\\ \tau\sim\pi|s_0}}\left(o_T \in O_{\text{leave}}\right) \tag{E.93}$$

$$= \sum_{o\in O_{\text{leave}}} \mathbb{P}_{\substack{\pi\sim\text{decide}(\theta),\\ \tau\sim\pi|s_0}}\left(o \text{ at step } T\right) \tag{E.94}$$

$$= \sum_{o\in O_{\text{leave}}} \mathbb{P}_{\substack{\pi\sim\text{decide}(\theta),\\ \tau\sim\pi|s_0}}\left(o \text{ at step } T\right) \mathbb{P}_{\substack{\pi\sim\text{decide}(\theta),\\ \tau\sim\pi|s_0}}\big(\tau \text{ leaves by step } T \mid o \text{ at step } T\big) \tag{E.95}$$

$$= \sum_{o\in \mathcal{O}} \mathbb{P}_{\substack{\pi\sim\text{decide}(\theta),\\ \tau\sim\pi|s_0}}\left(o \text{ at step } T\right) \mathbb{P}_{\substack{\pi\sim\text{decide}(\theta),\\ \tau\sim\pi|s_0}}\big(\tau \text{ leaves by step } T \mid o \text{ at step } T\big) \tag{E.96}$$

$$= \mathbb{P}_{\substack{\pi\sim\text{decide}(\theta),\\ \tau\sim\pi|s_0}}\left(\tau \text{ has left the first room by step } T\right) \tag{E.97}$$

$$=: p_{\text{decide}}(\text{leave} \mid \theta). \tag{E.98}$$

Equation (E.95) follows because, since $o \in O_{\text{leave}}$ are only realizable by leaving the first room, this implies $\mathbb{P}_{\substack{\pi\sim\text{decide}(\theta),\\ \tau\sim\pi|s_0}}\big(\tau \text{ leaves by step } T \mid o \text{ at step } T\big) = 1$. Equation (E.96) follows because $O_{\text{leave}} \subseteq \mathcal{O}$, and probabilities are non-negative. Then we have shown eq. (E.83). ◻

Corollary E.32 (Final reward maximizers tend to leave the first room in MR)**.**

$$p_{\max}(\textit{leave} \mid R) \geq^n_{\text{most: } \mathbb{R}^{\mathcal{O}}} p_{\max}(\textit{stay} \mid R). \tag{E.99}$$

Proof. Using lemma E.31 and proposition E.30, apply lemma E.14 with $f_0(R) := p_{\max}(\text{leave} \mid R)$, $f_1(R) := f_{\max}(O_{\text{leave}} \mid R)$, $f_2(R) := f_{\max}(O_{\text{stay}} \mid R)$, $f_3(R) := p_{\max}(\text{stay} \mid R)$ to conclude that $p_{\max}(\text{leave} \mid R) \geq^n_{\text{most: } \mathbb{R}^{\mathcal{O}}} p_{\max}(\text{stay} \mid R)$. ◻

E.3.3 Featurized reward maximization

Consider the featurization function which takes as input an observation $o \in \mathcal{O}$:

$$\operatorname{feat}(o) := \begin{pmatrix} \text{\# of keys in inventory shown by } o \\ \text{\# of swords in inventory shown by } o \\ \text{\# of torches in inventory shown by } o \\ \text{\# of amulets in inventory shown by } o \end{pmatrix}. \tag{E.100}$$

Consider $A_{\text{feat}} := \left\{\operatorname{feat}(o) \mid o \in O_{\text{stay}}\right\}, B_{\text{feat}} := \left\{\operatorname{feat}(o) \mid o \in O_{\text{leave}}\right\}$.

Let $\mathbf{e}_i \in \mathbb{R}^4$ be the standard basis vector with a 1 in entry i and 0 elsewhere. When restricted to the room shown in fig. 6.1, the agent can either acquire the key in the first room and retain it until step T ($\mathbf{e}_1$), or reach time step T empty-handed ($\mathbf{0}$). We conclude that $A_{\text{feat}} = \{\mathbf{e}_1, \mathbf{0}\}$.

For B_{feat}, recall that in section 6.4.2 we assumed the rollout length T to be reasonably large. Then by leaving the room, some realizable trajectory induces o_T displaying an inventory containing only a sword ($\mathbf{e}_2$), or only a torch ($\mathbf{e}_3$), or only an amulet ($\mathbf{e}_4$), or nothing at all ($\mathbf{0}$). Therefore, $\{\mathbf{e}_2, \mathbf{e}_3, \mathbf{e}_4, \mathbf{0}\} \subseteq B_{\text{feat}}$. B_{feat} contains 3 copies of A_{feat} (definition E.7) via involutions $\phi_i : 1 \leftrightarrow i$, $i \neq 1$. Suppose all feature coefficient vectors $\alpha \in \mathbb{R}^4$ are plausible. Then $\Theta := \mathbb{R}^4$.

Let us be more specific about what is entailed by featurized reward maximization. The $\text{decide}_{\text{max}}(\alpha)$ procedure takes α as input and then considers the reward function $o \mapsto \operatorname{feat}(o)^\top \alpha$. Then, $\text{decide}_{\text{max}}$ uniformly randomly chooses an observation $o_T \in O_{T\text{-reach}}$ which maximizes this featurized reward, and then uniformly randomly chooses a policy which implements o_T.

Lemma E.33 (FracOptimal inequalities)**.** *Let $X \subseteq Y' \subseteq Y \subsetneq \mathbb{R}^d$ be finite, and let $\mathbf{u} \in \mathbb{R}^d$. Then*

$$\operatorname{FracOptimal}\left(X \mid Y, \mathbf{u}\right) \leq \operatorname{FracOptimal}\left(X \mid Y', \mathbf{u}\right) \leq \operatorname{FracOptimal}\left(X \cup (Y \setminus Y') \mid Y, \mathbf{u}\right). \tag{E.101}$$

Proof. For finite $X_1 \subsetneq \mathbb{R}^d$, let $\operatorname{Best}\left(X_1 \mid \mathbf{u}\right) := \arg\max_{\mathbf{x}_1 \in X_1} \mathbf{x}_1^\top \mathbf{u}$. Suppose $\mathbf{y}' \in$

$\text{Best}\left(Y' \mid \mathbf{u}\right)$, but $\mathbf{y}' \notin \text{Best}\left(Y \mid \mathbf{u}\right)$. Then for all $\mathbf{a} \in \text{Best}\left(Y' \mid \mathbf{u}\right)$,

$$\mathbf{a}^\top \mathbf{u} = \mathbf{y}'^\top \mathbf{u} < \max_{\mathbf{y} \in Y} \mathbf{y}^\top \mathbf{u}. \tag{E.102}$$

So $\mathbf{a} \notin \text{Best}\left(Y \mid \mathbf{u}\right)$. Then either $\text{Best}\left(Y' \mid \mathbf{u}\right) \subseteq \text{Best}\left(Y \mid \mathbf{u}\right)$, or the two sets are disjoint.

$$\text{FracOptimal}\left(X \mid Y, \mathbf{u}\right) := \frac{\left|\text{Best}\left(Y \mid \mathbf{u}\right) \cap X\right|}{\left|\text{Best}\left(Y \mid \mathbf{u}\right)\right|} \tag{E.103}$$

$$\leq \frac{\left|\text{Best}\left(Y' \mid \mathbf{u}\right) \cap X\right|}{\left|\text{Best}\left(Y' \mid \mathbf{u}\right)\right|} =: \text{FracOptimal}\left(X \mid Y', \mathbf{u}\right) \tag{E.104}$$

If $\text{Best}\left(Y' \mid \mathbf{u}\right) \subseteq \text{Best}\left(Y \mid \mathbf{u}\right)$, then since $X \subseteq Y'$, we have $X \cap \text{Best}\left(Y' \mid \mathbf{u}\right) = X \cap \text{Best}\left(Y \mid \mathbf{u}\right)$. Then in this case, eq. (E.103) has equal numerator and larger denominator than eq. (E.104). On the other hand, if $\text{Best}\left(Y' \mid \mathbf{u}\right) \cap \text{Best}\left(Y \mid \mathbf{u}\right) = \varnothing$, then since $X \subseteq Y'$, $X \cap \text{Best}\left(Y \mid \mathbf{u}\right) = \varnothing$. Then eq. (E.103) equals 0, and eq. (E.104) is non-negative. Either way, eq. (E.104)'s inequality holds. To show the second inequality, we handle the two cases separately.

Subset case. Suppose that $\text{Best}\left(Y' \mid \mathbf{u}\right) \subseteq \text{Best}\left(Y \mid \mathbf{u}\right)$.

$$\frac{\left|\text{Best}\left(Y' \mid \mathbf{u}\right) \cap X\right|}{\left|\text{Best}\left(Y' \mid \mathbf{u}\right)\right|} \leq \frac{\left|\text{Best}\left(Y' \mid \mathbf{u}\right) \cap X\right| + \left|\text{Best}\left(Y \setminus Y' \mid \mathbf{u}\right)\right|}{\left|\text{Best}\left(Y' \mid \mathbf{u}\right)\right| + \left|\text{Best}\left(Y \setminus Y' \mid \mathbf{u}\right)\right|} \tag{E.105}$$

$$= \frac{\left|\text{Best}\left(Y' \mid \mathbf{u}\right) \cap X\right| + \left|\text{Best}\left(Y \setminus Y' \mid \mathbf{u}\right) \cap (Y \setminus Y')\right|}{\left|\text{Best}\left(Y' \mid \mathbf{u}\right)\right| + \left|\text{Best}\left(Y \setminus Y' \mid \mathbf{u}\right)\right|} \tag{E.106}$$

$$= \frac{\left|\text{Best}\left(Y' \mid \mathbf{u}\right) \cap X\right| + \left|\text{Best}\left(Y \mid \mathbf{u}\right) \cap (Y \setminus Y')\right|}{\left|\text{Best}\left(Y' \mid \mathbf{u}\right)\right| + \left|\text{Best}\left(Y \setminus Y' \mid \mathbf{u}\right)\right|} \tag{E.107}$$

$$= \frac{\left|\text{Best}\left(Y' \mid \mathbf{u}\right) \cap X\right| + \left|\text{Best}\left(Y \mid \mathbf{u}\right) \cap (Y \setminus Y')\right|}{\left|\text{Best}\left(Y \mid \mathbf{u}\right)\right|} \tag{E.108}$$

$$= \frac{\left|\text{Best}\left(Y \mid \mathbf{u}\right) \cap X\right| + \left|\text{Best}\left(Y \mid \mathbf{u}\right) \cap (Y \setminus Y')\right|}{\left|\text{Best}\left(Y \mid \mathbf{u}\right)\right|} \tag{E.109}$$

$$= \frac{\left|\text{Best}\left(Y \mid \mathbf{u}\right) \cap (X \cup (Y \setminus Y'))\right|}{\left|\text{Best}\left(Y \mid \mathbf{u}\right)\right|} \tag{E.110}$$

$$=: \text{FracOptimal}\left(X \cup (Y \setminus Y') \mid Y, \mathbf{u}\right). \tag{E.111}$$

Equation (E.105) follows because when $n \leq d, k \geq 0$, we have $\frac{n}{d} \leq \frac{n+k}{d+k}$. For eq. (E.107), since Best $\left(Y' \mid \mathbf{u}\right) \subseteq$ Best $\left(Y \mid \mathbf{u}\right)$, we must have

$$\text{Best}\left(Y \mid \mathbf{u}\right) = \text{Best}\left(Y \setminus Y' \mid \mathbf{u}\right) \cup \text{Best}\left(Y' \mid \mathbf{u}\right).$$

But then

$$\text{Best}\left(Y \mid \mathbf{u}\right) \cap (Y \setminus Y') = \left(\text{Best}\left(Y \setminus Y' \mid \mathbf{u}\right) \cap (Y \setminus Y')\right) \cup \left(\text{Best}\left(Y' \mid \mathbf{u}\right) \cap (Y \setminus Y')\right) \tag{E.112}$$

$$= \text{Best}\left(Y \setminus Y' \mid \mathbf{u}\right) \cap (Y \setminus Y'). \tag{E.113}$$

Then eq. (E.107) follows. Equation (E.108) follows since

$$\text{Best}\left(Y \mid \mathbf{u}\right) = \text{Best}\left(Y \setminus Y' \mid \mathbf{u}\right) \cup \text{Best}\left(Y' \mid \mathbf{u}\right).$$

Equation (E.109) follows since $X \subseteq Y'$, and so

$$\text{Best}\left(Y' \mid \mathbf{u}\right) \cap X = \text{Best}\left(Y \mid \mathbf{u}\right) \cap X.$$

Equation (E.110) follows because $X \subseteq Y'$ is disjoint of $Y \setminus Y'$. We have shown that

$$\text{FracOptimal}\left(X \mid Y', \mathbf{u}\right) \leq \text{FracOptimal}\left(X \cup (Y \setminus Y') \mid Y, \mathbf{u}\right)$$

in this case.

Disjoint case. Suppose that $\text{Best}\left(Y' \mid \mathbf{u}\right) \cap \text{Best}\left(Y \mid \mathbf{u}\right) = \varnothing$.

$$\frac{\left|\text{Best}\left(Y' \mid \mathbf{u}\right) \cap X\right|}{\left|\text{Best}\left(Y' \mid \mathbf{u}\right)\right|} \leq 1 \tag{E.114}$$

$$= \frac{\left|\text{Best}\left(Y \setminus Y' \mid \mathbf{u}\right)\right|}{\left|\text{Best}\left(Y \setminus Y' \mid \mathbf{u}\right)\right|} \tag{E.115}$$

$$= \frac{\left|\text{Best}\left(Y \setminus Y' \mid \mathbf{u}\right) \cap (Y \setminus Y')\right|}{\left|\text{Best}\left(Y \setminus Y' \mid \mathbf{u}\right)\right|} \tag{E.116}$$

$$= \frac{\left|\text{Best}\left(Y \setminus Y' \mid \mathbf{u}\right) \cap (X \cup (Y \setminus Y'))\right|}{\left|\text{Best}\left(Y \setminus Y' \mid \mathbf{u}\right)\right|} \tag{E.117}$$

$$= \frac{\left|\text{Best}\left(Y \mid \mathbf{u}\right) \cap (X \cup (Y \setminus Y'))\right|}{\left|\text{Best}\left(Y \mid \mathbf{u}\right)\right|} \tag{E.118}$$

$$=: \text{FracOptimal}\left(X \cup (Y \setminus Y') \mid Y, \mathbf{u}\right). \tag{E.119}$$

Equation (E.114) follows because $\text{Best}\left(Y' \mid \mathbf{u}\right) \cap X \subseteq \text{Best}\left(Y' \mid \mathbf{u}\right)$. For eq. (E.117), note that we trivially have $\text{Best}\left(Y' \mid \mathbf{u}\right) \cap \text{Best}\left(Y \setminus Y' \mid \mathbf{u}\right) = \varnothing$, and also that $X \subseteq Y'$. Therefore, $\text{Best}\left(Y \setminus Y' \mid \mathbf{u}\right) \cap X = \varnothing$, and eq. (E.117) follows. Finally, the disjointness assumption implies that

$$\max_{\mathbf{y}' \in Y'} \mathbf{y}'^{\top} \mathbf{u} < \max_{\mathbf{y} \in Y} \mathbf{y}^{\top} \mathbf{u}.$$

Therefore, the optimal elements of Y must come exclusively from $Y \setminus Y'$; *i.e.* $\text{Best}\left(Y \mid \mathbf{u}\right) = \text{Best}\left(Y \setminus Y' \mid \mathbf{u}\right)$. Then eq. (E.118) follows, and we have shown that

$$\text{FracOptimal}\left(X \mid Y', \mathbf{u}\right) \leq \text{FracOptimal}\left(X \cup (Y \setminus Y') \mid Y, \mathbf{u}\right)$$

in this case. □

Conjecture E.34 (Generalizing lemma E.33)**.** Lemma E.33 and Turner et al. [99]'s Lemma E.26 have extremely similar functional forms. How can they be unified?

Proposition E.35 (Featurized reward maximizers tend to leave the first room in MR)**.**

$$p_{\max}(\textit{leave} \mid \alpha) \geq^{3}_{\text{most: } \mathbb{R}^4} p_{\max}(\textit{stay} \mid \alpha). \tag{E.120}$$

Proof. We want to show that $f_{\max}(O_{\text{leave}} \mid \alpha) \geq^{n}_{\text{most: } \mathbb{R}^4} f_{\max}(O_{\text{stay}} \mid \alpha)$. Recall that $A_{\text{feat}} = \{\mathbf{e}_1, \mathbf{0}\}, B'_{\text{feat}} := \{\mathbf{e}_2, \mathbf{e}_3, \mathbf{e}_4\} \subseteq B_{\text{feat}}$.

$$p_{\max}(\text{stay} \mid \alpha) \tag{E.121}$$
$$\leq f_{\max}(O_{\text{stay}} \mid \alpha) \tag{E.122}$$
$$:= \underset{\substack{\pi \sim \text{decide}_{\max}(\alpha),\\ \tau \sim \pi \mid s_0}}{\mathbb{P}} \left(o_T \in O_{\text{stay}}\right) \tag{E.123}$$
$$= \underset{\substack{\pi \sim \text{decide}_{\max}(\alpha),\\ \tau \sim \pi \mid s_0}}{\mathbb{P}} \left(o_T \in O_{\text{stay}}, \text{feat}(o_T) \neq \mathbf{0}\right) + \underset{\substack{\pi \sim \text{decide}_{\max}(\alpha),\\ \tau \sim \pi \mid s_0}}{\mathbb{P}} \left(o_T \in O_{\text{stay}}, \text{feat}(o_T) = \mathbf{0}\right) \tag{E.124}$$
$$\leq \text{FracOptimal}\left(\{\mathbf{e}_1\} \mid C_{\text{feat}}, \alpha\right) + \underset{\substack{\pi \sim \text{decide}_{\max}(\alpha),\\ \tau \sim \pi \mid s_0}}{\mathbb{P}} \left(o_T \in O_{\text{stay}}, \text{feat}(o_T) = \mathbf{0}\right) \tag{E.125}$$
$$\leq \text{FracOptimal}\left(\{\mathbf{e}_1\} \mid \{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, \mathbf{e}_4\}, \alpha\right) + \underset{\substack{\pi \sim \text{decide}_{\max}(\alpha),\\ \tau \sim \pi \mid s_0}}{\mathbb{P}} \left(o_T \in O_{\text{stay}}, \text{feat}(o_T) = \mathbf{0}\right) \tag{E.126}$$
$$\leq^{3}_{\text{most: } \mathbb{R}^4_{>0}} \text{FracOptimal}\left(\{\mathbf{e}_2, \mathbf{e}_3, \mathbf{e}_4\} \mid \{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, \mathbf{e}_4\}, \alpha\right) + \underset{\substack{\pi \sim \text{decide}_{\max}(\alpha),\\ \tau \sim \pi \mid s_0}}{\mathbb{P}} \left(o_T \in O_{\text{leave}}, \text{feat}(o_T) = \mathbf{0}\right) \tag{E.127}$$
$$\leq \text{FracOptimal}\left(\{\mathbf{e}_2, \mathbf{e}_3, \mathbf{e}_4\} \cup (C_{\text{feat}} \setminus \{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, \mathbf{e}_4\}) \mid C_{\text{feat}}, \alpha\right) \tag{E.128}$$
$$= \text{FracOptimal}\left(C_{\text{feat}} \setminus \{\mathbf{e}_1\} \mid C_{\text{feat}}, \alpha\right) \tag{E.129}$$
$$\leq \underset{\substack{\pi \sim \text{decide}_{\max}(\alpha),\\ \tau \sim \pi \mid s_0}}{\mathbb{P}} \left(o_T \in O_{\text{leave}}\right) \tag{E.130}$$
$$=: f_{\max}(O_{\text{leave}} \mid \alpha) \tag{E.131}$$
$$\leq p_{\max}(\text{leave} \mid \alpha). \tag{E.132}$$

Equation (E.121) and eq. (E.132) hold by lemma E.31. If $o_T \in O_{\text{stay}}$ is realized by $f_{\max}$

and $\text{feat}(o_T) \neq \mathbf{0}$, then we must have $\text{feat}(o_T) = \{\mathbf{e}_1\}$ be optimal and so the $\mathbf{e}_1$ inventory configuration is realized. Therefore, eq. (E.125) follows. Equation (E.126) follows by applying the first inequality of lemma E.33 with $X := \{\mathbf{e}_1\}, Y' := \{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, \mathbf{e}_4\}, Y := C_{\text{feat}}$.

By applying proposition E.11's item 2 with $A := A_{\text{feat}} = \{\mathbf{e}_1\}$, $B' := B'_{\text{feat}} = \{\mathbf{e}_2, \mathbf{e}_3, \mathbf{e}_4\}$, $C := A \cup B'$, we have

$$\text{FracOptimal}\left(\{\mathbf{e}_1\} \mid \{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, \mathbf{e}_4\}, \alpha\right) \leq^{3}_{\text{most: } \mathbb{R}^4_{>0}} \text{FracOptimal}\left(\{\mathbf{e}_2, \mathbf{e}_3, \mathbf{e}_4\} \mid \{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, \mathbf{e}_4\}, \alpha\right). \tag{E.133}$$

Furthermore, observe that

$$\mathbb{P}_{\substack{\pi \sim \text{decide}_{\max}(\alpha), \\ \tau \sim \pi \mid s_0}}\left(o_T \in O_{\text{stay}}, \text{feat}(o_T) = \mathbf{0}\right) \leq \mathbb{P}_{\substack{\pi \sim \text{decide}_{\max}(\alpha), \\ \tau \sim \pi \mid s_0}}\left(o_T \in O_{\text{leave}}, \text{feat}(o_T) = \mathbf{0}\right) \tag{E.134}$$

because either $\mathbf{0}$ is not optimal (in which case both sides equal 0), or else $\mathbf{0}$ is optimal, in which case the right side is strictly greater. This can be seen by considering how $\text{decide}_{\max}(\alpha)$ uniformly randomly chooses an observation in which the agent ends up with an empty inventory. As argued previously, the vast majority of such observations can only be induced by leaving the first room.

Combining eq. (E.133) and eq. (E.134), eq. (E.127) follows. Equation (E.128) follows by applying the second inequality of lemma E.33 with $X := \{\mathbf{e}_2, \mathbf{e}_3, \mathbf{e}_4\}$, $Y' := \{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, \mathbf{e}_4\}$, $Y := C_{\text{feat}}$. If $\text{feat}(o_T) \in B_{\text{feat}}$ is realized by $f_{\max}$, then by the definition of B_{feat}, $o_T \in O_{\text{leave}}$ is realized, and so eq. (E.130) follows.

Then by applying lemma E.14 with

$$f_0(\alpha) := p_{\max}(\text{leave} \mid \alpha), \tag{E.135}$$

$$f_1(\alpha) := \text{FracOptimal}\left(\{\mathbf{e}_1\} \mid \{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, \mathbf{e}_4\}, \alpha\right), \tag{E.136}$$

$$f_2(\alpha) := \text{FracOptimal}\left(\{\mathbf{e}_2, \mathbf{e}_3, \mathbf{e}_4\} \mid \{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, \mathbf{e}_4\}, \alpha\right), \tag{E.137}$$

$$f_3(\alpha) := p_{\max}(\text{stay} \mid \alpha), \tag{E.138}$$

we conclude that $p_{\max}(\text{leave} \mid \alpha) \geq^3_{\text{most: } \mathbb{R}^4_{>0}} p_{\max}(\text{stay} \mid \alpha)$. □

Lastly, note that if $\mathbf{0} \in \Theta$ and $f(A \mid \mathbf{0}) > f(B \mid \mathbf{0})$, f cannot be even be simply retargetable for the Θ parameter set. This is because $\forall \phi \in S_d$, $\phi \cdot \mathbf{0} = \mathbf{0}$. For example, inductive bias ensures that, absent a reward signal, learned policies tend to stay in the initial room in MR. This is one reason why section 6.4.4's analysis of the policy tendencies of reinforcement learning excludes the all-zero reward function.

E.4 Lower bounds on MDP power-seeking incentives for optimal policies

Turner et al. [99] prove conditions under which *at least half* of the orbit of every reward function incentivizes power-seeking behavior. For example, in fig. E.2, they prove that avoiding $\varnothing$ maximizes average per-timestep reward for at least half of reward functions. Roughly, there are more self-loop states ($\varnothing$, $\ell_{\swarrow}$, $r_{\searrow}$, $r_{\nearrow}$) available if the agent goes `left` or `right` instead of up towards $\varnothing$. We strengthen this claim, with corollary E.47 showing that for *at least three-quarters* of the orbit of every reward function, it is average-optimal to avoid $\varnothing$.

Therefore, we answer Turner et al. [99]'s open question of whether increased number of environmental symmetries quantitatively strengthens the degree to which power-seeking is incentivized. The answer is *yes*. In particular, it may be the case that only one in a million state-based reward functions makes it average-optimal for Pac-Man to die immediately.

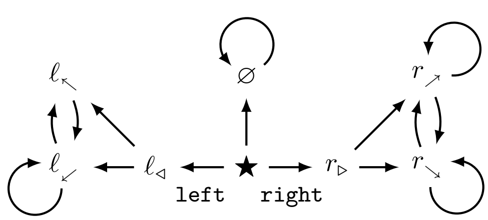


Figure E.2: A toy MDP for reasoning about power-seeking tendencies. *Reproduced from Turner et al. [99].*

We will briefly restate several definitions needed for our key results, theorem E.46 and corollary E.47. For explanation, see Turner et al. [99].

Definition E.36 (Non-dominated linear functionals)**.** Let $X \subsetneq \mathbb{R}^{|\mathcal{S}|}$ be finite. $\operatorname{ND}(X) := \left\{\mathbf{x} \in X \mid \exists \mathbf{r} \in \mathbb{R}^{|\mathcal{S}|} : \mathbf{x}^\top \mathbf{r} > \max_{\mathbf{x}' \in X \setminus \{\mathbf{x}\}} \mathbf{x}'^\top \mathbf{r}\right\}$.

Definition E.37 (Bounded reward function distribution)**.** $\mathfrak{D}_{\text{bound}}$ is the set of bounded-support probability distributions $\mathcal{D}_{\text{bound}}$.

Remark. When $n = 1$, lemma E.38 reduces to the first part of Turner et al. [99]'s lemma D.29, and lemma E.40 reduces to the first part of Turner et al. [99]'s lemma D.33.

Lemma E.38 (Quantitative expectation superiority lemma)**.** *Let $A, B \subsetneq \mathbb{R}^d$ be finite and let $g : \mathbb{R} \to \mathbb{R}$ be a (total) increasing function. Suppose B contains n copies of* $\operatorname{ND}(A)$*. Then*

$$\mathop{\mathbb{E}}_{\mathbf{r} \sim \mathcal{D}_{bound}}\left[g\left(\max_{\mathbf{b} \in B} \mathbf{b}^\top \mathbf{r}\right)\right] \geq^n_{\text{most: } \mathfrak{D}_{bound}} \mathop{\mathbb{E}}_{\mathbf{r} \sim \mathcal{D}_{bound}}\left[g\left(\max_{\mathbf{a} \in A} \mathbf{a}^\top \mathbf{r}\right)\right]. \tag{E.139}$$

Proof. Because $g : \mathbb{R} \to \mathbb{R}$ is increasing, it is measurable (as is max).

Let $L := \inf_{\mathbf{r} \in \operatorname{supp}(\mathcal{D}_{\text{bound}})} \max_{\mathbf{x} \in X} \mathbf{x}^\top \mathbf{r}, U := \sup_{\mathbf{r} \in \operatorname{supp}(\mathcal{D}_{\text{bound}})} \max_{\mathbf{x} \in X} \mathbf{x}^\top \mathbf{r}$. Both exist because $\mathcal{D}_{\text{bound}}$ has bounded support. Furthermore, since g is monotone increasing, it is bounded $[g(L), g(U)]$ on $[L, U]$. Therefore, g is measurable and bounded on each $\operatorname{supp}(\mathcal{D}_{\text{bound}})$, and so the relevant expectations exist for all $\mathcal{D}_{\text{bound}}$.

For finite $X \subsetneq \mathbb{R}^d$, let $f(X \mid \mathbf{u}) := g(\max_{\mathbf{x} \in X} \mathbf{x}^\top \mathbf{u})$. By lemma E.24, f is invariant under joint permutation by S_d. Furthermore, f is measurable because g and max are. Therefore, apply lemma E.18 to conclude that $f(X \mid \mathcal{D}_{\text{bound}}) := \mathbb{E}_{\mathbf{u} \sim \mathcal{D}_{\text{bound}}}\left[g(\max_{\mathbf{x} \in X} \mathbf{x}^\top \mathbf{u})\right]$ is also invariant under joint permutation by S_d (with f being bounded when restricted to $\operatorname{supp}(\mathcal{D}_{\text{bound}})$). Lastly, if $X' \subseteq X$, $f(X' \mid \mathcal{D}_{\text{bound}}) \leq f(X \mid \mathcal{D}_{\text{bound}})$ because g is increasing.

$$\mathop{\mathbb{E}}_{\mathbf{u} \sim \mathcal{D}_{\text{bound}}}\left[g\left(\max_{\mathbf{a} \in A} \mathbf{a}^\top \mathbf{u}\right)\right] = \mathop{\mathbb{E}}_{\mathbf{u} \sim \mathcal{D}_{\text{bound}}}\left[g\left(\max_{\mathbf{a} \in \operatorname{ND}(A)} \mathbf{a}^\top \mathbf{u}\right)\right] \tag{E.140}$$

$$\leq^n_{\text{most: } \mathfrak{D}_{\text{any}}} \mathop{\mathbb{E}}_{\mathbf{r} \sim \mathcal{D}_{\text{bound}}}\left[g\left(\max_{\mathbf{b} \in B} \mathbf{b}^\top \mathbf{r}\right)\right]. \tag{E.141}$$

Equation (E.140) follows by corollary D.16 of [99]. Equation (E.141) follows by applying lemma E.22 with f as defined above with the $\phi_1, \ldots, \phi_n$ guaranteed by the copy assumption. $\square$

Definition E.39 (Linear functional optimality probability [99])**.** For finite $A, B \subsetneq \mathbb{R}^{|\mathcal{S}|}$, the *probability under $\mathcal{D}_{any}$ that A is optimal over B* is

$$p_{\mathcal{D}_{\text{any}}}(A \geq B) := \mathbb{P}_{\mathbf{r}\sim\mathcal{D}_{\text{any}}}\left(\max_{\mathbf{a}\in A} \mathbf{a}^\top \mathbf{r} \geq \max_{\mathbf{b}\in B} \mathbf{b}^\top \mathbf{r}\right).$$

Lemma E.40 (Quantitative optimality probability superiority lemma)**.** *Let $A, B, C \subsetneq \mathbb{R}^d$ be finite and let Z satisfy $\operatorname{ND}(C) \subseteq Z \subseteq C$. Suppose that B contains n copies of $\operatorname{ND}(A)$ via involutions ϕ_i. Furthermore, let $B_{extra} := B \setminus \left(\cup_{i=1}^n \phi_i \cdot \operatorname{ND}(A)\right)$; suppose that for all i, $\phi_i \cdot \left(Z \setminus B_{extra}\right) = Z \setminus B_{extra}$.*

Then $p_{\mathcal{D}_{any}}(B \geq C) \geq^n_{\text{most: } \mathfrak{D}_{any}} p_{\mathcal{D}_{any}}(A \geq C)$.

Proof. For finite $X, Y \subsetneq \mathbb{R}^d$, let

$$g(X, Y \mid \mathcal{D}_{\text{any}}) := p_{\mathcal{D}_{\text{any}}}(X \geq Y) = \mathbb{E}_{\mathbf{u}\sim\mathcal{D}_{\text{any}}}\left[\mathbb{1}_{\max_{\mathbf{x}\in X} \mathbf{x}^\top \mathbf{u} \geq \max_{\mathbf{y}\in Y} \mathbf{y}^\top \mathbf{u}}\right].$$

By the proof of item 1 of proposition E.11, g is the expectation of a $\mathbf{u}$-measurable function. g is an EU function, and so lemma E.24 shows that it is invariant to joint permutation by ϕ_i. Letting $f_Y(X \mid \mathcal{D}_{\text{any}}) := g(X, Y \mid \mathcal{D}_{\text{any}})$, lemma E.23 shows that $f_Y(X \mid \mathcal{D}_{\text{any}}) = f_Y(\phi_i \cdot X \mid \phi_i \cdot \mathcal{D}_{\text{any}})$ whenever the ϕ_i satisfy $\phi_i \cdot Y = Y$.

Furthermore, if $X' \subseteq X$, then $f_Y(X' \mid \mathcal{D}_{\text{any}}) \leq f_Y(X \mid \mathcal{D}_{\text{any}})$.

$$\begin{aligned}
p_{\mathcal{D}_{\text{any}}}(A \geq C) &= p_{\mathcal{D}_{\text{any}}}\left(\operatorname{ND}(A) \geq C\right) && \text{(E.142)}\\
&\leq p_{\mathcal{D}_{\text{any}}}\left(\operatorname{ND}(A) \geq Z \setminus B_{\text{extra}}\right) && \text{(E.143)}\\
&\leq^n_{\text{most: } \mathfrak{D}_{\text{any}}} p_{\mathcal{D}_{\text{any}}}\left(B \geq Z \setminus B_{\text{extra}}\right) && \text{(E.144)}\\
&\leq p_{\mathcal{D}_{\text{any}}}(B \cup B_{\text{extra}} \geq Z) && \text{(E.145)}\\
&= p_{\mathcal{D}_{\text{any}}}(B \geq Z) && \text{(E.146)}\\
&= p_{\mathcal{D}_{\text{any}}}(B \geq C). && \text{(E.147)}
\end{aligned}$$

Equation (E.142) follows by Turner et al. [99]'s lemma D.17's item 2 with $X \coloneqq A$, $X' \coloneqq \mathrm{ND}(A)$ (similar reasoning holds for C and Z in eq. (E.147)). Equation (E.143) follows by the first inequality of lemma D.31 of [99] with $X \coloneqq A, Y \coloneqq C, Y' \coloneqq Z \setminus B_{\text{extra}}$. Equation (E.144) follows by applying lemma E.22 with the $f_{Z \setminus B_{\text{extra}}}$ defined above. Equation (E.145) follows by the second inequality of lemma D.31 of [99] with $X \coloneqq A, Y \coloneqq Z, Y' \coloneqq Z \setminus B_{\text{extra}}$. Equation (E.146) follows because $B_{\text{extra}} \subseteq B$.

Letting $f_0(\mathcal{D}_{\text{any}}) \coloneqq p_{\mathcal{D}_{\text{any}}}(A \geq C)$, $f_1(\mathcal{D}_{\text{any}}) \coloneqq p_{\mathcal{D}_{\text{any}}}\left(\mathrm{ND}(A) \geq Z \setminus B_{\text{extra}}\right)$, $f_2(\mathcal{D}_{\text{any}}) \coloneqq p_{\mathcal{D}_{\text{any}}}\left(B \geq Z \setminus B_{\text{extra}}\right)$, $f_3(\mathcal{D}_{\text{any}}) \coloneqq p_{\mathcal{D}_{\text{any}}}(B \geq C)$, apply lemma E.14 to conclude that

$$p_{\mathcal{D}_{\text{any}}}(A \geq C) \leq^n_{\text{most: } \mathfrak{D}_{\text{any}}} p_{\mathcal{D}_{\text{any}}}(B \geq C).$$

□

Definition E.41 (Rewardless MDP [99])**.** $\langle \mathcal{S}, \mathcal{A}, T \rangle$ is a rewardless MDP with finite state and action spaces $\mathcal{S}$ and $\mathcal{A}$, and stochastic transition function $T : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$. We treat the discount rate γ as a variable with domain $[0, 1]$.

Definition E.42 (1-cycle states [99])**.** Let $\mathbf{e}_s \in \mathbb{R}^{|\mathcal{S}|}$ be the standard basis vector for state s, such that there is a 1 in the entry for state s and 0 elsewhere. State s is a *1-cycle* if $\exists a \in \mathcal{A} : T(s, a) = \mathbf{e}_s$. State s is a *terminal state* if $\forall a \in \mathcal{A} : T(s, a) = \mathbf{e}_s$.

Definition E.43 (State visit distribution [91])**.** $\Pi \coloneqq \mathcal{A}^{\mathcal{S}}$, the set of stationary deterministic policies. The *visit distribution* induced by following policy π from state s at discount rate $\gamma \in [0, 1)$ is $\mathbf{f}^{\pi,s}(\gamma) \coloneqq \sum_{t=0}^{\infty} \gamma^t \mathbb{E}_{s_t \sim \pi | s}\left[\mathbf{e}_{s_t}\right]$. $\mathbf{f}^{\pi,s}$ is a *visit distribution function*; $\mathcal{F}(s) \coloneqq \{\mathbf{f}^{\pi,s} \mid \pi \in \Pi\}$.

Definition E.44 (Recurrent state distributions [68])**.** The *recurrent state distributions* which can be induced from state s are $\mathrm{RSD}(s) \coloneqq \left\{\lim_{\gamma \to 1}(1-\gamma)\mathbf{f}^{\pi,s}(\gamma) \mid \pi \in \Pi\right\}$. $\mathrm{RSD}_{\text{nd}}(s)$ is the set of RSDs which strictly maximize average reward for some reward function.

Definition E.45 (Average-optimal policies [99])**.** The *average-optimal policy set* for reward function R is $\Pi^{\text{avg}}(R) \coloneqq \left\{\pi \in \Pi \mid \forall s \in \mathcal{S} : \mathbf{d}^{\pi,s} \in \arg\max_{\mathbf{d} \in \mathrm{RSD}(s)} \mathbf{d}^\top \mathbf{r}\right\}$ (the policies which induce optimal RSDs at all states). For $D \subseteq \mathrm{RSD}(s)$, the *average optimality probability* is $\mathbb{P}_{\mathcal{D}_{\text{any}}}(D, \text{average}) \coloneqq \mathbb{P}_{R \sim \mathcal{D}_{\text{any}}}\left(\exists \mathbf{d}^{\pi,s} \in D : \pi \in \Pi^{\text{avg}}(R)\right)$.

Remark. Theorem E.46 generalizes the first claim of Turner et al. [99]'s theorem 5.29, and corollary E.47 generalizes the first claim of Turner et al. [99]'s corollary 5.30.

Theorem E.46 (Quantitatively, average-optimal policies tend to end up in "larger" sets of RSDs)**.** *Let $D', D \subseteq \mathrm{RSD}(s)$. Suppose that D contains n copies of D' and that the sets $D' \cup D$ and $\mathrm{RSD}_{\mathrm{nd}}(s) \setminus \left(D' \cup D\right)$ have pairwise orthogonal vector elements (*i.e. *pairwise disjoint vector support). Then $\mathbb{P}_{\mathcal{D}_{any}}\left(D', \text{average}\right) \leq^n_{\text{most: } \mathfrak{D}_{any}} \mathbb{P}_{\mathcal{D}_{any}}(D, \text{average})$.*

Proof. Let $D_i := \phi_i \cdot D'$, where $D_i \subseteq D$ by assumption.
Let $S := \left\{s' \in \mathcal{S} \mid \max_{\mathbf{d} \in D' \cup D} \mathbf{d}^\top \mathbf{e}_{s'} > 0\right\}$.
Define

$$\phi_i'(s') := \begin{cases} \phi_i(s') & \text{if } s' \in S \\ s' & \text{else.} \end{cases} \tag{E.148}$$

Since ϕ_i is an involution, ϕ_i' is also an involution. Furthermore, $\phi_i' \cdot D' = D_i$, $\phi_i' \cdot D_i = D'$, and $\phi_i' \cdot D_j = D_j$ for $j \neq i$ because we assumed that these equalities hold for ϕ_i, and $D', D_i, D_j \subseteq D' \cup D$ and so the vectors of these sets have support contained in S.

Let $D^* := D' \cup_{i=1}^n D_i \cup \Big(\mathrm{RSD}_{\mathrm{nd}}(s) \setminus \left(D' \cup D\right)\Big)$. By an argument mirroring that in the proof of theorem 5.29 in Turner et al. [99] and using the fact that $\phi_i' \cdot D_j = D_j$ for all $i \neq j$, $\phi_i' \cdot D^* = D^*$. Consider $Z := \left(\mathrm{RSD}_{\mathrm{nd}}(s) \setminus (D' \cup D)\right) \cup D' \cup D$. First, $Z \subseteq \mathrm{RSD}(s)$ by definition. Second, $\mathrm{RSD}_{\mathrm{nd}}(s) = \mathrm{RSD}_{\mathrm{nd}}(s) \setminus (D' \cup D) \cup \left(\mathrm{RSD}_{\mathrm{nd}}(s) \cap D'\right) \cup \left(\mathrm{RSD}_{\mathrm{nd}}(s) \cap D\right) \subseteq Z$. Note that $D^* = Z \setminus (D \setminus \cup_{i=1}^n D_i)$.

$$\mathbb{P}_{\mathcal{D}_{\text{any}}}\left(D', \text{average}\right) = p_{\mathcal{D}_{\text{any}}}\left(D' \geq \mathrm{RSD}(s)\right) \tag{E.149}$$
$$\leq^n_{\text{most: } \mathfrak{D}_{\text{any}}} p_{\mathcal{D}_{\text{any}}}\left(D \geq \mathrm{RSD}(s)\right) \tag{E.150}$$
$$= \mathbb{P}_{\mathcal{D}_{\text{any}}}(D, \text{average}). \tag{E.151}$$

Since $\phi_i' \cdot D' \subseteq D$ and $\mathrm{ND}\left(D'\right) \subseteq D'$, $\phi_i' \cdot \mathrm{ND}\left(D'\right) \subseteq D$ and so D contains n copies of $\mathrm{ND}\left(D'\right)$ via involutions ϕ_i'. Then eq. (E.150) holds by applying lemma E.40 with $A := D'$, $B_i := D_i$ for all $i = 1, \ldots, n$, $B := D, C := \mathrm{RSD}(s)$, Z as defined above, and involutions ϕ_i' which satisfy $\phi_i' \cdot \left(Z \setminus (B \setminus \cup_{i=1}^n B_i)\right) = \phi_i' \cdot D^* = D^* = Z \setminus (B \setminus \cup_{i=1}^n B_i)$. $\square$

Corollary E.47 (Quantitatively, average-optimal policies tend not to end up in any given 1-cycle)**.** *Let* $D' := \left\{\mathbf{e}_{s'_1}, \ldots, \mathbf{e}_{s'_k}\right\}, D_r := \left\{\mathbf{e}_{s_1}, \ldots, \mathbf{e}_{s_{n \cdot k}}\right\} \subseteq \mathrm{RSD}\left(s\right)$ *be disjoint, for* $n \geq 1, k \geq 1$*. Then* $\mathbb{P}_{\mathcal{D}_{any}}\left(D', \text{average}\right) \leq^n_{\text{most: } \mathfrak{D}_{any}} \mathbb{P}_{\mathcal{D}_{any}}\left(\mathrm{RSD}\left(s\right) \setminus D', \text{average}\right)$*.*

Proof. For each $i \in \{1, \ldots, n\}$, let

$$\begin{aligned}
\phi_i &:= (s'_1 \;\; s_{(i-1)\cdot k+1}) \cdots (s'_k \;\; s_{(i-1)\cdot k+k}),\\
D_i &:= \left\{\mathbf{e}_{s_{(i-1)\cdot k+1}}, \ldots, \mathbf{e}_{s_{(i-1)\cdot k+k}}\right\},\\
D &:= \mathrm{RSD}\left(s\right) \setminus D'.
\end{aligned}$$

Each $D_i \subseteq D_r \subseteq \mathrm{RSD}\left(s\right) \setminus D'$ by disjointness of D' and D_r.

D contains n copies of D' via involutions $\phi_1, \ldots, \phi_n$. $D' \cup D = \mathrm{RSD}\left(s\right)$ and $\mathrm{RSD}_{\text{nd}}\left(s\right) \setminus \mathrm{RSD}\left(s\right) = \emptyset$ trivially have pairwise orthogonal vector elements.

Apply theorem E.46 to conclude that

$$\mathbb{P}_{\mathcal{D}_{\text{any}}}\left(D', \text{average}\right) \leq^n_{\text{most: } \mathfrak{D}_{\text{any}}} \mathbb{P}_{\mathcal{D}_{\text{any}}}\left(\mathrm{RSD}\left(s\right) \setminus D', \text{average}\right).$$

□

Let $A := \{\mathbf{e}_1, \mathbf{e}_2\}, B \subseteq \mathbb{R}^5$, $C := A \cup B$. Conjecture E.48 conjectures that *e.g.*

$$p_{\mathcal{D}'}\left(B \geq C\right) \geq^{\frac{3}{2}}_{\text{most: } \mathfrak{D}_{\text{any}}} p_{\mathcal{D}'}\left(A \geq C\right).$$

Conjecture E.48 (Fractional quantitative optimality probability superiority lemma)**.** Let A, B, $C \subsetneq \mathbb{R}^d$ be finite. If $A = \bigcup_{j=1}^m A_j$ and $\bigcup_{i=1}^n B_i \subseteq B$ such that for each A_j, B contains n copies $(B_1, \ldots, B_n)$ of A_j via involutions ϕ_{ji} which *also* fix $\phi_{ji} \cdot A_{j'} = A_{j'}$ for $j' \neq j$, then

$$p_{\mathcal{D}_{\text{any}}}\left(B \geq C\right) \geq^{\frac{n}{m}}_{\text{most: } \mathfrak{D}_{\text{any}}} p_{\mathcal{D}_{\text{any}}}\left(A \geq C\right).$$

We suspect that any proof of the conjecture should generalize lemma E.20 to the fractional set copy containment case.

Table E.1: Orbit-level incentives across 4 decision-making functions.

Utility function $\mathbf{u}'$	$\spadesuit^{B}\ \heartsuit^{B}\ \diamondsuit^{A}$ 10, 5, 0	$\spadesuit^{B}\ \heartsuit^{B}\ \diamondsuit^{A}$ 10, 0, 5	$\spadesuit^{B}\ \heartsuit^{B}\ \diamondsuit^{A}$ 5, 10, 0	$\spadesuit^{B}\ \heartsuit^{B}\ \diamondsuit^{A}$ 5, 0, 10	$\spadesuit^{B}\ \heartsuit^{B}\ \diamondsuit^{A}$ 0, 10, 5	$\spadesuit^{B}\ \heartsuit^{B}\ \diamondsuit^{A}$ 0, 5, 10
$\operatorname{Optimal}\left(\left\{\mathbf{e}_{\spadesuit^B}, \mathbf{e}_{\heartsuit^B}\right\} \mid C_{\text{cards}}, \mathbf{u}'\right)$	1	1	1	0	1	0
$\operatorname{Optimal}\left(\left\{\mathbf{e}_{\diamondsuit^A}\right\} \mid C_{\text{cards}}, \mathbf{u}'\right)$	0	0	0	1	0	1

(a) Dark gray columns indicate utility function permutations $\mathbf{u}'$ for which $\operatorname{Optimal}\left(B_{\text{cards}} \mid C_{\text{cards}}, \mathbf{u}'\right) > \operatorname{Optimal}\left(A_{\text{cards}} \mid C_{\text{cards}}, \mathbf{u}'\right)$, while white indicates that the opposite strict inequality holds.

Utility function $\mathbf{u}'$	$\spadesuit^{B}\ \heartsuit^{B}\ \diamondsuit^{A}$ 10, 5, 0	$\spadesuit^{B}\ \heartsuit^{B}\ \diamondsuit^{A}$ 10, 0, 5	$\spadesuit^{B}\ \heartsuit^{B}\ \diamondsuit^{A}$ 5, 10, 0	$\spadesuit^{B}\ \heartsuit^{B}\ \diamondsuit^{A}$ 5, 0, 10	$\spadesuit^{B}\ \heartsuit^{B}\ \diamondsuit^{A}$ 0, 10, 5	$\spadesuit^{B}\ \heartsuit^{B}\ \diamondsuit^{A}$ 0, 5, 10
$\operatorname{AntiOpt}\left(\left\{\mathbf{e}_{\spadesuit^B}, \mathbf{e}_{\heartsuit^B}\right\} \mid C_{\text{cards}}, \mathbf{u}'\right)$	0	1	0	1	1	1
$\operatorname{AntiOpt}\left(\left\{\mathbf{e}_{\diamondsuit^A}\right\} \mid C_{\text{cards}}, \mathbf{u}'\right)$	1	0	1	0	0	0

(b) Utility-minimizing outcome selection probability.

Utility function $\mathbf{u}'$	$\spadesuit^{B}\ \heartsuit^{B}\ \diamondsuit^{A}$ 10, 5, 0	$\spadesuit^{B}\ \heartsuit^{B}\ \diamondsuit^{A}$ 10, 0, 5	$\spadesuit^{B}\ \heartsuit^{B}\ \diamondsuit^{A}$ 5, 10, 0	$\spadesuit^{B}\ \heartsuit^{B}\ \diamondsuit^{A}$ 5, 0, 10	$\spadesuit^{B}\ \heartsuit^{B}\ \diamondsuit^{A}$ 0, 10, 5	$\spadesuit^{B}\ \heartsuit^{B}\ \diamondsuit^{A}$ 0, 5, 10
$\operatorname{Boltzmann}_1\left(\left\{\mathbf{e}_{\spadesuit^B}, \mathbf{e}_{\heartsuit^B}\right\} \mid C_{\text{cards}}, \mathbf{u}'\right)$	1	.993	1	.007	.993	.007
$\operatorname{Boltzmann}_1\left(\left\{\mathbf{e}_{\diamondsuit^A}\right\} \mid C_{\text{cards}}, \mathbf{u}'\right)$	.000	.007	.000	.993	.007	.993

(c) Boltzmann selection probabilities for $T = 1$, rounded to three significant digits.

Utility function $\mathbf{u}'$	$\spadesuit^{B}\ \heartsuit^{B}\ \diamondsuit^{A}$ 10, 5, 0	$\spadesuit^{B}\ \heartsuit^{B}\ \diamondsuit^{A}$ 10, 0, 5	$\spadesuit^{B}\ \heartsuit^{B}\ \diamondsuit^{A}$ 5, 10, 0	$\spadesuit^{B}\ \heartsuit^{B}\ \diamondsuit^{A}$ 5, 0, 10	$\spadesuit^{B}\ \heartsuit^{B}\ \diamondsuit^{A}$ 0, 10, 5	$\spadesuit^{B}\ \heartsuit^{B}\ \diamondsuit^{A}$ 0, 5, 10
$\operatorname{Satisfice}_3\left(\left\{\mathbf{e}_{\spadesuit^B}, \mathbf{e}_{\heartsuit^B}\right\} \mid C_{\text{cards}}, \mathbf{u}'\right)$	1	.5	1	.5	.5	.5
$\operatorname{Satisfice}_3\left(\left\{\mathbf{e}_{\diamondsuit^A}\right\} \mid C_{\text{cards}}, \mathbf{u}'\right)$	0	.5	0	.5	.5	.5

(d) A satisficer uniformly randomly selects an outcome lottery with expected utility greater than or equal to the threshold t. Here, $t = 3$. When $\operatorname{Satisfice}_3\left(\left\{\mathbf{e}_{\spadesuit^B}, \mathbf{e}_{\heartsuit^B}\right\} \mid C_{\text{cards}}, \mathbf{u}'\right) = \operatorname{Satisfice}_3\left(\left\{\mathbf{e}_{\diamondsuit^A}\right\} \mid C_{\text{cards}}, \mathbf{u}'\right)$, the column is colored medium gray.

Targeting parameter θ	$f(\{\blacklozenge^{\text{A}}\} \mid \theta)$	$f(\{\heartsuit^{\text{B}}\} \mid \theta)$	$f(\{\spadesuit^{\text{B}}\} \mid \theta)$	$f(\{\heartsuit^{\text{B}}, \spadesuit^{\text{B}}\} \mid \theta)$
$\theta' \coloneqq 1\mathbf{e}_1 + 3\mathbf{e}_2 + 2\mathbf{e}_3$	1	0	0	0
$\phi_1 \cdot \theta' = \phi_2 \cdot \theta'' \coloneqq 3\mathbf{e}_1 + 1\mathbf{e}_2 + 2\mathbf{e}_3$	0	2	2	2
$\phi_2 \cdot \theta' \coloneqq 2\mathbf{e}_1 + 3\mathbf{e}_2 + 1\mathbf{e}_3$	0	2	2	2
$\theta'' \coloneqq 2\mathbf{e}_1 + 1\mathbf{e}_2 + 3\mathbf{e}_3$	1	0	0	0
$\phi_1 \cdot \theta'' \coloneqq 1\mathbf{e}_1 + 2\mathbf{e}_2 + 3\mathbf{e}_3$	0	2	2	2
$\theta^\star \coloneqq 3\mathbf{e}_1 + 2\mathbf{e}_2 + 1\mathbf{e}_3$	1	0	0	0

Table E.2: Let $\phi_1 \coloneqq \blacklozenge^{\text{A}} \leftrightarrow \heartsuit^{\text{B}}, \phi_2 \coloneqq \blacklozenge^{\text{A}} \leftrightarrow \spadesuit^{\text{B}}$. We tabularly define a function f which meets all requirements of lemma E.20, except for item 4: letting $j \coloneqq 2$, $f(B_2^\star \mid \phi_1 \cdot \theta') = 2 > 0 = f(B_2^\star \mid \theta')$. Although $f(B \mid \theta) \geq^1_{\text{most: } S_3 \cdot \theta} f(A \mid \theta)$, it is not true that $f(B \mid \theta^*) \geq^2_{\text{most: } S_3 \cdot \theta} f(A \mid \theta^*)$. Therefore, item 4 is generally required.

F

Additional Theoretical Results

This chapter contains results on:

- Mdp structure and representation (seemingly building towards a novel category-theoretic treatment of mdps),
- The expressivity of Markov reward [1],
- The properties of optimal value functions,
- Characterizing the mdps in which there exist reward functions whose optimal policy set depends on the discount rate,
- An optimal value-based distance metric on the state space,
- An improved rl regret formalism and a no-free-lunch theorem prohibiting simultaneously bounding this regret for all reward functions,

- Power-seeking incentives for agents with discount rate very close to 1, and
- Power-seeking incentives for ϵ-optimal agents.

F.1 Basic results on visit distributions

The traditional view of Markov decision processes (MDPs) takes for granted a reward function and discount rate, and considers the on-policy value function induced by solving the Bellman equations for a given policy. The "dual" view considers the available *state visit distribution functions.*

I think that the dual view deserves more prominence: each reward function and discount rate tuple merely imposes a preordering over policies *given* the existing dynamics. The dual view regards the environment structure as primary, and the optimization objective as secondary. Studying this environmental structure makes available unexplored areas of basic MDP theory.

We freely rely on the theorems and definitions of *Optimal Policies Tend to Seek Power* [99] and *Avoiding Side Effects in Complex Environments* [98], as they do not depend on these results. All MDPs are assumed to have finite state and action spaces.

Lemma F.1 (Each state has a visit distribution function)**.** $\forall s \in \mathcal{S} : 1 \leq \left|\mathcal{F}(s)\right|$.

Proof. $|\Pi| = |\mathcal{A}|^{|\mathcal{S}|}$. Each MDP must have at least one action and one state, and so Π is not empty.

Let $\pi \in \Pi, \gamma \in [0,1), s \in \mathcal{S}$ and consider $\mathbf{T}^\pi$, the transition matrix which π induces. $\mathbf{T}^\pi$ is left stochastic and hence has spectral radius at most 1. Therefore, $\gamma\mathbf{T}^\pi$ has spectral radius strictly less than 1. This means that its Neumann series $\sum_{t=0}^\infty \left(\gamma\mathbf{T}^\pi\right)^t = \left(\mathbf{I} - \gamma\mathbf{T}^\pi\right)^{-1}$, where $\mathbf{I}$ is the $|\mathcal{S}|$-dimensional identity matrix.

$$\mathbf{f}^{\pi,s}(\gamma) \coloneqq \sum_{t=0}^{\infty} \gamma^t \mathop{\mathbb{E}}_{s'}\left[\mathbf{e}_{s'} \mid \pi \text{ followed for } t \text{ steps from } s\right]. \tag{F.1}$$

By inspection, $\mathbf{f}^{\pi,s}(\gamma) = \mathbf{I}\mathbf{e}_s + \left(\gamma\mathbf{T}^\pi\right)\mathbf{e}_s + \left(\gamma\mathbf{T}^\pi\right)^2\mathbf{e}_s + \cdots$, so $\mathbf{f}^{\pi,s}(\gamma) = \left(\mathbf{I} - \gamma\mathbf{T}^\pi\right)^{-1}\mathbf{e}_s$. $\square$

Corollary F.2 ($\mathbf{f}^{\pi,s}$ identity [68])**.** $\mathbf{f}^{\pi,s}(\gamma) = (\mathbf{I} - \gamma\mathbf{T}^{\pi})^{-1}\,\mathbf{e}_s$.

Lemma F.3 (Strictly increasing visit frequency)**.** *If* $\exists\gamma \in [0,1) : \mathbf{f}^{\pi,s}_{s'}(\gamma) \notin \{0,1\}$*, then* $\mathbf{f}^{\pi,s}_{s'}(\gamma)$ *is strictly monotonically increasing on* $\gamma \in [0,1)$*.*

Proof. Suppose $\exists\gamma \in [0,1) : \mathbf{f}^{\pi,s}_{s'}(\gamma) \notin \{0,1\}$. This implies that there exists a summand in the Neumann series $\sum_{t=0}^{\infty} (\gamma\mathbf{T}^{\pi})^t\,\mathbf{e}_s$ in whose result s' has a positive entry. If the only such summand were at $t = 0$, $\mathbf{f}^{\pi,s}_{s'}(\gamma) = 0$, which is not the case. Therefore, $t > 0$ and the summand must strictly increase with $\gamma \in [0,1)$. □

The intuition for the following result is that if $s \neq s_1$, s can achieve strictly greater discounted s-visitation frequency than s_1 can.

Lemma F.4 (Each state has a unique visit distribution)**.** *If* $s \neq s_1$ *and* $\gamma \in [0,1)$*, then* $\exists\mathbf{f}^{\pi}(\gamma) \in \mathcal{F}_{\text{nd}}(s,\gamma) \setminus \mathcal{F}(s_1,\gamma)$*.*

Proof. Given the fixed $\gamma \in [0,1)$, consider the visit distributions of s whose policies always maximize discounted s-visitation frequency, no matter which state is the initial state. By corollary F.67, at least one such maximizing visit distribution $\mathbf{f}^{\pi}(\gamma)$ is non-dominated. Let $s_1 \neq s$ and $\mathbf{f}^{\pi'} \in \mathcal{F}(s_1)$.

$$\begin{aligned}
\mathbf{f}^{\pi',s_1}_s(\gamma) &\leq \mathbf{f}^{\pi,s_1}_s(\gamma) && \text{(F.2)}\\
&= \mathbb{1}_{s_1=s} + \gamma \mathop{\mathbb{E}}_{s_2\sim T(s_1,\pi(s_1))}\left[\mathbf{f}^{\pi,s_2}_s(\gamma)\right] && \text{(F.3)}\\
&= \gamma \mathop{\mathbb{E}}_{s_2\sim T(s_1,\pi(s_1))}\left[\mathbf{f}^{\pi,s_2}_s(\gamma)\right] && \text{(F.4)}\\
&= \gamma \mathop{\mathbb{E}}_{s_2\sim T(s_1,\pi(s_1))}\left[\mathbb{1}_{s_2=s} + \gamma \mathop{\mathbb{E}}_{s_3\sim T(s_2,\pi(s_2))}\left[\mathbf{f}^{\pi,s_3}_s(\gamma)\right]\right] && \text{(F.5)}\\
&\leq \gamma\mathbf{f}^{\pi,s}_s(\gamma) && \text{(F.6)}\\
&< \mathbf{f}^{\pi,s}_s(\gamma). && \text{(F.7)}
\end{aligned}$$

Equation (F.2) follows because π maximizes s-visitation frequency from every state. Equation (F.4) follows because $s_1 \neq s$ ($\mathbb{1}_{s_1=s}$ is the indicator function which returns 1 if $s_1 = s$ and 0 otherwise). Equation (F.6) follows because, at best, $s_2 = s$ with probability

1; until s is reached, no visitation frequency is accrued. Equation (F.7) follows because $\gamma \in (0,1)$.

Since $\forall \mathbf{f}^{\pi',s_1} \in \mathcal{F}(s_1) : \mathbf{f}_s^{\pi',s_1}(\gamma) < \mathbf{f}_s^{\pi,s}(\gamma)$, $\mathbf{f}^{\pi,s} \notin \mathcal{F}(s_1)$. $\square$

F.1.1 Child distributions

Definition F.5 (Child state distributions)**.** $T(s) \coloneqq \big\{T(s,a) \mid a \in \mathcal{A}\big\}$.

Definition F.6 (Non-dominated child state distributions)**.** The *non-dominated child state distributions* are $T_{\text{nd}}(s) \coloneqq \text{ND}\big(T(s)\big)$.

Lemma F.7 (When the dynamics are locally deterministic, $T(s) = T_{\text{nd}}(s)$)**.** *Suppose s is such that $\forall a \in \mathcal{A}, s' \in \mathcal{S} : T(s,a,s') \in \{0,1\}$. Then $T(s) = T_{\text{nd}}(s)$.*

Proof. Local determinism implies that $T(s)$ is a set of one-hot state vectors $\mathbf{e}_{s^*}$. Then each $\mathbf{e}_{s^*}$ strictly maximizes visitation frequency for s^*. Therefore, $\mathbf{e}_{s^*}$ is strictly optimal for reward function $\mathbf{r} \coloneqq \mathbf{e}_{s^*}$ when $\gamma \approx 0$, and so $\mathbf{e}_{s^*} \in T_{\text{nd}}(s)$. $\square$

Lemma F.8 (Dominated child state distributions induce dominated visit distribution functions)**.** *If $T(s,a) \in T(s) \setminus T_{\text{nd}}(s)$ and $\pi(s) = a$, then $\mathbf{f}^\pi \in \mathcal{F}(s) \setminus \mathcal{F}_{\text{nd}}(s)$.*

Proof. Let $\mathbf{d} \coloneqq T(s,a) \in T(s) \setminus T_{\text{nd}}(s)$ and $\pi(s) = a$ for some policy π. If $\mathbf{f}^\pi \in \mathcal{F}_{\text{nd}}(s)$, $\mathbf{f}^\pi$ has to be strictly optimal for some R at some discount rate $\gamma \in (0,1)$. In particular, in order for $\mathbf{f}^\pi$ to be *uniquely optimal* for R at γ, only actions equivalent to a at state s can be optimal.

Let $\mathbf{v}_\gamma^\pi \in \mathbb{R}^{|\mathcal{S}|}$ be $V_R^\pi(\cdot,\gamma)$ in column vector form. Then we must have

$$\mathbf{d}^\top \mathbf{v}_\gamma^\pi > \max_{\mathbf{d}' \in T(s) \setminus \{\mathbf{d}\}} \mathbf{d}'^\top \mathbf{v}_\gamma^\pi. \tag{F.8}$$

This contradicts the fact that $T(s,a) \notin T_{\text{nd}}(s)$. $\square$

Lemma F.9 (Dynamics characterize the nested structure of visit distribution functions)**.**

$$T(s,a) \in T(s) \text{ iff } \left\{\mathbf{e}_s + \gamma\, \mathbb{E}_{s'\sim T(s,a)}\left[\mathbf{f}^{\pi,s'}\right] \mid \mathbf{f}^{\pi,s'} \in \mathcal{F}(s' \mid \pi(s)=a)\right\} \subseteq \mathcal{F}(s).$$

Proof. $T(s) = \left\{\lim_{\gamma\to 0}\gamma^{-1}\left(\mathbf{f}_{s,\gamma} - \mathbf{e}_s\right) \mid \mathbf{f}_s \in \mathcal{F}(s)\right\}$. The lemma then follows from the linear independence of the canonical unit vectors: $\sum_i \alpha_i \mathbf{e}_{s_i}$ is an element of the right-hand set iff it is an element of $T(s)$. $\square$

Definition F.10 (Normalized value and action-value functions)**.** For policy π,

$$V^{\pi}_{R,\,\text{norm}}\left(R\right) s,\gamma \coloneqq \lim_{\gamma^*\to\gamma}(1-\gamma^*)V^{\pi}_{R}\left(s,\gamma^*\right);$$

this limit exists for all $\gamma \in [0,1]$ by lemma D.45. We similarly define $V^{*}_{R,\,\text{norm}}\left(s,\gamma\right)$ and $Q^{*}_{R,\,\text{norm}}\left(s,a,\gamma\right)$.

Lemma F.11 ($\textsc{Power}_{\mathcal{D}_{\text{bound}}}$ is the average normalized next-state optimal value)**.**

$$\textsc{Power}_{\mathcal{D}_{bound}}\left(s,\gamma\right) = \mathop{\mathbb{E}}_{R\sim\mathcal{D}}\left[\max_{\pi\in\Pi}\mathop{\mathbb{E}}_{s'\sim T\left(s,\pi(s)\right)}\left[V^{\pi}_{R,\,norm}\left(s',\gamma\right)\right]\right]. \tag{F.9}$$

Proof. The $\gamma \in (0,1)$ case follows from eq. (D.94) in the proof of lemma D.43 and the fact that when $\gamma \in (0,1)$, $V^{\pi}_{R,\,\text{norm}}\left(s',\gamma\right) \coloneqq \lim_{\gamma^*\to\gamma}(1-\gamma^*)V^{\pi}_{R}\left(s',\gamma^*\right) = (1-\gamma)V^{\pi}_{R}\left(s',\gamma\right)$ because on-policy value is continuous on $\gamma \in [0,1)$. The $\gamma = 0$ and $\gamma = 1$ cases follow by taking the appropriate limits and applying the continuity of $\textsc{Power}_{\mathcal{D}_{\text{bound}}}$. $\square$

Proposition F.12 (Identical $T_{\text{nd}}\left(s\right)$ implies equal $\textsc{Power}_{\mathcal{D}_{\text{bound}}}(s,\gamma)$)**.** *If* $T_{\text{nd}}\left(s\right) = T_{\text{nd}}\left(s'\right)$*, then* $\forall\gamma \in [0,1] : \textsc{Power}_{\mathcal{D}_{bound}}\left(s,\gamma\right) = \textsc{Power}_{\mathcal{D}_{bound}}\left(s',\gamma\right)$.

Proof. Let $\gamma \in (0,1)$. Lemma F.11 shows that

$$\textsc{Power}_{\mathcal{D}_{\text{bound}}}\left(s,\gamma\right) = (1-\gamma)\mathop{\mathbb{E}}_{R\sim\mathcal{D}}\left[\max_{\pi\in\Pi}\mathop{\mathbb{E}}_{s'\sim T\left(s,\pi(s)\right)}\left[V^{\pi}_{R}\left(s',\gamma\right)\right]\right] \tag{F.10}$$

$$= (1-\gamma) \mathop{\mathbb{E}}_{R\sim\mathcal{D}}\left[\max_a \mathop{\mathbb{E}}_{s'\sim T(s,a)}\left[V^*_R\left(s',\gamma\right)\right]\right] \tag{F.11}$$

$$= (1-\gamma) \mathop{\mathbb{E}}_{R\sim\mathcal{D}}\left[\max_{\mathbf{d}\in T(s)} \mathop{\mathbb{E}}_{s'\sim\mathbf{d}}\left[V^*_R\left(s',\gamma\right)\right]\right] \tag{F.12}$$

$$= (1-\gamma) \mathop{\mathbb{E}}_{R\sim\mathcal{D}}\left[\max_{\mathbf{d}\in T_{\text{nd}}(s)} \mathop{\mathbb{E}}_{s'\sim\mathbf{d}}\left[V^*_R\left(s',\gamma\right)\right]\right] \tag{F.13}$$

$$= (1-\gamma) \mathop{\mathbb{E}}_{R\sim\mathcal{D}}\left[\max_{\mathbf{d}'\in T_{\text{nd}}(s')} \mathop{\mathbb{E}}_{s'\sim\mathbf{d}'}\left[V^*_R\left(s',\gamma\right)\right]\right] \tag{F.14}$$

$$= \textsc{Power}_{\mathcal{D}_{\text{bound}}}\left(s',\gamma\right). \tag{F.15}$$

Equation (F.10) follows from lemma F.11. Since restriction to stationary policies leaves optimal value unchanged, the non-stationarity in eq. (F.11) leaves optimal value unchanged.

Lemma F.8 shows that dominated child distributions induce dominated visit distribution functions. By lemma D.39, restriction to non-dominated visit distribution functions leaves optimal value unchanged for all reward functions, and so only considering $\mathbf{d} \in T_{\text{nd}}(s)$ leaves optimal value unchanged. Therefore, eq. (F.13) follows. Equation (F.14) follows because we assumed that $T_{\text{nd}}(s) = T_{\text{nd}}\left(s'\right)$. Equation (F.15) follows by the reasoning for eq. (F.10) through eq. (F.14).

The equality holds in the limit as $\gamma \to 0$ or $\gamma \to 1$. □

Figure F.1 demonstrates how proposition F.12 establishes $\textsc{Power}_{\mathcal{D}_{\text{bound}}}$ equality, even when the equality is not intuitively obvious.

F.1.2 Optimal policy set transfer across discount rates

Definition F.13 (Blackwell optimal policies [11])**.** $\Pi^*(R,1) := \lim_{\gamma\to 1}\Pi^*(R,\gamma)$ is the *Blackwell optimal policy set* for reward function R.

Definition F.14 (Greedy optimality)**.** $\Pi^{\text{greedy}}(R)$ is the set of policies π for which $\forall s$:

$$\mathop{\mathbb{E}}_{s'\sim T(s,\pi(s))}\left[R(s')\right] = \max_a \mathop{\mathbb{E}}_{s'\sim T(s,a)}\left[R(s')\right].$$

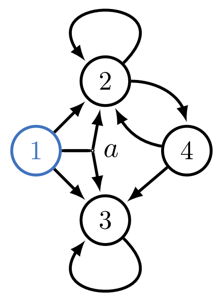


Figure F.1: The bifurcated action a is a stochastic transition, where $T(s_1, a, s_2) = .5 = T(s_1, a, s_3)$. Since $T(s_1, a) \in T(s_1) \setminus T_{\text{nd}}(s_1)$, $T_{\text{nd}}(s_1) = T_{\text{nd}}(s_4)$ and so proposition F.12 implies that $\forall \gamma \in [0,1] : \textsc{Power}_{\mathcal{D}_{\text{bound}}}(s_1, \gamma) = \textsc{Power}_{\mathcal{D}_{\text{bound}}}(s_4, \gamma)$.

Figure F.2 shows that Turner et al. [99]'s proposition D.35 does not always hold for $\gamma = 0$, but proposition F.18 shows that it almost always does hold.

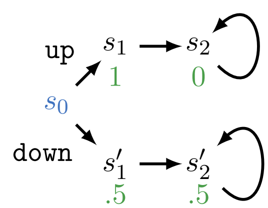


Figure F.2: When $\gamma = \frac{1}{2}$, going `up` and going `down` are both optimal, with $V^*_R\left(s_1, \frac{1}{2}\right) = V^*_R\left(s'_1, \frac{1}{2}\right) = 1$. However, proposition D.35 does not hold when the new discount rate is 0: only $\pi^{\texttt{down}} \in \Pi^*\left(V^*_R\left(\cdot, \frac{1}{2}\right), 0\right)$, with greedy policies preferring to "gradually" navigate through higher-value states. When the new discount rate $\gamma = 0$, the optimal policies for the constructed reward function are guaranteed to also be optimal for the original by proposition F.18, but the converse statement is not true in this MDP.

Lemma F.15 (When $\gamma \approx 0$, optimal policies are greedy)**.** $\Pi^*(R, 0) \subseteq \Pi^{greedy}(R)$.

Proof. Let $\pi^* \in \Pi^*(R, 0)$, $\pi \in \Pi$, and $\gamma \in (0, 1)$.

$$V^{\pi^*}_R(s, \gamma) \geq V^{\pi}_R(s, \gamma) \tag{F.16}$$

$$R(s)+\gamma \underset{s'\sim T\left(s,\pi^*(s)\right)}{\mathbb{E}}\left[V_R^{\pi^*}(s',\gamma)\right] \geq R(s)+\gamma \underset{s'\sim T\left(s,\pi(s)\right)}{\mathbb{E}}\left[V_R^{\pi}(s',\gamma)\right] \tag{F.17}$$

$$\underset{s'\sim T\left(s,\pi^*(s)\right)}{\mathbb{E}}\left[V_R^{\pi^*}(s',\gamma)\right] \geq \underset{s'\sim T\left(s,\pi(s)\right)}{\mathbb{E}}\left[V_R^{\pi}(s',\gamma)\right] \tag{F.18}$$

$$\underset{s'\sim T\left(s,\pi^*(s)\right)}{\mathbb{E}}\left[R(s')+\gamma \underset{s''}{\mathbb{E}}\left[V_R^{\pi^*}(s'',\gamma)\right]\right] \geq \underset{s'\sim T\left(s,\pi(s)\right)}{\mathbb{E}}\left[R(s')+\gamma \underset{s''}{\mathbb{E}}\left[V_R^{\pi}(s'',\gamma)\right]\right]. \tag{F.19}$$

Equation (F.18) is valid because $\gamma > 0$.

Let $b \coloneqq \min_{s^-\in\mathcal{S}} R(s^-), c \coloneqq \max_{s^+\in\mathcal{S}} R(s^+)$; these exist because $\mathcal{S}$ is finite. Then $\forall s'' : \frac{b}{1-\gamma} \leq V_R^*\left(s'',\gamma\right) \leq \frac{c}{1-\gamma}$. Because optimal value is thus bounded, eq. (F.19) is controlled by expected next-state reward when $\gamma \approx 0$. The result follows because definition 5.7 defines $\Pi^*\left(R,0\right) \coloneqq \lim_{\gamma\to 0} \Pi^*\left(R,\gamma\right)$. ☐

Corollary F.16 (At each $\gamma \in (0,1)$, almost all reward functions have optimal actions at each state which are unique up to action equivalence)**.**

Proof. Let s be a state. Lemma F.105 implies that, for any fixed $\gamma \in (0,1)$ and for almost all reward functions R, the optimal action at s is unique up to action equivalence (definition 5.23). Since there are only finitely many states, almost all reward functions R must have a unique-up-to-equivalence optimal action at all states. ☐

It is optimal to stochastically mix between two actions iff both actions are (deterministically) optimal. Corollary F.17 shows that almost no reward functions have optimal policies which mix between non-equivalent actions.

Corollary F.17 (Almost all reward functions do not have non-trivial stochastic optimal policies)**.** *Let $\gamma \in (0,1)$ and let $X \coloneqq \left\{R \in \mathbb{R}^{\mathcal{S}} \mid \exists s, \pi_1^*, \pi_2^* \in \Pi^*\left(R,\gamma\right) : \pi_1^*(s) \not\equiv_s \pi_2^*(s)\right\}$. Considered as a subset of $\mathbb{R}^{|\mathcal{S}|}$, X has zero Lebesgue measure.*

Proof. The result follows directly from corollary F.16. ☐

Proposition F.18 (Transferring optimal policy sets to $\gamma = 0$)**.** *Fix $\gamma \in (0,1)$ and let R be a reward function.* $\Pi^*\left(V_R^*(\cdot,\gamma),0\right) \subseteq \Pi^{greedy}\left(V_R^*(\cdot,\gamma)\right) = \Pi^*\left(R,\gamma\right)$. *Equality holds* *for almost all R; in particular, equality holds when* $\forall s \in \mathcal{S} : \exists \mathbf{f} \in \mathcal{F}_{\text{nd}}(s) : \mathbf{f}(\gamma)^\top \mathbf{r} > \max_{\mathbf{f}' \in \mathcal{F}_{\text{nd}}(s) \setminus \{\mathbf{f}\}} \mathbf{f}'(\gamma)^\top \mathbf{r}$.

Proof. By lemma F.15, $\Pi^*\left(V_R^*(\cdot, \gamma), 0\right) \subseteq \Pi^{\text{greedy}}\left(V_R^*(\cdot, \gamma)\right)$. By the definition of an optimal policy, $\pi \in \Pi^*(R, \gamma)$ iff π maximizes value $V_R^\pi(s, \gamma)$ at all states s, which holds iff π maximizes the rightmost term of $V_R^*(s, \gamma) := R(s) + \gamma \max_a \mathbb{E}_{s' \sim T(s,a)}\left[V_R^*(s', \gamma)\right]$ since $\gamma \in (0, 1)$. This is true iff $\pi \in \Pi^{\text{greedy}}\left(V_R^*(\cdot, \gamma)\right)$. Therefore, $\Pi^{\text{greedy}}\left(V_R^*(\cdot, \gamma)\right) = \Pi^*(R, \gamma)$.

Corollary F.16 implies that, for any fixed $\gamma \in (0, 1)$ and for almost all reward functions R, the optimal action at each state is unique up to action equivalence (definition 5.23). Therefore, $\forall s : \exists \mathbf{f} \in \mathcal{F}_{\text{nd}}(s) : \mathbf{f}(\gamma)^\top \mathbf{r} > \max_{\mathbf{f}' \in \mathcal{F}_{\text{nd}}(s) \setminus \{\mathbf{f}\}} \mathbf{f}'(\gamma)^\top \mathbf{r}$ for these reward functions R.

For such reward functions R, a policy $\pi' \in \Pi^*(R, \gamma)$ iff it induces the appropriate strictly optimal $\mathbf{f} \in \mathcal{F}_{\text{nd}}(s)$ at each state s. Let $\pi \in \Pi^*\left(V_R^*(\cdot, \gamma), 0\right)$. Since $\pi \in \Pi^*(R, \gamma)$ by the above and since optimal action for R at γ is unique up to action equivalence, π and π' must choose equivalent actions at all states. Then $\pi' \in \Pi^*(R, \gamma)$. Since π' was arbitrary, $\Pi^*\left(V_R^*(\cdot, \gamma), 0\right) = \Pi^*(R, \gamma)$. □

Question F.19 (Can proposition D.35 be generalized to $\gamma = 1$?)**.**

F.1.3 Optimal policy set characterization

Definition F.20 (Set of optimal policy sets)**.** The set of deterministic stationary optimal policy sets is

$$\mathcal{P}^{\text{opt}} := \left\{\Pi^*(R, \gamma) \mid R \in \mathbb{R}^{\mathcal{S}}, \gamma \in [0, 1]\right\}. \tag{F.20}$$

Theorem F.21 (Characterization of $\mathcal{P}^{\text{opt}}$)**.** *Let* $\Pi' \subseteq \Pi$ *be a set of deterministic stationary policies. The following are equivalent:*

1. $\forall \gamma_1 \in [0, 1] : \exists R_1 \in \mathbb{R}^{\mathcal{S}} : \Pi^*(R_1, \gamma_1) = \Pi'$.

2. $\exists \gamma_2 \in [0, 1], R_2 \in \mathbb{R}^{\mathcal{S}} : \Pi^*(R_2, \gamma_2) = \Pi'$ *(*i.e. $\Pi' \in \mathcal{P}^{opt}$*).*

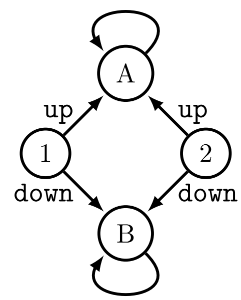


Figure F.3: In general, not all policy sets are valid optimal policy sets ($\mathcal{P}^{\text{opt}} \subsetneq \mathcal{P}(\Pi)$). In the above, consider the policy $\pi(s_1) := \texttt{up}, \pi(s_2) := \texttt{down}$. Then $\Pi' := \{\pi\} \notin \mathcal{P}^{\text{opt}}$. If $R(s_A) > R(s_B)$, then no $\pi \in \Pi'$ would choose up over down; similar reasoning holds for $R(s_A) < R(s_B)$. If $R(s_A) = R(s_B)$, then Π' would contain some $\pi_2(s_1) = \texttt{down}$.

3. $\exists R_3 \in \mathbb{R}^{\mathcal{S}} : \Pi^{greedy}(R_3) = \Pi'$.

Item 1 and item 2 say that the set of feasible optimal policy sets is invariant to the discount rate. In particular, any optimal policy set can be rationalized as asymptotically greedy or Blackwell-optimal.

Item 2 and item 3 show that a policy set is an optimal policy set for some R_2 at some γ_2 iff that policy set can be rationalized as greedily optimizing some R_3.

Proof. 1 $\implies$ 2 by definition F.20.

2 $\implies$ 3. Suppose item 2 holds. If $\gamma_2 \in (0,1)$, define $R_3(s) := V^*_{R_2}(s, \gamma_2)$. Then apply proposition F.18 to conclude that $\Pi^{\text{greedy}}(R_3) = \Pi^*(R_2, \gamma_2) = \Pi'$.

Since $\Pi^*(R_2, 1) := \lim_{\gamma \to 1} \Pi^*(R_2, \gamma)$ (definition 5.7) always exists by corollary F.124 and only finitely many optimal policy shifts occur (lemma D.40), $\forall R : \exists \gamma \in (0,1) : \Pi^*(R, \gamma) = \Pi^*(R, 1)$. Select such a γ for R_2. Since $\gamma \in (0,1)$, we have just shown that $\exists R_3 : \Pi' = \Pi^*(R_2, 1) = \Pi^*(R_2, \gamma) = \Pi^{\text{greedy}}(R_3)$.

The $\gamma_2 = 0$ case follows by the same reasoning. Then item 2 implies item 3.

3 $\implies$ 1. Suppose item 3 holds. Let $\pi \in \Pi^{\text{greedy}}(R_3)$ and $\gamma_1 \in (0,1)$. Define $R_1 := (\mathbf{I} - \gamma_1 \mathbf{T}^\pi) R_3$. By reasoning identical to that in the proof of proposition D.35, $\exists R_1 \in \mathbb{R}^{\mathcal{S}} : \Pi^*(R_1, \gamma_1) = \Pi^{\text{greedy}}(R_3) = \Pi'$.

Now suppose $\gamma_1 = 1$ and consider the function $R_1(s, \gamma_1) \coloneqq (\mathbf{I} - \gamma_1 \mathbf{T}^\pi)\, R_3(s)$. Since $\forall \gamma \in (0,1) : \Pi^*\left(R_1(\cdot, \gamma), \gamma\right) = \Pi'$ by the above reasoning, $\Pi^*\left(R_1, 1\right) = \lim_{\gamma \to 1} \Pi^*\left(R_1(\cdot, \gamma), \gamma\right) = \Pi'$ (by definition 5.7).

Similar reasoning holds when $\gamma_1 = 0$. Since this equality can be satisfied for any $\gamma_1 \in [0,1]$, item 3 implies item 1. □

Recall the following definition:

Definition F.5 (Child state distributions)**.** $T(s) \coloneqq \left\{T(s,a) \mid a \in \mathcal{A}\right\}$.

Given a starting state s, we can consider its available child state distributions $T(s)$ to be lotteries over the next state the agent will visit. A set of policies Π' can be seen as preferring some state lotteries (by containing policies which take a certain action at s) over others (lotteries which are not induced by any $\pi \in \Pi'$ at state s).

Definition F.22 (State lottery relation implied by a policy set)**.** The relations implied by policy set $\Pi' \subseteq \Pi$ is defined as follows. Consider s such that $L_1, L_2 \in T(s)$ (*i.e.* the lotteries are realizable at state s), where $\exists \pi_1 \in \Pi' : T(s, \pi_1(s)) = L_1$.

$L_1 \sim^{\Pi'} L_2$ if $\exists \pi_2 \in \Pi' : T(s, \pi_2(s)) = L_2$. In other words, L_1 and L_2 are both induced by $\pi_1, \pi_2 \in \Pi'$, respectively. This implies that L_1 and L_2 are "equally good," since Π' induces both of them.

$L_1 \succ^{\Pi'} L_2$ if $\neg\exists \pi_2 \in \Pi' : T(s, \pi_2(s)) = L_2$. Since no policy induces L_2, L_2 is considered to be strictly worse than L_1.

However, some Π' imply incoherent preferences over states: in fig. F.3, Π' implies that $s_A \succ^{\Pi'} s_B$ (via s_1) and also $s_B \succ^{\Pi'} s_A$ (via s_2's).

Definition F.23 (Coherent state lottery relations)**.** $(\sim^{\Pi'}, \succ^{\Pi'})$ is *coherent* when $\exists \mathbf{r} \in \mathbb{R}^{|\mathcal{S}|}$:

1. $L_1 \sim^{\Pi'} L_2$ implies that $L_1^\top \mathbf{r} = L_2^\top \mathbf{r}$.
2. $L_1 \succ^{\Pi'} L_2$ implies that $L_1^\top \mathbf{r} > L_2^\top \mathbf{r}$.

Theorem F.24 (Optimal policy sets imply coherent preference relations)**.** *If* $\Pi' \in \mathcal{P}^{opt}$, *then* $(\sim^{\Pi'}, \succ^{\Pi'})$ *is coherent.*

Proof. If $\Pi' \in \mathcal{P}^{\text{opt}}$, then let R be a reward function for which $\Pi^{\text{greedy}}(R) = \Pi'$ (such a reward function exists by theorem F.21). By the definition of $\Pi^{\text{greedy}}(R)$ (definition F.14), $\pi \in \Pi^{\text{greedy}}(R)$ iff $\forall s \in \mathcal{S} : \pi(s)$ maximizes expected next-step reward.

Suppose $L_1 \sim^{\Pi'} L_2$. Then by definition F.22, $\exists \pi_1, \pi_2 \in \Pi' : \big(T(s, \pi_1(s)) = L_1\big) \wedge \big(T(s, \pi_2(s)) = L_2\big)$. By the greediness of Π', this implies that both L_1 and L_2 maximize next-step reward. Therefore, $T(s, \pi_1(s))^\top \mathbf{r} = T(s, \pi_2(s))^\top \mathbf{r}$ and so $L_1^\top \mathbf{r} = L_2^\top \mathbf{r}$.

Suppose $L_1 \succ^{\Pi'} L_2$. Then by definition F.22, there $\exists s \in \mathcal{S}, \pi_1 \in \Pi' : T(s, \pi_1(s)) = L_1$ and $L_2 \in T(s)$, but $\neg \exists \pi_2 \in \Pi' : T(s, \pi_2(s)) = L_2$. Since there is no $\pi_2 \in \Pi^{\text{greedy}}(R)$ taking some action a_2 such that $T(s, a_2) = L_2$, and since such an action exists ($L_2 \in T(s)$), we conclude that a' is not a greedy action at state s for R. In other words, $T(s, \pi_1(s))^\top \mathbf{r} > T(s, a')^\top \mathbf{r}$ and so $L_1^\top \mathbf{r} > L_2^\top \mathbf{r}$.

Since $L_1 \sim^{\Pi'} L_2$ and $L_1 \succ^{\Pi'} L_2$ were arbitrary, $(\sim^{\Pi'}, \succ^{\Pi'})$ is coherent for R. □

The VNM utility theorem implies that agents with a coherent preference ordering over state lotteries can be rationalized as maximizing the expected utility of some utility function. Somewhat similarly, theorem F.24 shows that optimal policy sets can be rationalized as coherently trading off the greedy values of environment states.

For incoherent Π', it is not generally possible to rectify the incoherence of $(\sim^{\Pi'}, \succ^{\Pi'})$ by deducing a reward function for which Π' "should" be greedy. For example, in fig. F.3, it is unclear what preferences over state lotteries Π' "should" have.

Conjecture F.25 (Coherent relations imply greedy policy subset)**.** If $(\sim^{\Pi'}, \succ^{\Pi'})$ is coherent for R, then $\Pi' \subseteq \Pi^{\text{greedy}}(R)$.

F.1.3.1 Optimal policy sets factorize

We now further constrain the structure of optimal policy sets.

Definition F.26 (Actions taken by a policy set at a state)**.** Let $\Pi' \subseteq \Pi$.

$$\mathcal{A}_s^{\Pi'} \coloneqq \left\{a \in \mathcal{A} \mid \exists \pi \in \Pi' : \pi(s) = a\right\}. \tag{F.21}$$

Lemma F.27 (Optimal policy sets mix-and-match optimal actions)**.** *Let* $\pi_1, \pi_2 \in \Pi' \in \mathcal{P}^{opt}$ *and consider state-space partition* $\mathcal{S} = S_1 \sqcup S_2$. *Then* $\exists \pi_3 \in \Pi'$ *such that* $\forall s_1 \in S_1 : \pi_3(s_1) = \pi_1(s_1)$ *and* $\forall s_2 \in S_2 : \pi_3(s_2) = \pi_2(s_2)$.

Proof. Since $\Pi' \in \mathcal{P}^{\text{opt}}$, $\exists R \in \mathbb{R}^{\mathcal{S}} : \Pi' = \Pi^{\text{greedy}}(R)$ by theorem F.21. By the definition of $\Pi^{\text{greedy}}(R)$ (definition F.14), each action of π_1 and π_2 maximizes expected next-state reward. Therefore, any π_3 agreeing with π_1 on S_1 and with π_2 on S_2 must also maximize expected next-state reward for all states in the MDP. Then $\pi_3 \in \Pi^{\text{greedy}}(R) = \Pi'$. $\square$

Proposition F.28 (If $\Pi' \in \mathcal{P}^{\text{opt}}$, then $\Pi' \cong \prod_s \mathcal{A}_s^{\Pi'}$)**.** *There exists a set bijection between* Π' *and* $\prod_s \mathcal{A}_s^{\Pi'}$.

Proof. Apply lemma F.27 to conclude that any combination of actions in $\prod_s \mathcal{A}_s^{\Pi'}$ maps to an optimal policy in Π'. By inverting the same map, any policy in Π' must map to a unique element of $\prod_s \mathcal{A}_s^{\Pi'}$ by the definition of $\mathcal{A}_s^{\Pi'}$ (definition F.26). $\square$

This implies that optimal policy sets admit compressed representations: instead of explicitly storing the outputs of (up to) $|\mathcal{A}|^{|\mathcal{S}|}$ optimal policies at $|\mathcal{S}|$ states, we can simply record (up to) $|\mathcal{A}|$ optimal actions at $|\mathcal{S}|$ states. Then, we regenerate the full set Π' by taking the Cartesian product of the optimal actions.

This immediately suggests divisibility tests which rule out certain policy sets from being optimal policy sets.

Corollary F.29 (If $\Pi' \in \mathcal{P}^{\text{opt}}$, then $\forall s : \left|\mathcal{A}_s^{\Pi'}\right|$ divides $\left|\Pi'\right|$)**.**

Lemma F.30 (Optimal policy sets take all equivalent actions)**.** *Let* $\Pi' \in \mathcal{P}^{opt}$. *For any state* s *and action* a, *if* $a \in \mathcal{A}_s^{\Pi'}$ *and* $a' \equiv_s a$, *then* $a' \in \mathcal{A}_s^{\Pi'}$.

Proof. Since $\Pi' \in \mathcal{P}^{\text{opt}}$, $\exists R \in \mathbb{R}^{\mathcal{S}} : \Pi' = \Pi^{\text{greedy}}(R)$ by theorem F.21. If $a \in \mathcal{A}_s^{\Pi'}$ and $a' \equiv_s a$, a is greedily optimal at s. Then $a' \equiv_s a$ must also be greedily optimal because $T(s,a) = T(s,a')$ by definition 5.23. Therefore, $a' \in \mathcal{A}_s^{\Pi'}$. $\square$

Proposition F.31 (Multiple optimal actions at multiple states implies that the optimal policy set has composite cardinality)**.** *Let* $\Pi' \in \mathcal{P}^{opt}$*. If* $\exists s \neq s' : \left|\mathcal{A}_s^{\Pi'}\right| \cdot \left|\mathcal{A}_{s'}^{\Pi'}\right|$*, then* $\left|\Pi'\right|$ *is composite.*

Proof. By corollary F.29, $\left|\mathcal{A}_s^{\Pi'}\right| \cdot \left|\mathcal{A}_{s'}^{\Pi'}\right|$ divides $\left|\Pi'\right|$. Since both of these factors are greater than 1, $\left|\Pi'\right|$ is a composite number. □

Corollary F.32 (If $\forall s \in \mathcal{S}, a \in \mathcal{A} : \exists a' \neq a : a \equiv_s a'$ and if $|\mathcal{S}| > 1$, then no optimal policy sets have prime cardinality)**.**

Proof. Let $\Pi' \in \mathcal{P}^{\text{opt}}$. Let $s \neq s'$ be distinct states. Π' is nonempty by the existence of a stationary deterministic optimal policy for all reward functions and discount rates. Let $\pi \in \Pi'$. Since $\forall s'' \in \mathcal{S}, a \in \mathcal{A} : \exists a' \neq a : a \equiv_{s''} a'$, apply lemma F.30 to conclude that $\left|\mathcal{A}_s^{\Pi'}\right|, \left|\mathcal{A}_{s'}^{\Pi'}\right| > 1$. Apply proposition F.31. □

F.1.3.2 Deciding whether a policy set is optimal

Theorem F.24 allows us to efficiently decide whether a policy set is optimal for some (R, γ) tuple (since enumeration is impossible for the uncountably many such tuples). I initially conjectured conjecture F.33 in 2020. Abel et al. [1]'s Theorem 4.3 has since proven a generalization of it.

Conjecture F.33 ($\Pi' \in \mathcal{P}^{\text{opt}}$? can be efficiently decided)**.** The decision problem corresponding to deciding whether $\Pi' \in \mathcal{P}^{\text{opt}}$ is in **P**.

F.1.4 How reward function combination affects optimality

Lemma F.34 (Non-negative combination of reward functions preserves value ordering agreement)**.** *If* $V_{R_1}^{\pi}(s, \gamma) \geq V_{R_1}^{\pi'}(s, \gamma)$ *and* $V_{R_2}^{\pi}(s, \gamma) \geq V_{R_2}^{\pi'}(s, \gamma)$*, then for any* $\alpha, \beta \geq 0$*,*

$$V_{\alpha R_1 + \beta R_2}^{\pi}(s, \gamma) \geq V_{\alpha R_1 + \beta R_2}^{\pi'}(s, \gamma). \tag{F.22}$$

Proof. The premise implies that $\mathbf{f}^{\pi,s}(\gamma)^\top \mathbf{r}_1 \geq \mathbf{f}^{\pi',s}(\gamma)^\top \mathbf{r}_1$ and $\mathbf{f}^{\pi,s}(\gamma)^\top \mathbf{r}_2 \geq \mathbf{f}^{\pi',s}(\gamma)^\top \mathbf{r}_2$.

Then

$$V^{\pi}_{\alpha R_1+\beta R_2}(s,\gamma) = \mathbf{f}^{\pi,s}(\gamma)^\top(\alpha\mathbf{r}_1+\beta\mathbf{r}_2) \tag{F.23}$$
$$\geq \mathbf{f}^{\pi',s}(\gamma)^\top(\alpha\mathbf{r}_1+\beta\mathbf{r}_2) \tag{F.24}$$
$$= V^{\pi'}_{\alpha R_1+\beta R_2}(s,\gamma). \tag{F.25}$$

□

Corollary F.35 (Non-negative combination of reward functions preserves optimal policy agreement)**.** *If π^* is optimal for R_1 and R_2 at discount rate γ, it is also optimal for $\alpha R_1+\beta R_2$ at discount rate γ for any $\alpha,\beta\geq 0$.*

However, R_1 and R_2 having the same optimal policy set doesn't mean they will incentivize the same distribution over learned policies.

Question F.36 (If two reward functions induce the same ordering over policies, when do they tend to incentivize similar learned policies?)**.**

Remark. With respect to a distribution over network initializations and a fixed learning process, the distribution over learned policies is sometimes invariant to positive affine transformation of the reward function (so long as no *e.g.* underflow or overflow errors occur, or instability does not occur). In contrast, reward shaping [58] often accelerates learning and changes which policies get tend to get learned after a fixed number of policy improvement steps. Both transformations preserve policy ordering, but only shaping improves learning. What is the fundamental difference?

F.1.5 Visit distribution function agreement

Figure F.4 illustrates how two distinct visit distribution functions can output the same visit distribution for certain γ.

Lemma F.37 (Distinct visit distribution functions agree finitely many times)**.** *Let s,s' be any two states and let $\mathbf{f}\in\mathcal{F}(s),\mathbf{f}'\in\mathcal{F}(s')$ be distinct visit distribution functions. For all but finitely many $\gamma\in(0,1)$, $\mathbf{f}(\gamma)\neq\mathbf{f}'(\gamma)$.*

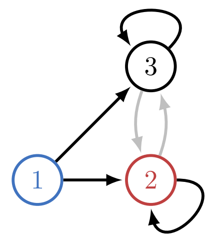


Figure F.4: When $\gamma = \frac{1}{2}$, the trajectories $s_1 s_2 s_3 s_3 \ldots$ and $s_1 s_3 s_2 s_2 \ldots$ induce the same state visitation distribution of $\begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{2} \end{pmatrix}^\top$. Both trajectories are induced by dominated state visit distribution functions; lemma D.52 shows that this is no coincidence.

Proof. By lemma D.9, each visit distribution function is multivariate rational on γ. If the functions are distinct, then they must disagree on at least one output dimension. Their difference along this dimension is a rational non-zero function, which has finitely many roots by the fundamental theorem of algebra. $\square$

Lemma F.38 (Different states have disjoint visit distribution function sets)**.** *If* $s \neq s'$, $\neg\exists \mathbf{f} \in \mathcal{F}(s), \mathbf{f}' \in \mathcal{F}(s') : \forall \gamma \in (0,1) : \mathbf{f}(\gamma) = \mathbf{f}(\gamma)$.

Proof. When $\gamma \approx 0$, the visit distributions approximate the appropriate unit vectors: $\mathbf{f}(\gamma) \approx \mathbf{e}_s$ and $\mathbf{f}'(\gamma) \approx \mathbf{e}_{s'}$. Since $s \neq s'$, $\mathbf{f}(\gamma) \neq \mathbf{f}'(\gamma)$ when $\gamma \approx 0$. Therefore, $\mathbf{f}$ and $\mathbf{f}'$ are distinct visit distribution functions. Apply lemma F.37. $\square$

Corollary F.39 (Visit distributions are distinct at all but finitely many γ)**.** $\forall s, s' \in \mathcal{S}, \mathbf{f} \in \mathcal{F}(s), \mathbf{f}' \in \mathcal{F}(s')$: *either* $\mathbf{f} = \mathbf{f}'$ *or* $\mathbf{f}(\gamma) \neq \mathbf{f}'(\gamma)$ *holds for all but finitely many* $\gamma \in (0,1)$.

Proof. By lemma F.37, the distinct visit distributions at s agree with each other for at most finitely many γ. Likewise, the visit distributions of other states s' agree with $\mathbf{f}$ for at most finitely many γ. Since there are only finitely many visit distribution pairs, there is a finite set of the γ at which any two distinct visit distributions agree. $\square$

Proposition F.40 (Cross-state linear independence of visit distribution functions)**.** *$\mathbf{f} \in \mathcal{F}(s)$ cannot be written as a linear combination of the visit distribution functions of other states.*

Proof. Suppose $\mathbf{f} = \sum_i \alpha_i \mathbf{f}_{s_i}$ such that $\forall i : \alpha_i \in \mathbb{R} \wedge s_i \neq s \wedge \mathbf{f}_{s_i} \in \mathcal{F}(s_i)$. Then

$$\lim_{\gamma \to 0} \sum_i \alpha_i \mathbf{f}_{s_i} = \sum_i \alpha_i \mathbf{e}_{s_i} \tag{F.26}$$

$$\neq \mathbf{e}_s \tag{F.27}$$

$$= \lim_{\gamma \to 0} \mathbf{f}(\gamma), \tag{F.28}$$

where eq. (F.27) follows because $s_i \neq s$ and the canonical basis vectors are linearly independent. □

Figure F.5 shows that proposition F.40 does not hold amongst the visit distribution functions of a single state.

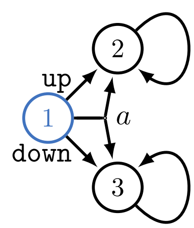


Figure F.5: The bifurcated action a is a stochastic transition, where $T(s_1, a, s_2) = .5 = T(s_1, a, s_3)$. $\mathbf{f}^{\pi_a, s_1} = .5(\mathbf{f}^{\pi_{\texttt{up}}, s_1} + \mathbf{f}^{\pi_{\texttt{down}}, s_1})$.

Lemma D.52 is generalized by proposition F.41.

Proposition F.41 (At every $\gamma \in (0,1)$, non-dominated visit distributions are outside of the convex hull of any set of other visit distributions)**.** *Let $\mathbf{f} \in \mathcal{F}(s)$. If there exists $\gamma \in (0,1)$ and $\mathbf{f}_1, \ldots, \mathbf{f}_k \in \mathcal{F}(s) \setminus \{\mathbf{f}\}$ such that $\mathbf{f}$ can be expressed as the convex combination $\mathbf{f}(\gamma) = \sum_{i=1}^{k} \alpha_i \mathbf{f}_i(\gamma)$, then $\mathbf{f} \notin \mathcal{F}_{\mathrm{nd}}(s)$.*

Proof. If some $\alpha_i = 1$, then $\mathbf{f} \notin \mathcal{F}_{\text{nd}}(s)$ by lemma D.52. Suppose at least two $\alpha_i > 0$. Then the conclusion follows by proposition F.59 (letting $X \coloneqq \left\{\mathbf{f}(\gamma) \mid \mathbf{f} \in \mathcal{F}(s)\right\}, \mathbf{x} \coloneqq \mathbf{f}(\gamma)$). $\square$

Lemma F.42 (Geometry of optimality support)**.** *At any* $\gamma \in (0,1)$, supp $\left(\mathbf{f}(\gamma) \geq \mathcal{F}(s,\gamma)\right)$ *(the set of reward functions for which* $\mathbf{f} \in \mathcal{F}(s)$ *is optimal) is both a closed convex polytope and a pointed convex cone.*

Proof. supp $\left(\mathbf{f}(\gamma) \geq \mathcal{F}(s,\gamma)\right)$ is a closed convex polytope because it is the intersection of half-spaces: $\mathbf{f}(\gamma)^\top \mathbf{r} \geq \max_{\mathbf{f}' \in \mathcal{F}(s)\setminus\{\mathbf{f}\}} \mathbf{f}'(\gamma)^\top \mathbf{r}$. The set is a pointed cone because for any $\alpha \geq 0$, $\mathbf{f}(\gamma)^\top \mathbf{r} \geq \max_{\mathbf{f}' \in \mathcal{F}(s)\setminus\{\mathbf{f}\}} \mathbf{f}'(\gamma)^\top \mathbf{r}$ implies $\mathbf{f}(\gamma)^\top (\alpha\mathbf{r}) \geq \max_{\mathbf{f}' \in \mathcal{F}(s)\setminus\{\mathbf{f}\}} \mathbf{f}'(\gamma)^\top (\alpha\mathbf{r})$. $\square$

Proposition F.43 (Visit dist. function convex hull intersection at any γ implies shared optimality status)**.** *Let* $\mathbf{f} \in \mathcal{F}(s)$*. Suppose there exists* $\gamma \in (0,1)$ *and* $\mathbf{f}_1, \ldots, \mathbf{f}_k \in \mathcal{F}(s)\setminus\{\mathbf{f}\}$ *such that* $\mathbf{f}$ *can be expressed as the convex combination* $\mathbf{f}(\gamma) = \sum_{i=1}^k \alpha_i \mathbf{f}_i(\gamma)$ *(where each* $\alpha_i > 0$*).*

$\mathbf{f}$ *is optimal for reward function* R *at* $\gamma^* \in (0,1)$ *iff* $\forall i : \mathbf{f}_i$ *is optimal for* R *at* γ^**.*

Proof. If $\forall i : \mathbf{f}_i$ is optimal for R at γ^*, then $\mathbf{f}$ is also optimal for R at γ^* by lemma F.42.

Suppose that $\mathbf{f}$ is optimal for R at $\gamma^* \in (0,1)$, while for some i, $\mathbf{f}_i$ is not optimal for R at γ^*. In particular, suppose R induces optimal policy set Π_{γ^*} at discount rate γ^*. Then by proposition D.35, we can construct a reward function R' which has optimal policy set Π_{γ^*} at discount rate γ.

Suppose $\alpha_i = 1$. Since $\mathbf{f}_i$ was not optimal at γ^*, $\mathbf{f}_i$ isn't optimal at γ either. But this is absurd, since then $\mathbf{f}(\gamma) = \mathbf{f}_i(\gamma)$ and so they must both be optimal at γ, a contradiction. This means that $\mathbf{f}_i$ is optimal for R at γ^*.

Suppose $\alpha_i < 1$. As $\mathbf{f}_i$ is not optimal for R at γ but $\mathbf{f}$ is, there must be another $\mathbf{f}_{i'}$ such that $\mathbf{f}_{i'}(\gamma)^\top \mathbf{r} \geq \mathbf{f}(\gamma)^\top \mathbf{r} > \mathbf{f}_i(\gamma)^\top \mathbf{r}$. Since we assumed that $\alpha_i > 0$, the first inequality must be strict. This means that $\mathbf{f}$ cannot be optimal for R at γ, a contradiction. Therefore, $\mathbf{f}_i$ must be optimal for R at γ^*. $\square$

Corollary F.44 (Visit distribution functions which agree at any γ, must be optimal together)**.** *Let* $\mathbf{f}, \mathbf{f}' \in \mathcal{F}(s)$*. Suppose* $\exists \gamma \in (0,1) : \mathbf{f}(\gamma) = \mathbf{f}'(\gamma)$*. Then* $\forall \gamma^* \in (0,1)$*:*

$$\operatorname{supp}\left(\mathbf{f}(\gamma^*) \geq \mathcal{F}(s, \gamma^*)\right) = \operatorname{supp}\left(\mathbf{f}'(\gamma^*) \geq \mathcal{F}(s, \gamma^*)\right).$$

Proof. Apply proposition F.43. □

Figure F.6 illustrates the power of corollary F.44.

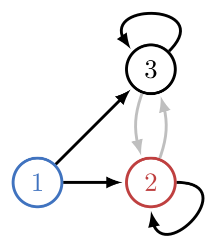


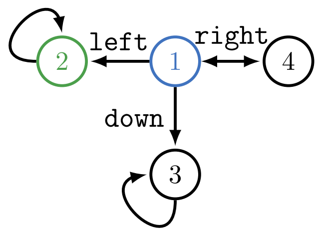
Figure F.6: When $\gamma = \frac{1}{2}$, the trajectories $s_1 s_2 s_3 s_3 \ldots$ and $s_1 s_3 s_2 s_2 \ldots$ induce the same state visitation distribution. Therefore, corollary F.44 shows that at *any* $\gamma \in (0,1)$, one trajectory is optimal iff the other trajectory is. Basic algebra confirms the point: one of these trajectories is optimal iff $R(s_2) = R(s_3)$; in that case, every policy is optimal.

Consider again proposition D.35:

Theorem (A means of transferring optimal policy sets across discount rates)**.** Suppose reward function R has optimal policy set $\Pi^*(R, \gamma)$ at discount rate $\gamma \in (0,1)$. For any $\gamma^* \in (0,1)$, we can construct a reward function R' such that $\Pi^*\left(R', \gamma^*\right) = \Pi^*(R, \gamma)$.

Figure F.7 shows that $V_R^\pi(s, \gamma) \geq V_R^{\pi'}(s, \gamma)$ does not imply $V_{R'}^\pi(s, \gamma^*) \geq V_{R'}^{\pi'}(s, \gamma^*)$. The new reward function R' is only guaranteed to have the same optimal policy set, not the same policy ordering.

Let R be a reward function and $\gamma^* \in (0,1), \gamma \in [0,1)$. Figure F.7 demonstrates that there does not always exist R' such that $\forall s, \pi : V_R^\pi(s, \gamma^*) = V_{R'}^\pi(s, \gamma)$. By preserving the optimal value function across discount rates, R' necessarily differs on its other value functions.



State	R	R'
s_1	0	$\frac{\gamma - \gamma^*}{1-\gamma}$
s_2	1	1
s_3	0	0
s_4	0	$\frac{\gamma(\gamma - \gamma^*)}{1-\gamma}$

Figure F.7: R and the transformed $R' \coloneqq \left(\mathbf{I} - \gamma^* \mathbf{T}^{\pi_{\texttt{left}}}\right) V^*_R(\cdot, \gamma)$ for which proposition D.35's reward transformation does not preserve the policy ordering. Even though $V_R^{\pi_{\texttt{down}}}(s_1, \gamma) = V_R^{\pi_{\texttt{right}}}(s_1, \gamma) = 0$, when $\gamma^* > \gamma$, $V_{R'}^{\pi_{\texttt{down}}}(s_1, \gamma) > V_{R'}^{\pi_{\texttt{right}}}(s_1, \gamma)$.

Question F.45 (Does there exist a transformation like that of proposition D.35 which preserves the entire policy ordering of R?)**.**

Question F.46 (Can we characterize the permissible policy orderings in a given MDP? Does the discount rate affect the permissible orderings?)**.**

Proposition F.47 (Almost all reward functions induce strict non-dominated visitation distribution orderings)**.** *Let s be a state and $\gamma \in (0, 1)$.*

$$\left\{\mathbf{r} \mid \mathbf{r} \in \mathbb{R}^{|\mathcal{S}|}, \exists \mathbf{f}, \mathbf{f}' \in \mathcal{F}_{\text{nd}}(s) : \mathbf{f} \neq \mathbf{f}' \wedge \mathbf{f}(\gamma)^\top \mathbf{r} = \mathbf{f}'(\gamma)^\top \mathbf{r}\right\} \tag{F.29}$$

has Lebesgue measure zero. The same statement holds for $\mathcal{F}$ instead of $\mathcal{F}_{\text{nd}}$ for all but finitely many $\gamma \in (0, 1)$.

Proof. Since $\mathbf{f}, \mathbf{f}' \in \mathcal{F}_{\text{nd}}(s)$ are distinct, $\forall \gamma \in (0, 1) : \mathbf{f}(\gamma) \neq \mathbf{f}'(\gamma)$ by lemma D.52. Apply lemma D.12.

All distinct $\mathbf{f}, \mathbf{f}' \in \mathcal{F}(s)$ disagree for all but finitely many γ by corollary F.39. Then apply lemma D.12 for these γ at which they all disagree, and the second claim follows. □

Remark. For almost all reward functions, proposition F.47 prohibits value equality anywhere in the non-dominated visit distribution ordering. In contrast, lemma F.106 only shows that almost all reward functions have a unique *optimal* visit distribution.

Lemma F.48 (Reward negation flips the visit distribution ordering)**.** *Let s be a state, $\mathbf{f}, \mathbf{f}' \in \mathcal{F}(s)$, $\mathbf{r} \in \mathbb{R}^{|\mathcal{S}|}$, and $\gamma \in (0,1)$. $\mathbf{f}(\gamma)^\top \mathbf{r} \geq \mathbf{f}'(\gamma)^\top \mathbf{r}$ iff $\mathbf{f}(\gamma)^\top(-\mathbf{r}) \leq \mathbf{f}'(\gamma)^\top(-\mathbf{r})$. In particular, optimal policies for $\mathbf{r}$ minimize value for $-\mathbf{r}$.*

Proof. This follows directly from the fact that $\forall a, b \in \mathbb{R} : a \geq b$ iff $-a \leq -b$. □

F.1.6 Visit distributions functions induced by non-stationary policies

Definition F.49 (Non-stationary visit distribution functions)**.** Let Π^{HD} be the set of history-dependent (or non-stationary) deterministic policies [68]. $\mathcal{F}^{\text{HD}}(s) := \left\{\mathbf{f}^{\pi,s} \mid \pi \in \Pi^{\text{HD}}\right\}$.

We continue using Π to denote the space of deterministic stationary policies. Clearly, $\mathcal{F}(s) \subseteq \mathcal{F}^{\text{HD}}(s)$.

Proposition F.50 (Visit distribution functions induced by non-stationary or non-deterministic policies are dominated)**.**

Proof. This follows from the fact that every reward function has a stationary, deterministic optimal policy. □

Proposition F.51 ($\mathcal{F}^{\text{HD}}(s)$ is finite iff stationarity is irrelevant)**.** *$\mathcal{F}^{HD}(s)$ is finite iff $\mathcal{F}(s) = \mathcal{F}^{HD}(s)$.*

Proof. Suppose $\mathcal{F}^{\text{HD}}(s)$ is finite but $\mathcal{F}(s) \neq \mathcal{F}^{\text{HD}}(s)$. Since $\mathcal{F}(s) \subseteq \mathcal{F}^{\text{HD}}(s)$, let $\mathbf{f}^\pi \in \mathcal{F}^{\text{HD}}(s) \setminus \mathcal{F}(s)$. π must exhibit a non-stationarity which affects the induced visit distribution, which means that starting from s, following π must induce a positive probability of visiting some state s' twice. If $\left|T(s')\right| = 1$, π cannot affect the induced visit distribution with a non-stationarity at s', so $\left|T(s')\right| > 1$.

Since s' can reach itself with positive probability, there exists some stationary policy π' which has positive probability of visiting s' k times, for any natural number k. Let action a be such that $T(s', \pi'(s')) \neq T(s', a)$ (such an action exists because $\left|T(s')\right| > 1$). Let π_k

be a non-stationary policy which agrees with π' for the first k visits to s', but which takes action a after visit k.

Each π_k induces a different state visit distribution function; since there are infinitely many of them, $\mathcal{F}^{\text{HD}}(s)$ cannot be finite if $\mathcal{F}(s) \neq \mathcal{F}^{\text{HD}}(s)$. Then if $\mathcal{F}^{\text{HD}}(s)$, then $\mathcal{F}(s) = \mathcal{F}^{\text{HD}}(s)$.

The reverse direction follows because $\left|\mathcal{F}(s)\right| \leq |\mathcal{S}|^{|\mathcal{A}|}$, which is finite. ☐

Proposition F.52 (Characterization of when stationarity is relevant)**.** *$\mathcal{F}^{HD}(s) = \mathcal{F}(s)$ iff for all states s' reachable from s with positive probability, if s' can reach itself with positive probability, $\left|T(s')\right| = 1$.*

Proof. For the forward direction, apply the reasoning from the proof of proposition F.51.

Reverse direction: if non-stationarity is to affect the induced visit distribution function, there must be some s' which can reach itself with positive probability, and which is reachable from s. But if $\left|T(s')\right| = 1$ for all such s', then non-stationarity does not affect the induced visitation distribution function (because all actions have the same local dynamics at s'). ☐

F.1.7 Generalized non-domination results

Lemma F.53 (Idempotence of non-domination)**.** *Let $X, Y \subsetneq \mathbb{R}^d$ be finite.*

$$\text{ND}\left(\text{ND}\left(X\right) \setminus Y\right) = \text{ND}\left(X\right) \setminus Y.$$

Proof. Trivially, $\text{ND}\left(\text{ND}\left(X\right) \setminus Y\right) \subseteq \text{ND}\left(X\right) \setminus Y$. Let $\mathbf{x} \in \text{ND}\left(X\right) \setminus Y$. Then $\exists \mathbf{r} \in \mathbb{R}^d :$ $\mathbf{x}^\top \mathbf{r} > \max_{\mathbf{x}' \in X \setminus \{\mathbf{x}\}} \mathbf{x}'^\top \mathbf{r} \geq \max_{\mathbf{x}' \in X \setminus \left(Y \cup \{\mathbf{x}\}\right)} \mathbf{x}'^\top \mathbf{r}$, and so $\mathbf{x} \in \text{ND}\left(\text{ND}\left(X\right) \setminus Y\right)$ by the definition of non-domination (definition 5.6).

Then $\text{ND}\left(\text{ND}\left(X\right) \setminus Y\right) \supseteq \text{ND}\left(X\right) \setminus Y$, and so $\text{ND}\left(\text{ND}\left(X\right) \setminus Y\right) = \text{ND}\left(X\right) \setminus Y$. ☐

Lemma F.54 (Non-dominated inclusion relation)**.** *Let $A, B \subsetneq \mathbb{R}^{|\mathcal{S}|}$ be finite. If $A \subseteq B$, then* $\text{ND}\left(B\right) \cap A \subseteq \text{ND}\left(A\right)$.

Proof. Suppose $\mathbf{a} \in \mathrm{ND}\,(B) \cap A$.

$$\exists \mathbf{r} \in \mathbb{R}^{|\mathcal{S}|} : \mathbf{a}^\top \mathbf{r} > \max_{\mathbf{b} \in B \setminus \{\mathbf{a}\}} \mathbf{b}^\top \mathbf{r} \tag{F.30}$$

$$\geq \max_{\mathbf{b} \in A \setminus \{\mathbf{a}\}} \mathbf{b}^\top \mathbf{r}. \tag{F.31}$$

Equation (F.30) follows by the definition of ND (definition D.14). Equation (F.31) follows because $A \subseteq B$. Therefore, $\exists \mathbf{r} \in \mathbb{R}^{|\mathcal{S}|} : \mathbf{a}^\top \mathbf{r} > \max_{\mathbf{b} \in A \setminus \{\mathbf{a}\}} \mathbf{b}^\top \mathbf{r}$. Since $\mathbf{a} \in A$, $\mathbf{a} \in \mathrm{ND}\,(A)$. ☐

Lemma F.55 (Permutation commutes with non-dominance)**.** *Let $X \subsetneq \mathbb{R}^{|\mathcal{S}|}$ be finite and let $\phi \in S_{|\mathcal{S}|}$. $\phi \cdot \mathrm{ND}\,(X) = \mathrm{ND}\,(\phi \cdot X)$.*

Proof. $\mathbf{x} \in \mathrm{ND}\,(X)$ iff $\exists \mathbf{r} \in \mathbb{R}^{|\mathcal{S}|}$ for which

$$\mathbf{x}^\top \mathbf{r} > \max_{\mathbf{x}' \in X \setminus \{\mathbf{x}\}} \mathbf{x}'^\top \mathbf{r} \tag{F.32}$$

$$\mathbf{x}^\top \mathbf{P}_\phi^{-1} \mathbf{P}_\phi \mathbf{r} > \max_{\mathbf{x}' \in X \setminus \{\mathbf{x}\}} \mathbf{x}'^\top \mathbf{P}_\phi^{-1} \mathbf{P}_\phi \mathbf{r} \tag{F.33}$$

$$\left(\mathbf{P}_\phi \mathbf{x}\right)^\top \mathbf{P}_\phi \mathbf{r} > \max_{\mathbf{x}' \in X \setminus \{\mathbf{x}\}} \left(\mathbf{P}_\phi \mathbf{x}'\right)^\top \mathbf{P}_\phi \mathbf{r} \tag{F.34}$$

$$\left(\mathbf{P}_\phi \mathbf{x}\right)^\top \mathbf{P}_\phi \mathbf{r} > \max_{\mathbf{x}'_\phi \in \phi \cdot X \setminus \left\{\mathbf{P}_\phi \mathbf{x}\right\}} \mathbf{x}'^\top_\phi \mathbf{P}_\phi \mathbf{r}. \tag{F.35}$$

Equation (F.33) follows because the identity matrix $\mathbf{I} = \mathbf{P}_\phi^{-1} \mathbf{P}_\phi$, since permutation matrices are invertible. Equation (F.34) follows because permutation matrices are orthogonal, and so $\left(\mathbf{P}_\phi^{-1}\right)^\top = \mathbf{P}_\phi$. By eq. (F.35), $\phi \cdot \mathrm{ND}\,(X) = \mathrm{ND}\,(\phi \cdot X)$. ☐

Lemma F.56 (Number of non-dominated linear functionals)**.** *Let $X \subsetneq \mathbb{R}^{|\mathcal{S}|}$ be finite. $1 < \left|\mathrm{ND}\,(X)\right|$ iff $1 < |X|$.*

Proof. Since $\mathrm{ND}\,(X) \subseteq X$, $1 < \left|\mathrm{ND}\,(X)\right| \leq |X|$ implies that $1 < \left|\mathrm{ND}\,(X)\right| \implies 1 < |X|$.

Suppose $1 < |X|$. Since X is non-empty, $\mathbf{0}$ the zero vector has some optimal $\mathbf{x} \in \mathrm{ND}\,(X)$ by lemma D.15. Since $1 < |X|$, let $\mathbf{x}' \in X$ be such that $\mathbf{x}' \neq \mathbf{x}$. Since $\mathbf{x} \in \mathrm{ND}\,(X)$, let $\mathbf{r} \in \mathbb{R}^{|\mathcal{S}|}$ be such that $\mathbf{x}^\top \mathbf{r} > \max_{\mathbf{x}'' \in X \setminus \{\mathbf{x}\}} \mathbf{x}''^\top \mathbf{r} \geq \mathbf{x}'^\top \mathbf{r}$.

Then for $\mathbf{r}' := -\mathbf{r}$, $\mathbf{x}'^\top \mathbf{r}' > \mathbf{x}^\top \mathbf{r}'$. By lemma D.15, $\exists \mathbf{x}_{\text{nd}} \in \text{ND}(X)$ such that $\mathbf{x}_{\text{nd}}^\top \mathbf{r}' > \max_{\mathbf{x}'' \in X \setminus \{\mathbf{x}_{\text{nd}}\}} \mathbf{x}''^\top \mathbf{r}'$. But $\mathbf{x}_{\text{nd}}^\top \mathbf{r}' \geq \mathbf{x}'^\top \mathbf{r}' > \mathbf{x}^\top \mathbf{r}'$, and so $\mathbf{x}_{\text{nd}} \neq \mathbf{x}$. Then $1 < \left|\text{ND}(X)\right|$. □

Lemma F.57 (Sufficient conditions for a linear functional being non-dominated)**.** *Let $X \subsetneq \mathbb{R}^{|\mathcal{S}|}$ be finite and let $\mathbf{x} \in X$. If*

1. $\exists \mathbf{e}_i : \mathbf{x}^\top \mathbf{e}_i > \max_{\mathbf{x}' \in X \setminus \{\mathbf{x}\}} \mathbf{x}'^\top \mathbf{e}_i$, *or*
2. $\exists \mathbf{e}_i : \mathbf{x}^\top \mathbf{e}_i < \min_{\mathbf{x}' \in X \setminus \{\mathbf{x}\}} \mathbf{x}'^\top \mathbf{e}_i$,

then $\mathbf{x} \in \text{ND}(X)$.

Proof. The first item follows directly from the definition of $\text{ND}(X)$ (definition D.14), where $\mathbf{x}$ is strictly maximal for vector $\mathbf{r} := \mathbf{e}_i$. For the second item, $\mathbf{x}$ is strictly maximal for vector $\mathbf{r} := -\mathbf{e}_i$. □

Conjecture F.58 (Expanded sufficient conditions for a linear functional being non-dominated)**.** Lemma F.57 can be expanded to account for reasoning like "$\mathbf{x}$ has non-strictly maximal entries along dimensions 1 and 2; among those $\mathbf{x}' \in X$ tied with $\mathbf{x}$ on both dimensions, $\mathbf{x}$ has strictly maximal entry value on dimension 3."

Proposition F.59 (Non-dominated linear functionals are convex independent of other functionals)**.** *Let $X \subsetneq \mathbb{R}^{|\mathcal{S}|}$ be finite and let $\mathbf{x} \in X$. If $\mathbf{x} \in \text{ND}(X)$, then $\mathbf{x}$ cannot be written as a convex combination of $\mathbf{x}_1, \ldots, \mathbf{x}_n \in X \setminus \{\mathbf{x}\}$.*

Proof. Suppose $\mathbf{x} \in \text{ND}(X)$ can be written as the convex combination $\mathbf{x} = \sum_{i=1}^n \theta_i \mathbf{x}_i$. By non-domination, $\exists \mathbf{r} \in \mathbb{R}^{|\mathcal{S}|} : \mathbf{x}^\top \mathbf{r} > \max_{\mathbf{x}' \in X \setminus \{\mathbf{x}\}} \mathbf{x}'^\top \mathbf{r}$. Therefore,

$$\left(\sum_{i=1}^n \theta_i \mathbf{x}_i\right)^\top \mathbf{r} = \mathbf{x}^\top \mathbf{r} \tag{F.36}$$

$$> \max_{\mathbf{x}' \in X \setminus \{\mathbf{x}\}} \mathbf{x}'^\top \mathbf{r} \tag{F.37}$$

$$\geq \max_i \mathbf{x}_i^\top \mathbf{r}. \tag{F.38}$$

But this is a contradiction; a convex combination of values cannot be strictly greater than all of its constituent values. $\square$

The reverse direction of proposition F.59 is not true. Let $X := \left\{ \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} .5 \\ .5 \end{pmatrix}, \begin{pmatrix} .4 \\ .6 \end{pmatrix} \right\}$. $\mathrm{ND}(X) = \left\{ \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} .4 \\ .6 \end{pmatrix} \right\}$. $\mathbf{x} := \begin{pmatrix} .5 \\ .5 \end{pmatrix}$ is dominated, even though it cannot be written as a convex combination of functionals $X \setminus \{\mathbf{x}\}$.

Conjecture F.60 (Characterizing when $X = \mathrm{ND}(X)$)**.** Let $X \subsetneq \mathbb{R}^{|\mathcal{S}|}$ be finite. $X = \mathrm{ND}(X)$ iff $|X| \leq 1$ or $\forall \mathbf{x}, \mathbf{x}' \in X : \exists c > 0 : \mathbf{x} = c\mathbf{x}'$.

Conjecture F.61 (ND argmax equality implies argmax equality)**.** Let $X \subsetneq \mathbb{R}^{|\mathcal{S}|}$ be finite and let $\mathbf{r}, \mathbf{r}' \in \mathbb{R}^{|\mathcal{S}|}$. If $\arg\max_{\mathbf{x} \in \mathrm{ND}(X)} \mathbf{x}^\top \mathbf{r} = \arg\max_{\mathbf{x} \in \mathrm{ND}(X)} \mathbf{x}^\top \mathbf{r}'$, then $\arg\max_{\mathbf{x} \in X} \mathbf{x}^\top \mathbf{r} = \arg\max_{\mathbf{x} \in X} \mathbf{x}^\top \mathbf{r}'$.

Lemma F.62 (Invariances of linear functional optimality probability)**.** *Let $A, B \subsetneq \mathbb{R}^{|\mathcal{S}|}$ be finite.*

1. $p_{\mathcal{D}_{any}}(A \geq B) = p_{\mathcal{D}_{any}}\big(A \geq \big(B \setminus A\big)\big)$.
2. $p_{\mathcal{D}_{any}}(A \geq B) = p_{\mathcal{D}_{any}}\big(A \geq \mathrm{ND}(B)\big) = p_{\mathcal{D}_{any}}\big(\mathrm{ND}(A) \geq \mathrm{ND}(B)\big)$.
3. *For any* $\mathbf{x} \in \mathbb{R}^{|\mathcal{S}|}$, $p_{\mathcal{D}_{any}}(A \geq B) = p_{\mathcal{D}_{any}}(A - \mathbf{x} \geq B - \mathbf{x})$.

Proof. Item 1:

$$p_{\mathcal{D}_{\text{any}}}(A \geq B) := \mathbb{P}_{\mathbf{r} \sim \mathcal{D}_{\text{any}}}\left(\max_{\mathbf{a} \in A} \mathbf{a}^\top \mathbf{r} \geq \max_{\mathbf{b} \in B} \mathbf{b}^\top \mathbf{r} \right) \tag{F.39}$$

$$= \mathbb{P}_{\mathbf{r} \sim \mathcal{D}_{\text{any}}}\left(\max_{\mathbf{a} \in A} \mathbf{a}^\top \mathbf{r} \geq \max\left(\max_{\mathbf{b} \in B \cap A} \mathbf{b}^\top \mathbf{r}, \max_{\mathbf{b} \in B \setminus A} \mathbf{b}^\top \mathbf{r} \right) \right) \tag{F.40}$$

$$= \mathbb{P}_{\mathbf{r} \sim \mathcal{D}_{\text{any}}}\left(\max_{\mathbf{a} \in A} \mathbf{a}^\top \mathbf{r} \geq \max_{\mathbf{b} \in B \setminus A} \mathbf{b}^\top \mathbf{r} \right) \tag{F.41}$$

$$=: p_{\mathcal{D}_{\text{any}}}\Big(A \geq \big(B \setminus A\big)\Big). \tag{F.42}$$

Equation (F.41) holds because $\max_{\mathbf{a}\in A}\mathbf{a}^\top\mathbf{r} \geq \max_{\mathbf{b}\in B\cap A}\mathbf{b}^\top\mathbf{r}$ for all $\mathbf{r}\in\mathbb{R}^{|\mathcal{S}|}$, and so the constraint is vacuous.

Item 2:

$$
\begin{aligned}
p_{\mathcal{D}_{\text{any}}}(A \geq B) &\coloneqq \mathbb{P}_{\mathbf{r}\sim\mathcal{D}_{\text{any}}}\left(\max_{\mathbf{a}\in A}\mathbf{a}^\top\mathbf{r} \geq \max_{\mathbf{b}\in B}\mathbf{b}^\top\mathbf{r}\right) && \text{(F.43)}\\
&= \mathbb{P}_{\mathbf{r}\sim\mathcal{D}_{\text{any}}}\left(\max_{\mathbf{a}\in A}\mathbf{a}^\top\mathbf{r} \geq \max_{\mathbf{b}\in \mathrm{ND}(B)}\mathbf{b}^\top\mathbf{r}\right) && \text{(F.44)}\\
&= \mathbb{P}_{\mathbf{r}\sim\mathcal{D}_{\text{any}}}\left(\max_{\mathbf{a}\in \mathrm{ND}(A)}\mathbf{a}^\top\mathbf{r} \geq \max_{\mathbf{b}\in \mathrm{ND}(B)}\mathbf{b}^\top\mathbf{r}\right) && \text{(F.45)}\\
&\eqqcolon p_{\mathcal{D}_{\text{any}}}\big(\mathrm{ND}(A) \geq \mathrm{ND}(B)\big). && \text{(F.46)}
\end{aligned}
$$

Equation (F.44) follows because corollary D.16 shows that $\max_{\mathbf{b}\in B}\mathbf{b}^\top\mathbf{r} = \max_{\mathbf{b}\in\mathrm{ND}(B)}\mathbf{b}^\top\mathbf{r}$. Similarly, eq. (F.45) follows because corollary D.16 shows that $\max_{\mathbf{a}\in A}\mathbf{a}^\top\mathbf{r} = \max_{\mathbf{a}\in\mathrm{ND}(A)}\mathbf{a}^\top\mathbf{r}$.

Item 3: For all $\mathbf{r}\in\mathbb{R}^{|\mathcal{S}|}$, $\max_{\mathbf{a}\in A}\mathbf{a}^\top\mathbf{r} \geq \max_{\mathbf{b}\in B}\mathbf{b}^\top\mathbf{r}$ iff $\mathbf{x}^\top\mathbf{r} + \max_{\mathbf{a}\in A}\mathbf{a}^\top\mathbf{r} \geq \mathbf{x}^\top\mathbf{r} + \max_{\mathbf{b}\in B}\mathbf{b}^\top\mathbf{r}$ iff $\max_{\mathbf{a}\in A}(\mathbf{a}+\mathbf{x})^\top\mathbf{r} \geq \mathbf{x}^\top\mathbf{r} + \max_{\mathbf{b}\in B}(\mathbf{b}+x)^\top\mathbf{r}$. ☐

Proposition F.63 (Additivity of linear functional optimality probability for $\mathcal{D}_{\text{cont}}$)**.** *For finite* $A, B \subsetneq \mathbb{R}^{|\mathcal{S}|}$, $p_{\mathcal{D}_{cont}}(A \geq B) = \sum_{\mathbf{a}\in A} p_{\mathcal{D}_{cont}}(\{\mathbf{a}\} \geq B)$.

Proof. Let $A = \{\mathbf{a}_1, \ldots, \mathbf{a}_n\}$. For each i between 1 and n, define the event $X_i \coloneqq \{\mathbf{r} \mid \mathbf{a}_i^\top\mathbf{r} \geq \max_{\mathbf{b}\in B}\mathbf{b}^\top\mathbf{r}\}$.

$$
\begin{aligned}
p_{\mathcal{D}_{\text{cont}}}(A \geq B) &\coloneqq \mathbb{P}_{\mathbf{r}\sim\mathcal{D}_{\text{cont}}}\left(\max_{\mathbf{a}\in A}\mathbf{a}^\top\mathbf{r} \geq \max_{\mathbf{b}\in B}\mathbf{b}^\top\mathbf{r}\right) && \text{(F.47)}\\
&= \mathbb{P}_{\mathcal{D}_{\text{cont}}}\left(\bigcup_{i=1}^{n} X_i\right) && \text{(F.48)}\\
&= \sum_{k=1}^{n}(-1)^{k-1}\sum_{\substack{I\subseteq\{1,\ldots,n\},\\ |I|=k}} \mathbb{P}_{\mathcal{D}_{\text{cont}}}\left(\bigcap_{i\in I} X_i\right) && \text{(F.49)}
\end{aligned}
$$

$$= \sum_{i=1}^{n} \mathbb{P}_{\mathcal{D}_{\text{cont}}}(X_i) \tag{F.50}$$

$$=: \sum_{\mathbf{a}\in A} p_{\mathcal{D}_{\text{cont}}}\left(\{\mathbf{a}\} \geq B\right). \tag{F.51}$$

Equation (F.49) follows by the inclusion-exclusion formula. Almost all $\mathbf{r}' \in \mathbb{R}^{|\mathcal{S}|}$ are maximized by a unique functional $\mathbf{a}^*$ by corollary D.13. By the same result, since $\mathcal{D}_{\text{cont}}$ is continuous, it assigns probability 0 to $\mathbf{r}'$ which are simultaneously maximized by at least two functionals. Therefore, all terms $k > 1$ in eq. (F.49) must vanish, and eq. (F.50) follows. □

Lemma F.64 (Positive probability under $\mathcal{D}_{\text{cont}}$ implies non-dominated functional)**.** *Let $A \subsetneq \mathbb{R}^{|\mathcal{S}|}$ be finite. Let $\mathbf{a} \in A$. If $p_{\mathcal{D}_{cont}}\left(\{\mathbf{a}\} \geq A\right) > 0$, then $\mathbf{a} \in \operatorname{ND}(A)$.*

Proof. Let $\mathbf{a} \in A$. If $\mathbf{a} \notin \operatorname{ND}(A)$, then by lemma D.15,

$$\forall \mathbf{r} \in \mathbb{R}^{|\mathcal{S}|} : \mathbf{a}^\top \mathbf{r} = \max_{\mathbf{a}'\in A} \mathbf{a}'^\top \mathbf{r} \implies \exists \mathbf{a}^* \in \operatorname{ND}(A) : \mathbf{a}^{*\top} \mathbf{r} = \mathbf{a}^\top \mathbf{r}. \tag{F.52}$$

$\mathbf{a} \neq \mathbf{a}^*$ since $\mathbf{a} \notin \operatorname{ND}(A)$. But corollary D.13 shows that continuous $\mathcal{D}_{\text{cont}}$ assign 0 probability to the set of $\mathbf{r}$ for which multiple $\mathbf{a}^*, \mathbf{a}$ are optimal. Then $p_{\mathcal{D}_{\text{cont}}}\left(\{\mathbf{a}\}, A\right) = 0$, a contradiction. So we conclude that $\mathbf{a} \in \operatorname{ND}(A)$. □

F.1.8 Non-dominated visit distribution functions

Corollary F.65 ($\left|\mathcal{F}_{\text{nd}}(s)\right| \geq 1$, with equality iff $\left|\mathcal{F}(s)\right| = 1$)**.**

Proof. We show $\left|\mathcal{F}_{\text{nd}}(s)\right| = 1$ implies $\left|\mathcal{F}(s)\right| = 1$ by proving the contrapositive. Suppose $\left|\mathcal{F}(s)\right| \geq 2$, let $\mathbf{f}, \mathbf{f}' \in \mathcal{F}(s)$, and let $\gamma \in (0, 1)$ be such that $\mathbf{f}(\gamma) \neq \mathbf{f}'(\gamma)$. Then $1 < \left|\mathcal{F}(s, \gamma)\right|$. Apply lemma F.56 to conclude that $1 < \left|\operatorname{ND}\left(\mathcal{F}(s, \gamma)\right)\right| = \left|\mathcal{F}_{\text{nd}}(s, \gamma)\right| \leq \left|\mathcal{F}_{\text{nd}}(s)\right|$, with the first equality following by lemma D.38, and the second following from the definition of $\mathcal{F}_{\text{nd}}(s, \gamma)$ (definition D.36).

If $\left|\mathcal{F}(s)\right| = 1$, then the sole visit distribution function is trivially strictly optimal for all $\mathbf{r} \in \mathbb{R}^{|\mathcal{S}|}$ at all $\gamma \in (0, 1)$. Thus $\left|\mathcal{F}_{\text{nd}}(s)\right| = 1$. □

Lemma F.66 (Each reward function has an optimal non-dominated visit distribution)**.** *Let* $\mathbf{r} \in \mathbb{R}^{|\mathcal{S}|}$ *and* $\gamma \in [0, 1)$. $\exists \mathbf{f} \in \arg\max_{\mathbf{f} \in \mathcal{F}(s)} \mathbf{f}(\gamma)^\top \mathbf{r} : \mathbf{f} \in \mathcal{F}_{\text{nd}}(s)$.

Proof. Apply lemma D.15 with $X \coloneqq \mathcal{F}(s, \gamma)$. ☐

Corollary F.67 (Strict visitation optimality is sufficient for non-domination)**.** *Let* $\gamma \in [0, 1)$ *and let* $s, s' \in \mathcal{S}$. *At least one element of* $\left\{\arg\max_{\mathbf{f} \in \mathcal{F}(s)} \mathbf{f}(\gamma)^\top \mathbf{e}_{s'}\right\}$ *is non-dominated.*

Proof. Apply lemma F.66 to the reward function $\mathbf{e}_{s'}$. ☐

Remark. Corollary F.67 implies that if s' is reachable with positive probability from s, then there is at least one non-dominated visit distribution function which assigns s' positive visitation frequency. In this sense, $\mathcal{F}_{\text{nd}}(s)$ "covers" the states reachable from s.

Definition F.68 (Surely reachable children)**.** The *surely reachable children* of s are $Ch_{\text{sure}}(s) \coloneqq \left\{s' \mid \exists a : T(s, a) = \mathbf{e}_{s'}\right\}$. Note that determinism implies that $Ch(s) = Ch_{\text{sure}}(s)$.

Corollary F.69 (Minimum number of non-dominated visit distribution functions)**.** *Suppose* $\mathbf{f}_1, \ldots, \mathbf{f}_k \in \mathcal{F}(s)$ *place strictly greater visitation frequency on some corresponding states* $s_1, \ldots, s_k$ *than do other visitation distributions. Then* $\mathbf{f}_1, \ldots, \mathbf{f}_k \in \mathcal{F}_{\text{nd}}(s)$ *and* $\left|\mathcal{F}_{\text{nd}}(s)\right| \geq k$. *In particular,* $\left|\mathcal{F}_{\text{nd}}(s)\right| \geq \left|Ch_{sure}(s)\right|$.

Proof. $\mathbf{f}_1, \ldots, \mathbf{f}_k \in \mathcal{F}_{\text{nd}}(s)$ by corollary F.67. $\left|\mathcal{F}_{\text{nd}}(s)\right| \geq \left|Ch_{\text{sure}}(s)\right|$ because each $s' \in Ch_{\text{sure}}(s)$ must have at least one visitation frequency-maximizing visit distribution function $\mathbf{f} \in \mathcal{F}_{\text{nd}}(s)$. ☐

Corollary F.70 (When $\left|\mathcal{F}(s)\right| \leq 2$, $\mathcal{F}(s) = \mathcal{F}_{\text{nd}}(s)$)**.**

Proof. $\mathcal{F}(s) = \mathcal{F}_{\text{nd}}(s)$ trivially when $\left|\mathcal{F}(s)\right| = 1$. When $\left|\mathcal{F}(s)\right| = 2$, each visitation distribution must visit at least one state with strictly greater frequency than does the other visitation distribution; apply corollary F.69. ☐

Lemma F.71 (Initial-state non-domination implies non-dom. at visited states)**.** *If* $\mathbf{f}^{\pi,s} \in \mathcal{F}_{\text{nd}}(s)$ *is strictly optimal for reward function* R *at discount rate* γ *and* $\mathbf{f}^{\pi,s}(\gamma)^\top \mathbf{e}_{s'} > 0$*, then* $\mathbf{f}^{\pi,s'} \in \mathcal{F}_{\text{nd}}(s')$*.*

Proof. Because $\mathbf{f}^{\pi,s}(\gamma)^\top \mathbf{e}_{s'} > 0$, π can induce some trajectory prefix $(s, s_1, \ldots s_{n-1}, s')$ with positive probability. By definition 5.3,

$$\mathbf{f}^{\pi,s}(\gamma) = \mathbf{e}_s + \gamma \mathop{\mathbb{E}}_{s_1 \sim T(s,\pi(s))}\left[\mathbf{e}_{s_1} + \ldots + \gamma \mathop{\mathbb{E}}_{s' \sim T(s_{n-1},\pi(s_{n-1}))}\left[\mathbf{f}^{\pi,s'}(\gamma)\right]\right]. \tag{F.53}$$

Since $\mathbf{f}^\pi$ is the strictly optimal visit distribution from state s for reward function R at discount rate γ, eq. (F.53) shows that π must in particular induce a strictly optimal visit distribution for R at γ starting from state s'. If not, another visit distribution would induce optimality starting from s', contradicting the strict optimality of $\mathbf{f}^\pi$. □

Corollary F.72 (Domination at visited state implies domination at initial state)**.** *If* $\mathbf{f}^{\pi,s'}$ *is dominated at* s'*, then either* $\mathbf{f}^{\pi,s}$ *is dominated at* s *or* $\mathbf{f}^{\pi,s\top}\mathbf{e}_{s'} = 0$*.*

Proof. This statement is the contrapositive of lemma F.71. □

Lemma F.73 (Dominated child state distributions induce dominated visit distributions)**.** *Let action* a *induce a dominated child state distribution* $\mathbf{d} \coloneqq T(s_1, a)$ *at state* s_1*, and let* s *be the initial state. If a policy* π *has* $\pi(s_1) = a$ *and* $\mathbf{f}^\pi(\gamma)^\top \mathbf{e}_{s_1} \neq 0$*, then* π *induces a dominated visit distribution* $\mathbf{f}^\pi \in \mathcal{F}(s) \setminus \mathcal{F}_{\text{nd}}(s)$*.*

Proof. By lemma F.8, $\mathbf{f}^{\pi,s_1} \in \mathcal{F}(s_1) \setminus \mathcal{F}_{\text{nd}}(s_1)$. Since $\mathbf{f}^\pi(\gamma)^\top \mathbf{e}_{s_1} \neq 0$, $\mathbf{f}^\pi \in \mathcal{F}(s) \setminus \mathcal{F}_{\text{nd}}(s)$ by corollary F.72. □

The following extends lemma F.75 to account for non-domination.

Kulkarni et al. [44] learn to estimate $\text{Succ}(s,a) \coloneqq \mathbb{E}_{s' \sim T(s,a)}\left[\mathbf{f}^{\pi,s'}(\gamma)\right]$ in order to infer state-space bottlenecks. We prove that bottlenecks "factorize" state visit distribution functions.

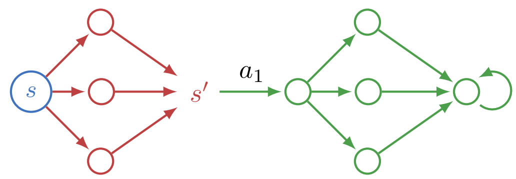


Figure F.8: Informally, lemma F.75 shows that a bottleneck at s' "factorizes" $\mathcal{F}(s \mid \pi(s') = a_1)$ into combinations of "what happens before acting at the bottleneck" and "what happens after acting at the bottleneck." In this rewardless MDP, $s \to s' \xrightarrow{a_1} \textsc{Reach}\left(s', a_1\right)$. Therefore, $\left|\mathcal{F}(s)\right| = 9$: each of the three red "prefix" visit distribution functions (induced before reaching s') can combine with the three green "suffix" visit distribution functions (induced after reaching s').

Definition F.74 (State-space bottleneck)**.** Starting from s, state s' is a *bottleneck* for $X \subseteq \mathcal{S}$ via action a when state s can reach the states of X with positive probability, but only by taking actions equivalent to a at state s'. We write this as $s \to s' \xrightarrow{a} X$.

Lemma F.75 ($\mathcal{F}(s)$ "factorizes" across bottlenecks)**.** *Suppose* $\forall 1 \leq i \leq k : s \to s' \xrightarrow{a_i} \textsc{Reach}\left(s', a_i\right)$. *Then let* $\mathbb{1}_{reach} := \sum_{i=1}^k \sum_{s_j \in \textsc{Reach}(s', a_i)} \mathbf{e}_{s_j}$ *and* $\mathbb{1}_{rest} := \mathbf{1} - \mathbb{1}_{reach}$ *(where* $\mathbf{1} \in \mathbb{R}^{|\mathcal{S}|}$ *is the all-ones vector). Let* $F^b_{rest} := \left\{\mathbf{f}^{\pi,s} \odot \mathbb{1}_{rest} \mid \pi \in \Pi\right\}$ *(with* $\odot$ *the Hadamard product),* $F^b_{rest,a_i} := \left\{\mathbf{f}^{\pi,s} \odot \mathbb{1}_{rest} \mid \pi \in \Pi : \pi(s') \equiv_{s'} a_i\right\}$*, and* $F^b_{a_i} := \left\{\mathbb{E}_{s_{a_i} \sim T(s', a_i)}\left[\mathbf{f}^{\pi, s_{a_i}}\right] \mid \pi \in \Pi\right\}$*. In the following,* γ *is left variable on* $[0,1)$*.*

$$\mathcal{F}(s \mid \pi(s') = a_i) = \left\{\mathbf{f}_{rest}(\gamma) + \left(1 - (1-\gamma)\left\|\mathbf{f}_{rest}(\gamma)\right\|_1\right)\mathbf{f}_{a_i}(\gamma) \mid \mathbf{f}_{rest} \in F^b_{rest,a_1}, \mathbf{f}_{a_i} \in F^b_{a_i}\right\}. \tag{F.54}$$

Proof. Keep in mind that in order to detail how state-space bottlenecks affect the structure of visit distribution functions, we hold γ variable on $[0,1)$.

$$\mathcal{F}(s \mid \pi(s') = a_i) \tag{F.55}$$
$$:= \left\{\mathbf{f}^{\pi,s} \mid \pi \in \Pi : \pi(s') = a_i\right\} \tag{F.56}$$
$$= \left\{\mathbf{f}^{\pi,s} \mid \pi \in \Pi : \pi(s') \equiv_{s'} a_i\right\} \tag{F.57}$$
$$= \left\{\mathbf{f}^{\pi,s}(\gamma) \odot (\mathbb{1}_{\text{rest}} + \mathbb{1}_{\text{reach}}) \mid \pi \in \Pi : \pi(s') \equiv_{s'} a_i\right\} \tag{F.58}$$

$$= \left\{ \mathbf{f}^{\pi,s}(\gamma) \odot \mathbb{1}_{\text{rest}} + \left(1 - (1-\gamma) \left\| \mathbf{f}^{\pi,s}(\gamma) \odot \mathbb{1}_{\text{rest}} \right\|_1 \right) \mathbf{f}^{\pi}_{a_i}(\gamma) \mid \pi \in \Pi : \pi(s') \equiv_{s'} a_i \right\} \tag{F.59}$$

$$= \left\{ \mathbf{f}_{\text{rest}}(\gamma) + \left(1 - (1-\gamma) \left\| \mathbf{f}_{\text{rest}}(\gamma) \right\|_1 \right) \mathbf{f}_{a_i}(\gamma) \mid \mathbf{f}_{\text{rest}} \in F^b_{\text{rest},a_i}, \mathbf{f}_{a_i} \in F^b_{a_i} \right\} \tag{F.60}$$

$$= \left\{ \mathbf{f}_{\text{rest}}(\gamma) + \left(1 - (1-\gamma) \left\| \mathbf{f}_{\text{rest}}(\gamma) \right\|_1 \right) \mathbf{f}_{a_i}(\gamma) \mid \mathbf{f}_{\text{rest}} \in F^b_{\text{rest},a_1}, \mathbf{f}_{a_i} \in F^b_{a_i} \right\}. \tag{F.61}$$

Equation (F.57) holds because by the definition of action equivalence (definition 5.23), equivalent actions induce identical state visit distribution functions $\mathbf{f}^{\pi,s}$. Equation (F.58) follows since $\mathbb{1}_{\text{rest}} + \mathbb{1}_{\text{reach}} = \mathbf{1}$. To see that eq. (F.59) follows, consider first that once the agent takes an action equivalent to a_i at state s', it induces state visit distribution $\mathbf{f}^{\pi}_{a_i}(\gamma) \in F^b_{a_i}(\gamma)$ (by the definition of $F^b_{a_i}$). Since the bottleneck assumption ensures that no other components of $\mathbf{f}^{\pi,s}$ visit the states of $\cup_{i=1}^k \textsc{Reach}\left(s', a_i\right)$, $\mathbf{f}^{\pi,s}(\gamma) \odot \mathbb{1}_{\text{reach}} = c(\gamma)\mathbf{f}^{\pi}_{a_i}(\gamma)$ for some scaling function $c \in \mathbb{R}^{[0,1)}$.

$$\left\| \mathbf{f}^{\pi,s}(\gamma) \right\|_1 = \frac{1}{1-\gamma} \tag{F.62}$$

$$\left\| \mathbf{f}^{\pi,s}(\gamma) \odot \mathbb{1}_{\text{rest}} + c(\gamma)\mathbf{f}^{\pi}_{a_i}(\gamma) \right\|_1 = \frac{1}{1-\gamma} \tag{F.63}$$

$$\left\| \mathbf{f}^{\pi,s}(\gamma) \odot \mathbb{1}_{\text{rest}} \right\|_1 + \left\| c(\gamma)\mathbf{f}^{\pi}_{a_i}(\gamma) \right\|_1 = \frac{1}{1-\gamma} \tag{F.64}$$

$$\left| c(\gamma) \right| = \left\| \mathbf{f}^{\pi}_{a_i}(\gamma) \right\|_1^{-1} \left(\frac{1}{1-\gamma} - \left\| \mathbf{f}^{\pi,s}(\gamma) \odot \mathbb{1}_{\text{rest}} \right\|_1 \right) \tag{F.65}$$

$$\left| c(\gamma) \right| = (1-\gamma) \left(\frac{1}{1-\gamma} - \left\| \mathbf{f}^{\pi,s}(\gamma) \odot \mathbb{1}_{\text{rest}} \right\|_1 \right) \tag{F.66}$$

$$c(\gamma) = 1 - (1-\gamma) \left\| \mathbf{f}^{\pi,s}(\gamma) \odot \mathbb{1}_{\text{rest}} \right\|_1. \tag{F.67}$$

Equation (F.62) and eq. (F.66) follow by proposition D.8 item 2. Equation (F.64) follows because $\mathbf{f}_{\text{rest}}(\gamma), \mathbf{f}^{\pi}_{a_i}(\gamma) \succeq 0$. In eq. (F.67), $\left|c(\gamma)\right| = c(\gamma)$ must hold because if $c(\gamma)$ were negative, $\mathbf{f}^{\pi,s}(\gamma)$ would contain negative entries (which is impossible by proposition D.8 item 1). Equation (F.67) demonstrates that eq. (F.59) holds.

To see that eq. (F.60) holds, consider that policy choices on $\textsc{Reach}\left(s', a_i\right)$ cannot affect the visit distribution function $\mathbf{f}^{\pi,s}(\gamma) \odot \mathbb{1}_{\text{rest}}$. This is because the definition of $\textsc{Reach}\left(s', a_i\right)$

ensures that once the agent reaches $\textsc{Reach}\left(s', a_i\right)$, it never leaves. Therefore, for all $\pi, \pi' \in \Pi$ which only disagree on $s_j \in \textsc{Reach}\left(s', a_i\right)$, $\mathbf{f}^{\pi,s}(\gamma) \odot \mathbb{1}_{\text{rest}} = \mathbf{f}^{\pi',s}(\gamma) \odot \mathbb{1}_{\text{rest}}$. This implies that any $\mathbf{f}^{\pi} \in F^b_{a_i}$ is compatible with any $\mathbf{f}^{\pi'} \in F^b_{\text{rest}}$. Therefore, eq. (F.60) holds.

Lastly, we show that $F^b_{\text{rest},a_i} = F^b_{\text{rest},a_1}$, which shows that eq. (F.61) holds. In the following, let $d(\gamma) \coloneqq 1 - (1-\gamma)\left\lVert \mathbf{f}^{\pi,s}(\gamma) \odot \mathbb{1}_{\text{rest}} \right\rVert_1$.

$$F^b_{\text{rest},a_i} \tag{F.68}$$
$$\coloneqq \left\{\mathbf{f}^{\pi,s} \odot \mathbb{1}_{\text{rest}} \mid \pi \in \Pi : \pi(s') \equiv_{s'} a_i\right\} \tag{F.69}$$
$$= \left\{\left(\mathbf{f}^{\pi,s}(\gamma) \odot \mathbb{1}_{\text{rest}} + d(\gamma)\mathbf{f}^{\pi}_{a_i}(\gamma)\right) \odot \mathbb{1}_{\text{rest}} \mid \pi \in \Pi : \pi(s') \equiv_{s'} a_i\right\} \tag{F.70}$$
$$= \left\{\left(\mathbf{f}^{\pi,s}(\gamma) \odot \mathbb{1}_{\text{rest}} + d(\gamma)\mathbf{f}^{\pi}_{a_1}(\gamma)\right) \odot \mathbb{1}_{\text{rest}} \mid \pi \in \Pi : \pi(s') \equiv_{s'} a_i\right\} \tag{F.71}$$
$$= \left\{\left(\mathbf{f}^{\pi,s}(\gamma) \odot \mathbb{1}_{\text{rest}} + d(\gamma)\mathbf{f}^{\pi}_{a_1}(\gamma)\right) \odot \mathbb{1}_{\text{rest}} \mid \pi \in \Pi : \pi(s') \equiv_{s'} a_1\right\} \tag{F.72}$$
$$= \left\{\mathbf{f}^{\pi,s} \odot \mathbb{1}_{\text{rest}} \mid \pi \in \Pi : \pi(s') \equiv_{s'} a_1\right\} \tag{F.73}$$
$$\eqqcolon F^b_{\text{rest},a_1}. \tag{F.74}$$

Equation (F.70) and eq. (F.73) follow by eq. (F.59) above. Equation (F.71) follows because by the definition of $\mathbb{1}_{\text{rest}}$ and of $\mathbf{f}^{\pi}_{a_i} \in F^b_{a_i}, \mathbf{f}^{\pi}_{a_1} \in F^b_{a_1}$, we have $d(\gamma)\mathbf{f}^{\pi}_{a_i}(\gamma) \odot \mathbb{1}_{\text{rest}} = d(\gamma)\mathbf{f}^{\pi}_{a_1}(\gamma) \odot \mathbb{1}_{\text{rest}} = \mathbf{0}$ (the all-zeros vector in $\mathbb{R}^{|\mathcal{S}|}$). Equation (F.72) follows because the definition of $\textsc{Reach}\left(s', a_i\right)$ ensures that once the agent takes actions equivalent to a_i at s', it only visits states $s_j \in \cup_{i'=1}^{k} \textsc{Reach}\left(s', a_i'\right)$. The same is true for a_1. Therefore, by the definition of $\mathbb{1}_{\text{rest}}$, $\mathbf{f}^{\pi,s}(\gamma) \odot \mathbb{1}_{\text{rest}}$ is invariant to the choice of action a_i versus a_1. We conclude that eq. (F.61) holds, which proves the desired equality. □

Definition F.76 (Non-dominated single-state $\mathcal{F}$ restriction)**.** $\mathcal{F}_{\text{nd}}(s \mid \pi(s') = a) \coloneqq \mathcal{F}(s \mid \pi(s') = a) \cap \mathcal{F}_{\text{nd}}(s)$.

Conjecture F.77 ($\mathcal{F}_{\text{nd}}$ factorizes across state bottlenecks)**.** In the following, γ is left variable on $[0, 1)$. Suppose that starting from s, state s' is a bottleneck for $\textsc{Reach}\left(s', a_i\right)$ via actions $\{a_i\}$, for $1 \leq i \leq k$. Then let $\mathbb{1}_{\text{reach}} \coloneqq \sum_{i=1}^{k} \sum_{s_j \in \textsc{Reach}(s', a_i)} \mathbf{e}_{s_j}$ and $\mathbb{1}_{\text{rest}} \coloneqq \mathbf{1} - \mathbb{1}_{\text{reach}}$ (where $\mathbf{1} \in \mathbb{R}^{|\mathcal{S}|}$ is the all-ones vector). Let $F^b_{\text{rest}} \coloneqq \left\{\mathbf{f}^{\pi,s} \odot \mathbb{1}_{\text{rest}} \mid \pi \in \Pi\right\}$ (with

$\odot$ the Hadamard product), $F^b_{\text{rest},a_i} \coloneqq \left\{\mathbf{f}^{\pi,s} \odot \mathbb{1}_{\text{rest}} \mid \pi \in \Pi : \pi(s') \equiv_{s'} a_i\right\}$, and $F^b_{a_i} \coloneqq \left\{\mathbb{E}_{s_{a_i} \sim T(s',a_i)}\left[\mathbf{f}^{\pi,s_{a_i}}\right] \mid \pi \in \Pi\right\}$.

1. $\mathcal{F}(s) = \left(F^b_{\text{rest}} \setminus F^b_{\text{rest},a_1}\right) \cup \left(\cup_{i=1}^k \mathcal{F}(s \mid \pi(s') = a_i)\right)$.
2. $\mathcal{F}_{\text{nd}}(s \mid \pi(s') = a_i) \subseteq$ $\left\{\mathbf{f}_{\text{rest}}(\gamma) + \left(1 - (1-\gamma)\left\|\mathbf{f}_{\text{rest}}(\gamma)\right\|_1\right)\mathbf{f}_{a_i}(\gamma) \mid \mathbf{f}_{a_i} \in \text{ND}\left(F^b_{a_i}\right), \mathbf{f}_{\text{rest}} \in \text{ND}\left(F^b_{\text{rest},a_1}\right)\right\}$.
3. $\mathcal{F}_{\text{nd}}(s \mid \pi(s') = a_i) \subseteq$ $\left\{\mathbf{f}_{\text{rest}}(\gamma) + \left(1 - (1-\gamma)\left\|\mathbf{f}_{\text{rest}}(\gamma)\right\|_1\right)\mathbf{f}_{a_i}(\gamma) \mid \mathbf{f}_{a_i} \in \text{ND}\left(F^b_{a_i}\right), \mathbf{f}_{\text{rest}} \in \text{ND}\left(F^b_{\text{rest},a_1}\right)\right\}$.
4. $\mathcal{F}_{\text{nd}}(s \mid \pi(s') = a_i) =$ $\left\{\mathbf{f}_{\text{rest}}(\gamma) + \left(1 - (1-\gamma)\left\|\mathbf{f}_{\text{rest}}(\gamma)\right\|_1\right)\mathbf{f}_{a_i} \mid \mathbf{f}_{a_i} \in \text{ND}\left(F^b_{a_i}\right), \mathbf{f}_{\text{rest}} \in \text{ND}\left(F^b_{\text{rest},a_1}\right)\right\}$.
5. $\mathcal{F}_{\text{nd}}(s) = \text{ND}\left(F^b_{\text{rest}} \setminus F^b_{\text{rest},a_1}\right) \cup$ $\left\{\mathbf{f}_{\text{rest}}(\gamma) + \left(1 - (1-\gamma)\left\|\mathbf{f}_{\text{rest}}(\gamma)\right\|_1\right)\mathbf{f}_{a_i} \mid \mathbf{f}_{a_i} \in \bigcup_{i=1}^k \text{ND}\left(F^b_{a_i}\right), \mathbf{f}_{\text{rest}} \in \text{ND}\left(F^b_{\text{rest},a_1}\right)\right\}$.

Partial proof sketch. **Item 1.**

$$
\begin{aligned}
&\mathcal{F}(s)\\
&\quad= \bigcup_{a\in\mathcal{A}} \mathcal{F}_{\text{nd}}(s \mid \pi(s') = a) &\text{(F.75)}\\
&\quad= \left(\bigcup_{\substack{a\in\mathcal{A}:\\ \forall i: a \not\equiv_{s'} a_i}} \mathcal{F}(s \mid \pi(s') = a)\right) \cup \left(\bigcup_{i=1}^k \mathcal{F}(s \mid \pi(s') = a_i)\right) &\text{(F.76)}\\
&\quad= \left(\bigcup_{\substack{a\in\mathcal{A}:\\ \forall i: a \not\equiv_{s'} a_i}} \left\{\mathbf{f}^{\pi,s} \mid \pi \in \Pi : \pi(s') = a\right\}\right) \cup \left(\bigcup_{i=1}^k \mathcal{F}(s \mid \pi(s') = a_i)\right) &\text{(F.77)}\\
&\quad= \left(\bigcup_{\substack{a\in\mathcal{A}:\\ \forall i: a \not\equiv_{s'} a_i}} \left\{\mathbf{f}^{\pi,s} \odot \mathbb{1}_{\text{rest}} \mid \pi \in \Pi : \pi(s') = a\right\}\right) \cup \left(\bigcup_{i=1}^k \mathcal{F}(s \mid \pi(s') = a_i)\right) &\text{(F.78)}
\end{aligned}
$$

$$= \left(F^b_{\text{rest}} \setminus \left(\cup_{i=1}^k F^b_{\text{rest},a_i}\right)\right) \cup \left(\bigcup_{i=1}^k \mathcal{F}(s \mid \pi(s') = a_i)\right) \tag{F.79}$$

$$= \left(F^b_{\text{rest}} \setminus F^b_{\text{rest},a_1}\right) \cup \left(\bigcup_{i=1}^k \mathcal{F}(s \mid \pi(s') = a_i)\right). \tag{F.80}$$

By the bottleneck assumption, s can only reach the states of $\cup_{i=1}^k \textsc{Reach}\left(s', a_i\right)$ by taking actions equivalent to some a_i at state s'. Therefore, eq. (F.78) holds by the definition of $\mathbb{1}_{\text{rest}}$. Equation (F.79) holds by the definition of F^b_{rest} and F^b_{rest,a_i}. Equation (F.80) because eq. (F.74) showed that $\forall i : F^b_{\text{rest},a_i} = F^b_{\text{rest},a_1}$. We have now shown item 1.

Item 2. Let $\mathbf{f} \in \mathcal{F}_{\text{nd}}(s \mid \pi(s') = a_i)$ be strictly optimal for $\mathbf{r}$ at discount rate $\gamma^* \in (0,1)$. By lemma F.75, for some $\mathbf{f}_{\text{rest}} \in F^b_{\text{rest},a_1}, \mathbf{f}_{a_i} \in F^b_{a_i}$, $\mathbf{f} = \mathbf{f}_{\text{rest}}(\gamma) + \left(1 - (1-\gamma)\left\|\mathbf{f}_{\text{rest}}(\gamma)\right\|_1\right)\mathbf{f}_{a_i}(\gamma)$. Suppose that $\mathbf{f}_{a_i} \not\in \text{ND}\left(F^b_{a_i}\right)$. Then there exists another $\mathbf{f}'_{a_i} \in F^b_{a_i}$ for which $\mathbf{f}_{a_i}(\gamma^*)^\top \mathbf{r} \geq \mathbf{f}'_{a_i}(\gamma^*)^\top \mathbf{r}$.

Consider $\mathbf{f}' := \mathbf{f}_{\text{rest}}(\gamma) + \left(1 - (1-\gamma)\left\|\mathbf{f}_{\text{rest}}(\gamma)\right\|_1\right)\mathbf{f}'_{a_i}(\gamma)$. By item 1, $\mathbf{f}' \in \mathcal{F}(s)$. However, since $\mathbf{f}_{a_i}(\gamma^*)^\top \mathbf{r} \geq \mathbf{f}'_{a_i}(\gamma^*)^\top \mathbf{r}$, $\mathbf{f}'(\gamma^*)^\top \mathbf{r} \geq \mathbf{f}(\gamma^*)^\top \mathbf{r}$. This contradicts the assumed strict optimality of $\mathbf{f}$. Therefore, $\mathbf{f}_{a_i} \in \text{ND}\left(F^b_{a_i}\right)$ (in particular, $\mathbf{f}_{a_i}(\gamma^*) \in \text{ND}\left(F^b_{a_i}(\gamma^*)\right)$). ☐

Conjecture F.78 (Action-restricted visit distribution function similarity requires action similarity)**.** If s can reach s' with positive probability and $\mathcal{F}_{\text{nd}}(s \mid \pi(s') = a')$ is similar to a subset of $\mathcal{F}(s \mid \pi(s') = a)$ via state permutation ϕ, then $\mathbf{P}_\phi T(s', a') = T(s', a)$.

F.1.9 Properties of optimality support

Definition F.79 (Optimality support)**.** Let $A, B \subsetneq \mathbb{R}^{|\mathcal{S}|}$ be finite.

$$\text{supp}\left(A \geq B\right) := \left\{\mathbf{r} \in \mathbb{R}^{|\mathcal{S}|} \mid \max_{\mathbf{a} \in A} \mathbf{a}^\top \mathbf{r} \geq \max_{\mathbf{b} \in B} \mathbf{b}^\top \mathbf{r}\right\}. \tag{F.81}$$

We sometimes abuse notation by replacing the set A with a vector $\mathbf{a}$, as in: $\text{supp}\left(\mathbf{a} \geq B\right)$.

$\text{supp}\left(\mathbf{f}(\gamma) \geq \mathcal{F}(s,\gamma)\right)$ represents the set of reward functions for which $\mathbf{f}(\gamma)$ is optimal at state s. $\operatorname{supp}\left(\mathbf{f}(\gamma) \geq \mathcal{F}(s, \gamma)\right)$ can be calculated by solving the relevant system of $\left|\mathcal{F}(s)\right| - 1$ inequalities.[1]

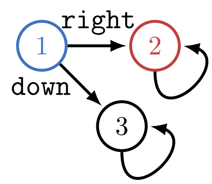


Figure F.9: A simple environment where it's easy to derive which reward functions make a trajectory optimal.

For example, consider fig. F.9.

$$\mathbf{f}^{\pi_{\texttt{right}}\top}\mathbf{r} \geq \mathbf{f}^{\pi_{\texttt{down}}\top}\mathbf{r}$$
$$R(s_1) + \frac{\gamma R(s_2)}{1-\gamma} \geq R(s_1) + \frac{\gamma R(s_3)}{1-\gamma},$$

so $R(s_2) \geq R(s_3)$.

Definition F.80 (Topological boundary)**.** $\operatorname{bd}(X)$ is the topological boundary of set X, equal to X's closure minus its interior ($\operatorname{Int}(X)$).

Remark. Unless otherwise stated, assume $\mathbb{R}^{|\mathcal{S}|}$ is endowed with the standard topology.

Lemma F.81 (A topological lemma)**.** *Let $X \subseteq \mathbb{R}^{|\mathcal{S}|}$ and suppose S is such that $S \cap \operatorname{Int}(X)$ is convex and has no interior. If $s \in S \cap \operatorname{Int}(X)$, then $\exists \theta \in (0,1), x_1, x_2 \in \operatorname{Int}(X) \setminus S$: $s = \theta x_1 + (1-\theta)x_2$.*

Proof. Since S has no interior, its restriction to $\operatorname{Int}(X)$ must equal a convex subset of some $(|\mathcal{S}|-1)$-dimensional hyperplane intersect $\operatorname{Int}(X)$. Let $\mathbf{x}$ be a unit-length vector orthogonal to this hyperplane. Since $R \in \operatorname{Int}(X)$, there exists $\epsilon > 0$ small enough such that $R + \epsilon\mathbf{x}, R - \epsilon\mathbf{x} \in \operatorname{Int}(X)$. Because $\mathbf{x}$ is perpendicular to the hyperplane, neither of these points belong to S. $R = .5(R + \epsilon\mathbf{x}) + .5(R - \epsilon\mathbf{x})$. □

[1] $V^*_{\mathcal{D}_\text{bound}}(s, \gamma)$ can sometimes be computed analytically. The Power and optimality probability in small deterministic MDPs can be computed using Mathematica code at https://github.com/loganriggs/Optimal-Policies-Tend-To-Seek-Power.

Lemma F.82 (Topological properties of optimality support)**.** *Let* $\mathbf{f}, \mathbf{f}' \in \mathcal{F}(s)$ *and* $\gamma \in (0,1)$.

1. $\operatorname{bd}(\operatorname{supp}\left(\mathbf{f}(\gamma) \geq \mathcal{F}(s,\gamma)\right)) \subseteq \operatorname{supp}\left(\mathbf{f}(\gamma) \geq \mathcal{F}(s,\gamma)\right)$, *with equality iff* $\mathbf{f}$ *is dominated.*
2. *If* $\mathbf{f} \neq \mathbf{f}'$ *are both optimal for* R *at discount rate* γ, *then* $R \in \operatorname{bd}(\operatorname{supp}\left(\mathbf{f}(\gamma) \geq \mathcal{F}(s,\gamma)\right)) \cap \operatorname{bd}(\operatorname{supp}\left(\mathbf{f}'(\gamma) \geq \mathcal{F}(s,\gamma)\right))$.

Proof. Item 1: $\operatorname{bd}(\operatorname{supp}\left(\mathbf{f}(\gamma) \geq \mathcal{F}(s,\gamma)\right)) \subseteq \operatorname{supp}\left(\mathbf{f}(\gamma) \geq \mathcal{F}(s,\gamma)\right)$ because optimality support is closed by lemma F.42. If equality holds, $\operatorname{supp}\left(\mathbf{f}(\gamma) \geq \mathcal{F}(s,\gamma)\right)$ has no interior, and $\mathbf{f}$ is therefore dominated by proposition F.151. Similarly, if $\mathbf{f}$ is dominated, then proposition F.151 dictates that $\operatorname{supp}\left(\mathbf{f}(\gamma) \geq \mathcal{F}(s,\gamma)\right)$ has no interior and thus equals its boundary.

Item 2: by lemma F.106, almost no reward functions have multiple optimal visit distributions, and so $\mathbf{f}$ and $\mathbf{f}'$ cannot both be optimal in an open neighborhood of R. Thus, R must lie on at least one boundary: $R \in \operatorname{bd}(\operatorname{supp}\left(\mathbf{f}(\gamma) \geq \mathcal{F}(s,\gamma)\right)) \cup \operatorname{bd}(\operatorname{supp}\left(\mathbf{f}'(\gamma) \geq \mathcal{F}(s,\gamma)\right))$.

Suppose $R \in \operatorname{Int}(\operatorname{supp}\left(\mathbf{f}(\gamma) \geq \mathcal{F}(s,\gamma)\right))$. Then $\mathbf{f}$ is optimal in an open neighborhood N around R and $R \in \operatorname{bd}(\operatorname{supp}\left(\mathbf{f}'(\gamma) \geq \mathcal{F}(s,\gamma)\right))$ by the above reasoning. By lemma F.106, $\mathbf{f}$ must be uniquely (and therefore strictly) optimal for almost all reward functions in N. Furthermore, $\operatorname{supp}\left(\mathbf{f}'(\gamma) \geq \mathcal{F}(s,\gamma)\right)$ has no interior and is convex (by lemma F.42), so by lemma F.81, $R \in \operatorname{supp}\left(\mathbf{f}'(\gamma) \geq \mathcal{F}(s,\gamma)\right) \cap \operatorname{Int}(\operatorname{supp}\left(\mathbf{f}(\gamma) \geq \mathcal{F}(s,\gamma)\right))$ can be written as the convex combination of reward functions for which $\mathbf{f}$ is optimal but $\mathbf{f}'$ is not. But corollary F.35 shows that convex combination of reward functions preserves optimal policy sets, and so $\mathbf{f}'$ cannot be optimal for R, a contradiction. So $R \notin \operatorname{Int}(\operatorname{supp}\left(\mathbf{f}(\gamma) \geq \mathcal{F}(s,\gamma)\right))$, and so $R \in \operatorname{bd}(\operatorname{supp}\left(\mathbf{f}(\gamma) \geq \mathcal{F}(s,\gamma)\right)) \cap \operatorname{bd}(\operatorname{supp}\left(\mathbf{f}'(\gamma) \geq \mathcal{F}(s,\gamma)\right))$. ☐

Theorem F.83 (If a dominated visit distribution is optimal, so are at least two non-dominated visit distributions)**.** *Suppose* $\mathbf{f}_d \in \mathcal{F}(s) \setminus \mathcal{F}_{\text{nd}}(s)$ *is optimal for reward function* R *at discount rate* $\gamma \in (0,1)$. *Then there exist distinct* $\mathbf{f}_1, \mathbf{f}_2 \in \mathcal{F}_{\text{nd}}(s)$ *which are also optimal for* R *at* γ.

Proof. If $\mathbf{f}_d$ were optimal for R at γ, then some non-dominated $\mathbf{f}_1$ must also be optimal by corollary F.65. $R \in \operatorname{bd}(\operatorname{supp}(\mathbf{f}_d)) \cap \operatorname{bd}(\operatorname{supp}(\mathbf{f}_1))$ by lemma F.82(2).

$\text{supp}\,(\mathbf{f}_d) \cup \text{supp}\,(\mathbf{f}_1) = \text{supp}\,(\mathbf{f}_1) \subsetneq \mathbb{R}^{|\mathcal{S}|}$ by corollary F.69. But

$$\bigcup_{\mathbf{f}\in\mathcal{F}(s)} \text{supp}\left(\mathbf{f}(\gamma) \geq \mathcal{F}(s,\gamma)\right) = \mathbb{R}^{|\mathcal{S}|}.$$

Therefore, since all $\text{supp}\,(\mathbf{f})$ are closed and convex (lemma F.42), there must be at least one more $\mathbf{f}_2$ such that $\text{supp}\,(\mathbf{f}_2)$ has non-empty interior (*i.e.* $\mathbf{f}_2 \in \mathcal{F}_{\text{nd}}(s)$ by proposition F.151) and $\mathbf{f}_2$ is optimal for R. Then $R \in \text{bd}(\text{supp}\,(\mathbf{f}_d)) \cap \text{bd}(\text{supp}\,(\mathbf{f}_1)) \cap \text{bd}(\text{supp}\,(\mathbf{f}_2))$. □

Conjecture F.84 (Geometry of dominated optimality support)**.** If $\mathbf{f}_d \in \mathcal{F}(s) \setminus \mathcal{F}_{\text{nd}}(s)$ and $\mathbf{f} \in \mathcal{F}_{\text{nd}}(s)$ are both optimal for R at $\gamma \in (0,1)$, then $\text{supp}\,(\mathbf{f}_d) \subseteq \text{bd}(\text{supp}\,(\mathbf{f}))$.

F.1.10 How geodesics affect visit distribution optimality

Definition F.85 (Geodesic trajectory)**.** In a directed graph, a path between two vertices is *geodesic* when it is a shortest path. In deterministic environments, a state trajectory $(s_0, s_1, \ldots)$ is geodesic when, for all $i \leq j$, the trajectory traces a geodesic path between s_i and s_j.

Intuitively, non-geodesic trajectories take "detours" (see fig. F.10).

Remark. We refer to a policy or trajectory as "dominated" when the corresponding visit distribution function is dominated.

Proposition F.86 (In deterministic MDPs, geodesic trajectories induce non-dominated visit distributions)**.** *Suppose the* MDP *is deterministic and that starting from state s_0, policy π induces geodesic trajectory $\tau^\pi = (s_0, s_1, \ldots)$. $\mathbf{f}^\pi \in \mathcal{F}_{\text{nd}}(s_0)$.*

Proof. Since the MDP is deterministic and the state space is finite, τ^π visits a finite number of states $s_0, \ldots, s_k$. Consider reward function R such that $\forall i \leq k : R(s_i) := \frac{i+1}{|\mathcal{S}|}$; for all states s' not visited by τ^π, $R(s') := 0$. Because τ^π is geodesic, $\mathbf{f}^\pi$ is strictly greedily optimal for R. Since optimal value changes continuously with γ, $\mathbf{f}^\pi$ is also strictly optimal for some $\gamma \approx 0$. Then $\mathbf{f}^\pi \in \mathcal{F}_{\text{nd}}(s_0)$. □

Corollary F.87 (In deterministic MDPs, dominated trajectories are not geodesic)**.**

Conversely, fig. F.10 shows that non-dominated visit distribution functions need not be geodesic.

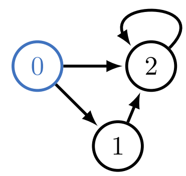


Figure F.10: $\left|\mathcal{F}_{\mathrm{nd}}(s_0)\right| = 2$, even though one of the visit distribution functions is not geodesic because it induces state trajectory $(s_0, s_1, s_2, s_2, \ldots)$.

Conjecture F.88 (Geodesics in stochastic environments)**.** Proposition F.86 can be generalized in some form to stochastic MDPs.

F.1.11 Number of visit distribution functions

Since $\mathcal{F}(s)$ only contains the visit distribution functions induced by *deterministic stationary* policies, fig. F.11 shows that s_3 being able to reach s_1 doesn't imply that $\left|\mathcal{F}(s_3)\right| \geq \left|\mathcal{F}(s_1)\right|$.

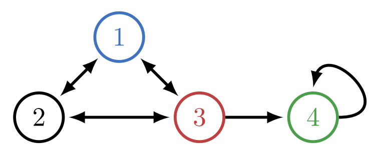


Figure F.11: $\left|\mathcal{F}(s_1)\right| = 8 > 7 = \left|\mathcal{F}(s_3)\right|$. In particular, considering only deterministic stationary policies, s_1 can reach s_4 via two different trajectories: $(s_1, s_2, s_3, s_4, s_4, \ldots)$ and $(s_1, s_3, s_4, s_4, \ldots)$. However, s_3 can only reach s_4 in one way.

Lemma F.89 (Maximum number of visit distribution functions (deterministic))**.** *Suppose the* MDP *is deterministic. For all s, $\left|\mathcal{F}(s)\right| \leq \sum_{j=1}^{|\mathcal{S}|} j \frac{\left(|\mathcal{S}|-1\right)!}{\left(|\mathcal{S}|-j\right)!}$, with equality iff $\forall s' : Ch\left(s'\right) = \mathcal{S}$.*

Proof. First assume $\forall s' : Ch\left(s'\right) = \mathcal{S}$. We count how many trajectories can be induced by deterministic stationary policies, starting from state s. Any trajectory visits at least 1 and at most $|\mathcal{S}|$ states. Since order matters, a j-state trajectory can be chosen in

$j!\binom{|\mathcal{S}|}{j} = \frac{|\mathcal{S}|!}{(|\mathcal{S}|-j)!}$ ways. However, we must always begin the trajectory at s, so divide by $|\mathcal{S}|$ possible starting states.

Since $\mathcal{S}$ is finite, the policy is stationary and deterministic, and the MDP is deterministic, each trajectory must have a cycle by the pigeonhole principle. For any list of j states, there are j locations at which the cycle can begin. Then there are $\sum_{j=1}^{|\mathcal{S}|} j \frac{(|\mathcal{S}|-1)!}{(|\mathcal{S}|-j)!}$ viable trajectories starting from state s. Two different trajectories induce different visitation distribution power series on γ (definition 5.3), and so different trajectories correspond to different visitation distribution functions.

Given $\forall s' : Ch\left(s'\right) = \mathcal{S}$, we showed that $\left|\mathcal{F}(s)\right| = \sum_{j=1}^{|\mathcal{S}|} j \frac{(|\mathcal{S}|-1)!}{(|\mathcal{S}|-j)!}$. If $\exists s', s'' : s'' \notin Ch\left(s'\right)$, this rules out trajectories containing an $s' \to s''$ transition. Therefore, we have $\left|\mathcal{F}(s)\right| \leq \sum_{j=1}^{|\mathcal{S}|} j \frac{(|\mathcal{S}|-1)!}{(|\mathcal{S}|-j)!}$ in general, with equality iff $\forall s' : Ch\left(s'\right) = \mathcal{S}$. $\square$

Remark. When $\forall s' : Ch\left(s'\right) = \mathcal{S}$, the non-dominated trajectories are the ones that immediately navigate to a state and stay there: $\forall s : \left|\mathcal{F}_{\text{nd}}(s)\right| = |\mathcal{S}|$. However, fig. F.12 demonstrates an MDP containing a state s for which $\left|\mathcal{F}_{\text{nd}}(s)\right| > |\mathcal{S}|$.

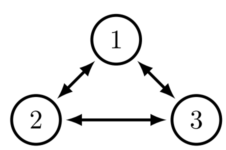


Figure F.12: $\forall s : \left|\mathcal{F}_{\text{nd}}(s)\right| = 4 > 3 = |\mathcal{S}|$.

Puterman [68] notes that often, multiple policies map to the same visit distribution function. This is always true in deterministic environments if there is more than one possible policy.

Theorem F.90 (In deterministic environments, $\pi \mapsto \mathbf{f}_s^\pi$ is non-injective unless $|\mathcal{A}| = 1$)**.** *Suppose the environment is deterministic.* $\exists s : \left|\mathcal{F}(s)\right| = |\Pi|$ *iff* $|\mathcal{A}| = 1$.

Proof. Suppose $|\mathcal{A}| > 1$. If $|\mathcal{S}| < |\mathcal{A}|$, then by determinism, at least two distinct actions must be equivalent. By the definition of action equivalence (definition 5.23), policies taking equivalent actions at all states induce the same visit distribution functions, and so then $\left|\mathcal{F}(s)\right| < |\Pi|$.

Suppose $|\mathcal{S}| \geq |\mathcal{A}|$; then $|\mathcal{S}| > 1$. If a cycle can be induced before visiting all states, the premise is contradicted by modifying the π in question in any unvisited states. So starting from s, let π induce the visitation of all states. Since $2 \leq |\mathcal{A}|, |\mathcal{S}|$, $2 < 2|\mathcal{S}| \leq |\mathcal{A}||\mathcal{S}|$ states can be reached over the course of π's state trajectory.

Therefore, by the time the trajectory has traversed $1 \leq \lceil \frac{|\mathcal{S}|}{2} \rceil < |\mathcal{S}|$ states, there must exist an action returning to a state which has already been visited. Modify π to take that action, and a cycle is formed before all states are visited. Since $2 \leq |\mathcal{A}|$ and the induced visit distribution function is not affected by the action taken at the unvisited state, there are at least two policies which induce the same visit distribution function. So $\left|\mathcal{F}(s)\right| < |\Pi|$.

Suppose $|\mathcal{A}| = 1$. By the definition of $\mathcal{F}(s)$ (definition 5.3), $\forall s : \left|\mathcal{F}(s)\right| \leq |\Pi| = |\mathcal{A}|^{|\mathcal{S}|} = 1$. $\forall s : 1 \leq \left|\mathcal{F}(s)\right|$ by lemma F.1. □

Theorem F.90 does not hold for stochastic environments.

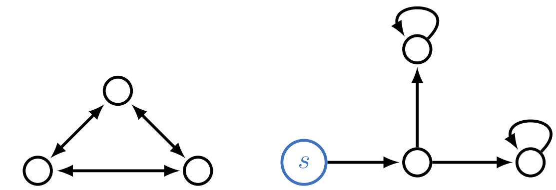
Conjecture F.91 (Sufficient condition for $\left|\mathcal{F}(s)\right| = |\Pi|$)**.** If $\forall s : \left|T(s)\right| = |\mathcal{A}|$ and $\forall s, s' \in \mathcal{S}, \mathbf{d} \in T(s) : \mathbf{d}^\top \mathbf{e}_{s'} > 0$, $\forall s : \left|\mathcal{F}(s)\right| = |\mathcal{A}|^{|\mathcal{S}|} = |\mathcal{S}|$.

Lemma F.92 ($|\mathcal{F}|$ bounds)**.**

$$1 \leq \left|\mathcal{F}(s)\right| \leq \prod_{s_i \in \textsc{Reach}(s)} \left|T(s_i)\right| \leq \prod_{i=1}^{|\mathcal{S}|} \left|T(s_i)\right| \leq |\mathcal{A}|^{|\mathcal{S}|} = |\Pi|.$$

Proof. $\forall s : 1 \leq \left|\mathcal{F}(s)\right|$ by lemma F.1. By the definition of visit distribution functions (definition 5.3), policy choices at unreachable states cannot affect the induced visit distribution function, so $\left|\mathcal{F}(s)\right| \leq \prod_{s_i \in \textsc{Reach}(s)} \left|T(s_i)\right|$. $\left|T(s_i)\right| \leq |\mathcal{A}|$ by the definition of child state distributions (definition F.5). $|\mathcal{A}|^{|\mathcal{S}|} = |\Pi|$ because Π is the set of deterministic stationary policies. □

By theorem F.90, if $|\mathcal{A}| > 1$, then $\left|\mathcal{F}(s)\right| < |\Pi|$ and so at least one of the intermediate inequalities must be strict. Figure F.13 demonstrates example cases.



(a) $\mathcal{F}(s) < \prod_{s_i \text{ reach. from } s} \left|T(s_i)\right|$

(b) $\prod_{i=1}^{|\mathcal{S}|} \left|T(s_i)\right| < |\Pi|$

Figure F.13: Suppose $|\mathcal{A}| = 2$ in both cases. a: $\forall s : \left|\mathcal{F}(s)\right| = 6 < 8 = \prod_{s_i \in \text{REACH}(s)} \left|T(s_i)\right| = \prod_{i=1}^{|\mathcal{S}|} \left|T(s_i)\right| = |\Pi|$. b: $\left|\mathcal{F}(s)\right| = 2 = \prod_{s_i \in \text{REACH}(s)} \left|T(s_i)\right| = \prod_{i=1}^{|\mathcal{S}|} \left|T(s_i)\right| < |\Pi| = 16$.

F.1.12 Variation distance of visit distributions in deterministic environments

Lemma F.93 (When $\gamma = 0$, the visit distributions of different states have 1 total variation)**.** *Let* $\mathbf{f} \in \mathcal{F}(s), \mathbf{f}' \in \mathcal{F}(s')$. $\text{TV}\left(\mathbf{f}(0), \mathbf{f}'(0)\right) = \mathbb{1}_{s \neq s'}$.

Proof. $\text{TV}\left(\mathbf{f}(0), \mathbf{f}'(0)\right) = \text{TV}\left(\mathbf{e}_s, \mathbf{e}_{s'}\right) = \mathbb{1}_{s \neq s'}$. $\square$

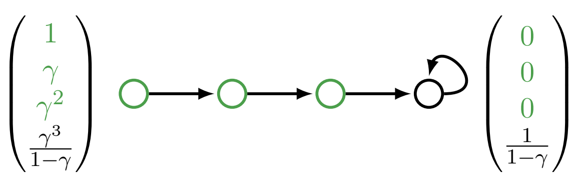


Figure F.14: Visit distribution functions $\mathbf{f}^\pi$ induced at different states along a path.

Lemma F.94 (Total variation along a graphical path)**.** *Let* $\gamma \in (0, 1)$. *Suppose that* π *travels a deterministic path from* $s_1, \ldots, s_{\ell+1}$ *and that* π *will not visit* s_ℓ *again.*

$$\text{TV}\left(\mathbf{f}^{\pi, s_1}(\gamma), \mathbf{f}^{\pi, s_{\ell+1}}(\gamma)\right) = \frac{1 - \gamma^\ell}{1 - \gamma}. \tag{F.82}$$

Proof. Since the path is deterministic, π never revisits any state in $s_1, \ldots, s_\ell$, since otherwise π would visit s_ℓ again. By definition 5.3, $\forall 1 \leq i < \ell + 1 : \mathbf{f}^{\pi, s_1}_{s_i}(\gamma) - \mathbf{f}^{\pi, s_{\ell+1}}_{s_i}(\gamma) = \gamma^{i-1}$: each s_i "loses" γ^{i-1} visitation frequency. Since all visitation distributions $\mathbf{f}$ have

$\left\|\mathbf{f}(\gamma)\right\|_1 = \frac{1}{1-\gamma}$, the total visitation frequency thus "lost" equals the total visitation frequency gained by other states (and therefore the total variation; see fig. F.14). Then $\text{TV}\left(\mathbf{f}^{\pi,s_1}(\gamma), \mathbf{f}^{\pi,s_{\ell+1}}(\gamma)\right) = \sum_{i=0}^{\ell-1} \gamma^i = \frac{1-\gamma^\ell}{1-\gamma}$. $\square$

Question F.95 (In stochastic environments, what general principles govern total variation among a policy's visit distributions?)**.**

Figure F.15 demonstrates how travelling along a deterministic cycle causes less total variation in the visit distributions.

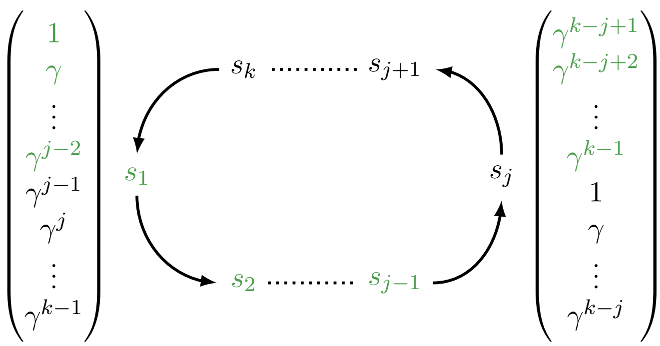


Figure F.15: Visitation distributions $\mathbf{f}^{\pi,s_1}(\gamma)$ and $\mathbf{f}^{\pi,s_j}$. Total variation is maximized at diametrically opposite states in the cycle. Factors of $\frac{1}{1-\gamma^k}$ left out to avoid clutter.

Lemma F.96 (Total variation along a graphical cycle)**.** *Let $\gamma \in (0,1)$ and suppose that starting from state s_1, π induces a deterministic k-cycle ($k > 1$).*

$$\max_{j\in[k]} \text{TV}\left(\mathbf{f}^{\pi,s_1}(\gamma), \mathbf{f}^{\pi,s_j}(\gamma)\right) \leq \frac{1-\gamma^{\frac{k}{2}}}{(1-\gamma)(1+\gamma^{\frac{k}{2}})} < \frac{1-\gamma^{\frac{k}{2}}}{1-\gamma}. \tag{F.83}$$

Proof.

$$\text{TV}\left(\mathbf{f}^{\pi,s_1}(\gamma), \mathbf{f}^{\pi,s_j}(\gamma)\right) = \sum_{i=0}^{j-1} \gamma^i - \gamma^{k-i-1} \tag{F.84}$$

$$= \frac{1-\gamma^j}{1-\gamma} \cdot \frac{1-\gamma^{k-j}}{1-\gamma^k} \tag{F.85}$$

$$= \frac{1-\gamma^j+\gamma^k-\gamma^{k-j}}{(1-\gamma)(1-\gamma^k)}. \tag{F.86}$$

Equation (F.84) can be verified by inspection of fig. F.15. Setting the derivative with respect to j to 0, we solve

$$0 = -\gamma^j + \gamma^{k-j} \tag{F.87}$$

$$j = \frac{k}{2}. \tag{F.88}$$

Equation (F.88) follows because $\gamma \in (0, 1)$. Solving via the derivative is justified because the function is strictly concave on $j \in [0, k]$ by the second-order test and the fact that $\gamma \in (0, 1)$. If k is even, we are done. If k is odd, then we need an integer solution. Plugging $j = \lfloor \frac{k}{2} \rfloor$ and $\lceil \frac{k}{2} \rceil$ into eq. (F.85) yields the same maximal result.

Therefore, in the odd case, both inequalities in the theorem statement are strict. In the even case, the first inequality is an equality. □

Proposition F.97 (Lower bound for total variation of a policy's visit distributions in deterministic environments)**.** *Suppose the environment is deterministic. For any* $\pi \in \Pi$*, if* $s \neq s'$*,* $\text{TV}\left(\mathbf{f}^{\pi,s}(\gamma), \mathbf{f}^{\pi,s'}(\gamma)\right) \geq \frac{1}{1+\gamma} \geq \frac{1}{2}$.

Proof. If π follows a path, it does so for at least 1 state. Plugging in $\ell = 1$ to lemma F.94 results in $\text{TV}\left(\mathbf{f}^{\pi,s}(\gamma), \mathbf{f}^{\pi,s'}(\gamma)\right) = 1$. If π follows a k-cycle ($k > 1$), it does so for at least one step. Then $j = 1, k = 2$ for lemma F.96, in which case $\text{TV}\left(\mathbf{f}^{\pi,s}(\gamma), \mathbf{f}^{\pi,s'}(\gamma)\right) = \frac{1}{1+\gamma} < 1$. □

F.2 Optimal value function theory

Lemma F.98 (Optimal value is piecewise rational on γ)**.** $V^*_R(s, \gamma)$ *is piecewise rational on* γ.

Proof. By lemma D.40, the optimal visit distribution changes a finite number of times for $\gamma \in [0, 1)$. Corollary D.10 implies $V^*_R(s, \gamma)$ is rational on each non-degenerate subinterval where the optimal visit distribution set is constant. □

Lemma F.99 ($V^*_R(s, \gamma)$ is piecewise linear on R)**.**

Proof. $V^*_R(s,\gamma) = \max_{\mathbf{f}\in\mathcal{F}(s)} \mathbf{f}(\gamma)^\top \mathbf{r}$ takes the maximum over a finite set of fixed $|\mathcal{S}|$-dimensional linear functionals. Therefore, the maximum is piecewise linear with respect to R. □

Lemma F.99 shows that optimal value is piecewise linear in the reward function. In unpublished work, Jacob Stavrianos showed that optimal value is globally sublinear in the reward function:

Lemma F.100 (Optimal value is sublinear in the reward function)**.** *Let R_1, R_2 be reward functions and let $\gamma \in [0,1)$.*

1. *Let $r \geq 0$. $V^*_{rR_1}(s,\gamma) = rV^*_{R_1}(s,\gamma)$.*
2. *$V^*_{R_1+R_2}(s,\gamma) \leq V^*_{R_1}(s,\gamma) + V^*_{R_2}(s,\gamma)$.*

Proof. $V^*_{rR_1}(s,\gamma) = \max_{\mathbf{f}\in\mathcal{F}(s)} \mathbf{f}(\gamma)^\top (r\mathbf{r}_1) = r\max_{\mathbf{f}\in\mathcal{F}(s)} \mathbf{f}(\gamma)^\top \mathbf{r}_1 = rV^*_{R_1}(s,\gamma)$.

For the second condition, we check that

$$V^*_{R_1+R_2}(s,\gamma) = \max_{\mathbf{f}\in\mathcal{F}(s)} \mathbf{f}(\gamma)^\top(\mathbf{r}_1+\mathbf{r}_2) \tag{F.89}$$

$$\leq \max_{\mathbf{f}_1\in\mathcal{F}(s)} \mathbf{f}(\gamma)^\top \mathbf{r}_1 + \max_{\mathbf{f}_2\in\mathcal{F}(s)} \mathbf{f}(\gamma)^\top \mathbf{r}_2 \tag{F.90}$$

$$= V^*_{R_1}(s,\gamma) + V^*_{R_2}(s,\gamma). \tag{F.91}$$

□

Corollary F.101 (Optimal value is concave in the reward function)**.**

Proof. Optimal value is sublinear in the reward function by lemma F.100; sublinearity is a sufficient condition for concavity. □

Lemma F.102 (Optimal value is monotonically increasing in the reward function)**.** *Let $s \in \mathcal{S}$, $\gamma \in [0,1)$, and suppose $\forall s' \in \mathcal{S} : R_1(s') \geq R_2(s')$. Then $V^*_{R_1}(s,\gamma) \geq V^*_{R_2}(s,\gamma)$.*

Proof. Let $\mathbf{f}_2 \in \arg\max_{\mathbf{f}\in\mathcal{F}(s)} \mathbf{f}(\gamma)^\top \mathbf{r}_2$.

$$\begin{aligned} V^*_{R_1}(s,\gamma) &= \max_{\mathbf{f}\in\mathcal{F}(s)} \mathbf{f}(\gamma)^\top \mathbf{r}_1 && \text{(F.92)}\\ &\geq \mathbf{f}_2(\gamma)^\top \mathbf{r}_1 && \text{(F.93)}\\ &\geq \mathbf{f}_2(\gamma)^\top \mathbf{r}_2 && \text{(F.94)}\\ &= V^*_{R_2}(s,\gamma). && \text{(F.95)} \end{aligned}$$

Equation (F.94) follows by the assumed component-wise domination $\mathbf{r}_1 \succeq \mathbf{r}_2$. Equation (F.95) follows by the definition of $\mathbf{f}_2$. □

Theorem F.103 (Reward functions map injectively to optimal value functions)**.** *$\forall \gamma \in [0,1)$, $R \mapsto V^*_R(\cdot,\gamma)$ is injective.*

Proof. Given $V^*_R(\cdot,\gamma)$ and the rewardless MDP, deduce an optimal policy π^* for R by choosing a $V^*_R(\cdot,\gamma)$-greedy action for each state.

$$\begin{aligned} V^*_R(\cdot,\gamma) &= R + \gamma \mathbf{T}^{\pi^*} V^*_R && \text{(F.96)}\\ \left(\mathbf{I} - \gamma \mathbf{T}^{\pi^*}\right) V^*_R(\cdot,\gamma) &= R. && \text{(F.97)} \end{aligned}$$

If two reward functions have the same optimal value function, then they have the same optimal policies. Then eq. (F.97) shows that the reward functions must be identical. □

Scott Emmons provided the proof sketch for theorem F.103.

Lemma F.104 (Linear independence of a policy's visit distributions)**.** *At any fixed $\gamma \in [0,1)$, the elements of $\left\{\mathbf{f}^{\pi,s}(\gamma) \mid s \in \mathcal{S}\right\}$ are linearly independent.*

Proof. Consider the all-zero optimal value function with optimal policy π^*. Theorem F.103 implies the following homogeneous system of equations has a unique solution for $\mathbf{r}$:

$$\begin{gathered} \mathbf{f}^{\pi^*,s_1}(\gamma)^\top \mathbf{r} = 0 \\ \vdots \end{gathered}$$

$$\mathbf{f}^{\pi^*,s_{|\mathcal{S}|}}(\gamma)^\top \mathbf{r} = 0.$$

Therefore, π^* induces linearly independent $\mathbf{f}$. But $\mathbf{r}$ must be the all-zero reward function (for which all policies are optimal), so the $\mathbf{f}^{\pi,s}$ are independent for *any* policy π. ☐

Lemma F.105 (Two distinct visit distributions differ in expected optimal value for almost all reward functions)**.** *Let* $\gamma \in (0,1)$*, and let* $\Delta, \Delta' \in \Delta(\mathcal{S})$*. If* $\Delta \neq \Delta'$*,*

$$\mathbb{P}_{R\sim\mathcal{D}_{cont}}\left(\underset{s\sim\Delta}{\mathbb{E}}\left[V^*_R\left(s,\gamma\right)\right] = \underset{s'\sim\Delta'}{\mathbb{E}}\left[V^*_R\left(s',\gamma\right)\right]\right) = 0. \tag{F.98}$$

Proof. Let $R \in \operatorname{supp}(\mathcal{D}_{\text{cont}})$ and $\pi^* \in \Pi^*(R,\gamma)$. By lemma F.104, $\mathbb{E}_\Delta\left[\mathbf{f}^{\pi^*,s}\right] = \mathbb{E}_{\Delta'}\left[\mathbf{f}^{\pi^*,s'}\right]$ iff $\Delta = \Delta'$. Therefore, $\mathbb{E}_\Delta\left[\mathbf{f}^{\pi^*,s}\right] \neq \mathbb{E}_{\Delta'}\left[\mathbf{f}^{\pi^*,s'}\right]$. Trivially, $\mathbb{E}_\Delta\left[V^*_R\left(s,\gamma\right)\right] = \mathbb{E}_{\Delta'}\left[V^*_R\left(s',\gamma\right)\right]$ iff $\mathbb{E}_\Delta\left[\mathbf{f}^{\pi^*,s}\right]^\top \mathbf{r} = \mathbb{E}_{\Delta'}\left[\mathbf{f}^{\pi^*,s'}\right]^\top \mathbf{r}$. Since $\mathbb{E}_\Delta\left[\mathbf{f}^{\pi^*,s}\right] \neq \mathbb{E}_{\Delta'}\left[\mathbf{f}^{\pi^*,s'}\right]$, lemma D.12 implies that the equality holds with 0 probability under $\mathcal{D}_{\text{cont}}$. ☐

No $\mathbf{f}$ is *sub*optimal for all reward functions: every visit distribution is optimal for a constant reward function. However, for any given γ, almost every reward function has a unique optimal visit distribution at each state.

Lemma F.106 (Optimal visit distributions are almost always unique)**.** *Let s be any state. For any* $\gamma \in (0,1)$*,* $\left\{\mathbf{r} \textit{ such that } \left|\arg\max_{\mathbf{f}\in\mathcal{F}(s)} \mathbf{f}(\gamma)^\top \mathbf{r}\right| > 1\right\}$ *has measure zero under any continuous reward function distribution.*

Proof. Let R be a reward function and let s be a state at which there is more than one optimal visit distribution for R at discount rate γ. Since R has more than one optimal visit distribution, there must exist a state s' reachable with positive probability from s such that actions a, a' are both optimal at s', where $a \not\equiv_{s'} a'$. Then $\mathbb{E}_{s''\sim T(s',a)}\left[V^*_R\left(s'',\gamma\right)\right] = \mathbb{E}_{s''\sim T(s',a')}\left[V^*_R(s'',\gamma)\right]$.

By lemma F.105, since $T(s',a) \neq T(s',a')$, this equation holds with probability 0 for reward functions drawn from any continuous reward function distribution. ☐

F.2.1 Discovering the true reward function

Definition F.107 (Value of reward information)**.** Let $\mathcal{D}_{\text{bound}}$ have mean reward function $\bar{R}$. For state s and $\gamma \in [0, 1)$, $\text{VOI}_{\mathcal{D}}(\mathcal{D}_{\text{bound}}) \coloneqq V^*_{\mathcal{D}_{\text{bound}}}(s, \gamma) - V^*_{\bar{R}}(s, \gamma)$.

Question F.108 (In what situations is $\text{VOI}_{\mathcal{D}}(s, \gamma)$ small?)**.**

Proposition F.109 (Value of reward information is non-negative)**.** *For state s and $\gamma \in [0, 1)$, $\text{VOI}_{\mathcal{D}}(\mathcal{D}_{bound}) \geq 0$.*

Proof.

$$
\begin{aligned}
V^*_{\bar{R}}(s, \gamma) &\coloneqq \max_{\pi} V^{\pi}_{\bar{R}}(s, \gamma) && \text{(F.99)}\\
&= \max_{\pi} \mathop{\mathbb{E}}_{R \sim \mathcal{D}_{\text{bound}}}\left[V^{\pi}_{R}(s, \gamma)\right] && \text{(F.100)}\\
&\leq \mathop{\mathbb{E}}_{R \sim \mathcal{D}_{\text{bound}}}\left[\max_{\pi} V^{\pi}_{R}(s, \gamma)\right] && \text{(F.101)}\\
&= V^*_{\mathcal{D}_{\text{bound}}}(s, \gamma)\,. && \text{(F.102)}
\end{aligned}
$$

The result follows since $\text{VOI}_{\mathcal{D}}(\mathcal{D}_{\text{bound}}) \coloneqq V^*_{\mathcal{D}_{\text{bound}}}(s, \gamma) - V^*_{\bar{R}}(s, \gamma)$ by definition F.107. $\square$

F.3 MDP Structure

Knowing the visit distribution functions for each state provides an enormous amount of information about the MDP. As $\gamma \to 0$, the local dynamics are revealed (lemma F.9). As $\gamma \to 1$, the renormalized visit distributions $(1 - \gamma)\mathbf{f}(\gamma)$ limit to the recurrent state distributions which can be induced from state s: $\text{RSD}(s)$ (definition 5.26).

As it turns out, $\mathcal{F}$ encodes the *entire* MDP (corollary F.114). First, we recap all of our visit distribution notation in table F.1.

Definition F.110 (Visitation function isomorphism)**.** Let $M \coloneqq \langle \mathcal{S}, \mathcal{A}, T, \gamma\rangle$ and $M' \coloneqq \langle \mathcal{S}', \mathcal{A}', T', \gamma'\rangle$ be two rewardless MDPs. $M \cong_{\mathcal{F}} M'$ (read "M and M' have *isomorphic visitation functions*") when there exists a bijection $\phi : \mathcal{S} \to \mathcal{S}'$ (with corresponding permutation matrix $\mathbf{P}_{\phi}$) satisfying $\forall s \in \mathcal{S} : \mathcal{F}_{M'}(\phi(s)) = \phi \cdot \mathcal{F}_M(s)$.

Notation	Meaning
Visit distribution function $\mathbf{f}^{\pi,s}$	Definition 5.3: The discounted state visit distribution function induced by following policy π starting from state s. A function from $\gamma \in [0,1)$ to $\mathbb{R}^{\lvert\mathcal{S}\rvert}$. π and s are often left implicit.
Visit distribution $\mathbf{f}^{\pi,s}(\gamma)$	Definition 5.3: $\mathbf{f}^{\pi,s}$ evaluated at discount rate γ. $\left\lVert\mathbf{f}^{\pi,s}(\gamma)\right\rVert_1 = \frac{1}{1-\gamma}$. π and s are often left implicit.
$\mathcal{F}(s)$	Definition 5.3: $\left\{\mathbf{f}^{\pi,s} \mid \pi \in \Pi\right\}$.
$\mathcal{F}(s,\gamma)$	Definition D.36: $\left\{\mathbf{f}(\gamma) \mid \mathbf{f} \in \mathcal{F}(s)\right\}$.
$\mathcal{F}_{\text{nd}}(s)$	Definition 5.6: The elements of $\mathcal{F}(s)$ which are strictly optimal for some reward function $\mathbf{r} \in \mathbb{R}^{\lvert\mathcal{S}\rvert}$ and discount rate $\gamma \in (0,1)$.
$\mathcal{F}_{\text{nd}}(s,\gamma)$	Definition D.36: $\left\{\mathbf{f}(\gamma) \mid \mathbf{f} \in \mathcal{F}_{\text{nd}}(s)\right\}$.
$\mathcal{F}_{\text{nd}}(s \mid \pi(s') = a)$	Definition 5.4: The elements of $\mathcal{F}_{\text{nd}}(s)$ whose policies take action a at state s'.
$\mathcal{F}_{\text{nd}}(s \mid \pi^*, S)$	Definition F.158: The elements of $\mathcal{F}_{\text{nd}}(s)$ whose policies agree with π^* on the states in $S \subseteq \mathcal{S}$.
$\mathcal{F}^{\text{HD}}(s)$	Definition F.49: $\left\{\mathbf{f}^{\pi,s} \mid \pi \in \Pi^{\text{HD}}\right\}$.

Table F.1: Summary of visit distribution notation.

This isomorphism is invariant to state representation, state labelling, action labelling, and the addition of superfluous actions (a such that $\forall s : \exists a' \neq a : T(s,a) = T(s,a')$).

Definition F.111 (Directed graph of a deterministic MDP)**.** The *directed graph* of a deterministic MDP is a directed graph with a vertex for each state such that there is an arrow from vertex s to vertex s' iff $s' \in Ch(s)$.

Directed graphs are deterministic special cases of so-called MDP *models* [68]. We introduce the following definition; when the dynamics are deterministic, it reduces to the standard directed graph isomorphism.

Definition F.112 (Stochastic model isomorphism)**.** ϕ is a stochastic model isomorphism between rewardless MDPs M, M' when ϕ is a bijection $\phi : \mathcal{S} \to \mathcal{S}'$ is such that for all $s \in \mathcal{S}$, $\left\{\mathbf{P}_\phi T(s,a) \mid a \in \mathcal{A}\right\} = \left\{T'(\phi(s),a') \mid a' \in \mathcal{A}'\right\}$. We then say that $M \cong_\phi M'$, which is read as read "M and M' have *isomorphic transitions*".

Theorem F.113 ($\cong_{\mathcal{F}}$ is equivalent to transition isomorphism)**.** *$M \cong_{\mathcal{F}} M'$ via bijection ϕ iff $M \cong_\phi M'$.*

Proof. Forward direction: let $s \in \mathcal{S}$.

$$\left\{T'(\phi(s), a') \mid a' \in \mathcal{A}'\right\} = \left\{\lim_{\gamma \to 0} \gamma^{-1}(\mathbf{f}^{\pi', \phi(s)}(\gamma) - \mathbf{e}_{\phi(s)}) \mid \pi' \in \Pi'\right\} \tag{F.103}$$

$$= \left\{\mathbf{P}_\phi \lim_{\gamma \to 0} \gamma^{-1}(\mathbf{f}^{\pi, s}(\gamma) - \mathbf{e}_s) \mid \pi \in \Pi\right\} \tag{F.104}$$

$$= \left\{\mathbf{P}_\phi T(s, a) \mid a \in \mathcal{A}\right\}. \tag{F.105}$$

Equation (F.104) follows because $M \cong_{\mathcal{F}} M'$. Then $M \cong_\phi M'$.

Suppose instead that $M \cong_\phi M'$. Let $\pi \in \Pi$ be a policy in M. Let π' be such that for all $s \in \mathcal{S}$: π' satisfies $T'(\phi(s), \pi'(\phi(s))) = \mathbf{P}_\phi T(s, \pi(s))$; such actions exist because we assumed that $M \cong_\phi M'$. Then by repeated application of lemma F.9, $\mathbf{f}^{\pi', \phi(s)} = \mathbf{P}_\phi \mathbf{f}^{\pi, s}$ for all $s \in \mathcal{S}$. Since π was arbitrary, $M \cong_{\mathcal{F}} M'$ via ϕ. $\square$

Corollary F.114 (Visit distribution functions encode MDPs)**.** *Given the function $\mathcal{F}$, the generating dynamics can be reconstructed up to transition isomorphism.*

Proof. Given $\mathcal{F}$, for each state s, deduce $T(s) = \left\{\lim_{\gamma \to 0} \gamma^{-1}(\mathbf{f}^s(\gamma) - \mathbf{e}_s) \mid \mathbf{f} \in \mathcal{F}(s)\right\}$. $\square$

In deterministic environments, the dynamics are encoded (up to transition isomorphism) by the visit distributions at a *single* $\gamma \in (0, 1)$.

Theorem F.115 (Visit distributions encode rewardless deterministic MDPs)**.** *Given the function $\mathcal{F}(\cdot, \gamma)$ generated by a deterministic rewardless MDP, the generating dynamics can be reconstructed up to transition isomorphism.*

Proof. Since $\forall s \in \mathcal{S}, \mathbf{f}(\gamma) \in \mathcal{F}(s, \gamma) : \left\|\mathbf{f}(\gamma)\right\|_1 = \frac{1}{1-\gamma}$ (by proposition D.8) and $\mathbf{f}(\gamma) \in \mathbb{R}^{|\mathcal{S}|}$, we can deduce γ and $|\mathcal{S}|$. Let $\mathcal{S}' \coloneqq \left\{1, \ldots, |\mathcal{S}|\right\}$. Using lemma F.9, deduce the children $Ch\,(s)$ of each state s. Define $\mathcal{A} \coloneqq \left\{1, \ldots, \max_s \left|Ch\,(s)\right|\right\}$. Construct a transition function T' using $Ch\,(s)$; if $|\mathcal{A}| > \left|Ch\,(s)\right|$ for some state s, map the redundant actions to any element of $Ch\,(s)$.

Let $M' \coloneqq \langle \mathcal{S}', \mathcal{A}', T', \gamma \rangle$. By construction, M and M' are transition isomorphic. $\square$

Conjecture F.116 (Theorem F.115 holds in stochastic environments)**.**

Theorem F.117 (Optimal value functions encode rewardless deterministic MDPs)**.** *Given the optimal value function/reward function pairs of a rewardless deterministic* MDP M*, M can be reconstructed up to $\mathcal{F}$-isomorphism.*

Proof. Suppose that for rewardless MDP M, we are given $\left\{(R, V^*_R(\cdot, \gamma)) \mid R \in [0,1]^{|\mathcal{S}|}\right\}$ for fixed γ. Let $\mathcal{S}' \coloneqq \left\{1, \ldots, |\text{domain of } R|\right\}$ for any reward function R.

For each s, we determine if it can reach itself. Let R_s be the indicator reward function on state s. s can reach itself iff $V^*_{R_s}(s, \gamma) > 1$. Because the MDP is finite, at least one state s must be able to reach itself.

If other states s' also have $V^*_{R_s}(s', \gamma) > 0$, consider $s_\text{pre} \in \arg\max_{s' \in \mathcal{S} \setminus \{s\}} V^*_{R_s}(s', \gamma)$. s_pre must be able to reach s in one step, so $V^*_{R_s}(s_\text{pre}, \gamma) = \gamma V^*_{R_s}(s, \gamma)$. Then $\gamma = \frac{V^*_{R_s}(s,\gamma)}{V^*_{R_s}(s_\text{pre},\gamma)}$.

If s is the only state with positive optimal value for R_s, $V^*_{R_s}(s, \gamma) > 1$ implies that s must be able to reach itself. Then $V^*_{R_s}(s, \gamma) = \frac{1}{1-\gamma}$; solve for γ.

The above reasoning explained how to test whether s can reach itself and how to determine which other states can reach s. This information allows us to construct a transition function T', setting the action space $\mathcal{A}'$ to be as large as necessary to accommodate the state with the most children. Because their directed graphs are isomorphic, $\langle \mathcal{S}', \mathcal{A}', \gamma, T' \rangle \cong_\mathcal{F} M$ by theorem F.113. $\square$

Remark. The proof of theorem F.117 shows that deterministic dynamics are fully determined by $|\mathcal{S}|$ optimal value functions (one for each state indicator reward function).

Corollary F.118 (Non-dominated visit distribution functions encode rewardless deterministic MDPs)**.** *Suppose the rewardless deterministic* MDP $M \coloneqq \langle \mathcal{S}, \mathcal{A}, T, \gamma \rangle$ *induces* $\mathcal{F}_\text{nd}$*. From $\mathcal{F}_\text{nd}$, M can be reconstructed up to $\mathcal{F}$-isomorphism.*

Proof. Restriction to non-dominated distributions leaves optimal value unchanged for all reward functions. Apply theorem F.117 to recover M up to $\mathcal{F}$-isomorphism. $\square$

Figure F.16 shows that neither theorem F.117 nor corollary F.118 hold for stochastic environments. Given known transition dynamics, theorem F.103 guarantees that R recovered from $V^*_R(\cdot, \gamma)$, but in the stochastic variant of theorem F.117, we would not know the transition dynamics *a priori*.

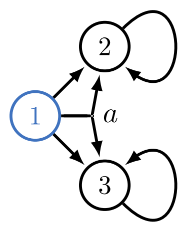


Figure F.16: The bifurcated action a is a stochastic transition, where $T(s_1, a, s_2) = p, T(s_1, a, s_3) = 1 - p$. For any $p \in (0, 1)$, a is a dominated action: $T(s_1, a) \in T(s_1) \setminus T_{\text{nd}}(s_1)$. Since there is no optimal value function for which it is strictly optimal to take action a, no optimal value function is affected by the presence of a. This ambiguity does not arise in deterministic MDPs, since $T(s) = T_{\text{nd}}(s)$ when the dynamics are deterministic.

Figure F.17 summarizes this section's results.

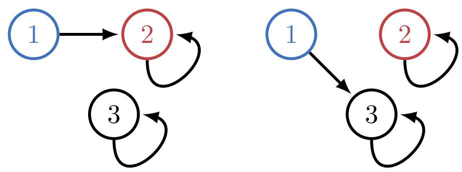
$$\begin{tikzcd}
& \mathcal{F}_{\text{nd}}(\cdot, \gamma) \arrow[dl, leftrightarrow] \arrow[dr, leftrightarrow] & \\
\langle \mathcal{S}, \mathcal{A}, \gamma, T \rangle \arrow[rr, leftrightarrow] & & \left\{ \left(R_s, V^*_{R_s}(\cdot, \gamma) \right) \mid s \in \mathcal{S} \right\}
\end{tikzcd}$$

Figure F.17: In deterministic MDPs, these three objects contain the same information (up to transition isomorphism).

Figure F.18 shows that theorem F.117 cannot be proven without knowing which reward functions generate which optimal value functions (although γ can still be deduced from the optimal value functions for all reward functions with reward bounded in $[0, 1]$).

Question F.119 (In what category of MDPs is $\cong_{\mathcal{F}}$ an isomorphism?)**.**

Question F.120 (Is $\cong_{\mathcal{F}}$ natural in the category-theoretic sense?)**.**



(a) $R(s_1) = 1$ (b) $R(s_1) = .8$

Figure F.18: Suppose $\gamma = .5$ and $V_R^*(s_1, .5) = 1.5, V_R^*(s_2, .5) = 1, V_R^*(s_3, .5) = 1.4$. This optimal value function is compatible with a (where $R(s_1) = 1$) and with b (where $R(s_1) = .8$).

Question F.121 (What properties would $\cong_{\mathcal{F}_{\text{nd}}}$ have?)**.**

Conjecture F.122 (Optimality probability and Power change "continuously" with respect to transition dynamics)**.**

F.4 Properties of optimal policy shifts

Definition F.123 (Optimal policy shift)**.** R has an *optimal policy shift* at $\gamma \in (0, 1)$ when $\lim_{\gamma^- \uparrow \gamma} \Pi^*\left(R, \gamma^-\right) \neq \Pi^*(R, \gamma)$. Similarly, R has an *optimal visit distribution shift* at γ and at state s.

Corollary F.124 (One-sided limits exist for $\Pi^*(R, \gamma)$)**.** *Let $L \in (0, 1)$ and let R be any reward function.* $\lim_{\gamma \uparrow L} \Pi^*(R, \gamma)$ *and* $\lim_{\gamma \downarrow L} \Pi^*(R, \gamma)$ *both exist.*

Proof. By lemma D.40, $\Pi^*(R, \gamma)$ can take on at most finitely many values for $\gamma \in (0, 1)$. Thus, infinite oscillation cannot occur in either one-sided limit, and so both one-sided limits exist. □

Corollary F.126 shows that definition F.123 loses no generality by defining optimal policy shifts with respect to the limit from below.

Lippman [48] showed that two visit distribution functions can trade off optimality status at most $2|\mathcal{S}| + 1$ times. We slightly improve this upper bound. We thank Max Sharnoff for contributions to lemma F.125.

Lemma F.125 (Upper bound on optimal visit distribution shifts)**.** *For any reward function R and $\mathbf{f}, \mathbf{f}' \in \mathcal{F}(s)$, $\left(\mathbf{f}(\gamma) - \mathbf{f}'(\gamma)\right)^\top \mathbf{r}$ is either the zero function, or it has at most $2\,|\mathcal{S}| - 1$ roots on $\gamma \in (0,1)$.*

Proof. Consider two policies π, π'. By lemma D.9, $(\mathbf{f}^\pi(\gamma) - \mathbf{f}^{\pi'}(\gamma))^\top \mathbf{r}$ is a rational function with degree at most $2\,|\mathcal{S}|$ by the sum rule for fractions. The fundamental theorem of algebra shows that $(\mathbf{f}^\pi(\gamma) - \mathbf{f}^{\pi'}(\gamma))^\top \mathbf{r}$ is either 0 for all γ or for at most $2\,|\mathcal{S}|$ values of $\gamma \in [0,1)$. Since $\mathbf{f}(0) = \mathbf{f}'(0) = \mathbf{e}_s$ (definition 5.3), one of the roots is at $\gamma = 0$. ☐

Corollary F.126 (Lower-limit optimal policy set inequality iff upper-limit inequality)**.** *Let $\gamma \in (0,1)$, and $\Pi^- := \lim_{\gamma^- \uparrow \gamma} \Pi^*\left(R, \gamma^-\right), \Pi^+ := \lim_{\gamma^+ \downarrow \gamma} \Pi^*\left(R, \gamma^+\right)$. $\Pi^- \neq \Pi^*\left(R, \gamma\right)$ iff $\Pi^+ \neq \Pi^*\left(R, \gamma\right)$.*

Proof. Suppose $\Pi^- \neq \Pi^*\left(R, \gamma\right)$ but $\Pi^*\left(R, \gamma\right) = \Pi^+$. By lemma F.127, if $\Pi^- \neq \Pi^*\left(R, \gamma\right)$, then $\Pi^- \subsetneq \Pi^*\left(R, \gamma\right)$. Let $\pi^* \in \Pi^*\left(R, \gamma\right) \setminus \Pi^-$ and $\pi^- \in \Pi^-$. Since $\pi^* \notin \Pi^-$, there exists some $\epsilon_1 > 0$ such that π^* isn't optimal for all $\gamma' \in (\gamma - \epsilon_1, \gamma]$. In particular, $(\mathbf{f}^{\pi^*}(\gamma') - \mathbf{f}^{\pi^-}(\gamma'))^\top \mathbf{r} < 0$ for such γ'. In particular, $(\mathbf{f}^{\pi^*}(\gamma^*) - \mathbf{f}^{\pi^-}(\gamma^*))^\top \mathbf{r}$ is not the zero function on γ^*.

Therefore, lemma F.125 implies that $(\mathbf{f}^{\pi^*}(\gamma^*) - \mathbf{f}^{\pi^-}(\gamma^*))^\top \mathbf{r}$ has finitely many roots on γ^*. But since $\pi^* \in \Pi^*\left(R, \gamma\right) = \Pi^+$, there exists $\epsilon_2 > 0$ such that $\forall \gamma' \in [\gamma, \gamma + \epsilon_2) :$ $(\mathbf{f}^{\pi^*}(\gamma') - \mathbf{f}^{\pi^-}(\gamma'))^\top \mathbf{r} = 0$. But this would imply that the expression has infinitely many roots, a contradiction. Therefore, if $\Pi^- \neq \Pi^*\left(R, \gamma\right)$, then $\Pi^*\left(R, \gamma\right) \neq \Pi^+$.

The proof of the reverse implication proceeds identically. ☐

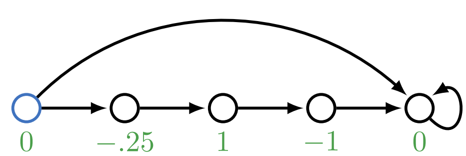


Figure F.19: In lemma F.127, Π^- can equal Π^+. Let R be the reward function whose rewards are shown in green. The shortcut is optimal for all γ. An optimal policy shift occurs at $\gamma = .5$. Since $\Pi^- = \Pi^+$ only contain policies which take the shortcut, $\Pi^- \cup \Pi^+ \subsetneq \Pi^*\left(R, \gamma\right)$.

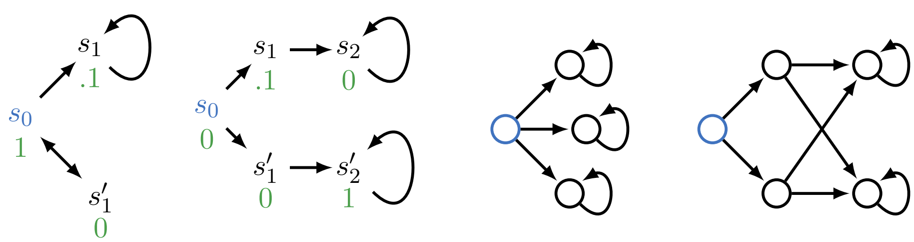
Lemma F.127 (Optimal policy sets overlap when shifts occur)**.** *Let R be a reward function and $\gamma \in (0,1)$. Let $\Pi^- := \lim_{\gamma^- \uparrow \gamma} \Pi^*\left(R, \gamma^-\right), \Pi^+ := \lim_{\gamma^+ \downarrow \gamma} \Pi^*\left(R, \gamma^+\right)$. Then $\Pi^- \cup \Pi^+ \subseteq \Pi^*(R,\gamma)$. Furthermore, if R has an optimal policy shift at γ, $\exists s \in \mathcal{S}$:*

$$\left| \underset{\mathbf{f} \in \mathcal{F}(s)}{\arg\max}\, \mathbf{f}(\gamma)^\top \mathbf{r} \right| \geq 2.$$

Proof. Since an optimal policy shift occurs at γ and since $V_R^*(s,\gamma)$ is continuous on γ by lemma F.98, $\forall \pi^- \in \Pi^-, \pi^+ \in \Pi^+, s \in \mathcal{S} : V_R^{\pi^-}(s,\gamma) = V_R^{\pi^+}(s,\gamma)$. Therefore, $\Pi^- \cup \Pi^+ \subseteq \Pi^*(R,\gamma)$.

By lemma D.6, for any $R \in \mathbb{R}^{\mathcal{S}}$, an optimal policy shift occurs at γ iff an optimal visit distribution shift occurs at γ for at least one state s. □

To better appreciate how optimal policy sets can be linked to the discount rate, consider the fact that some rewardless MDPs have no optimal policy shifts. In other words, for any reward function and for all $\gamma \in (0,1)$, greedy policies are optimal, as shown in fig. F.20. In deterministic environments, optimal policy shifts can occur if and only if the agent can be made to choose between lesser immediate reward and greater delayed reward.



(a) (b) (c) (d)

Figure F.20: a and b show reward functions whose optimal policies shift. No shifts occur in c or d.

Theorem F.128 suggests that the vast majority of deterministic rewardless MDPs allow optimal policy shifts, as the criterion is easily fulfilled.

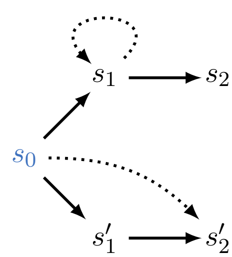
Theorem F.128 (Characterization of optimal policy shifts in deterministic rewardless MDPs)**.** *In deterministic environments, there exists a reward function whose optimal action* *at s_0 changes with γ iff $\exists s_1 \in Ch(s_0)$, $s_1' \in Ch(s_0)$, $s_2' \in Ch(s_1') \setminus Ch(s_1)$:*

$$s_2' \notin Ch(s_0) \vee \left(s_1 \notin Ch(s_1) \wedge s_1' \notin Ch(s_1)\right).$$

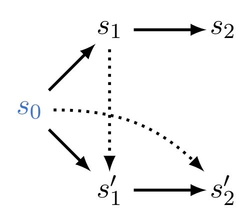
Proof. **Forward direction.** Without loss of generality, suppose the optimal policy set of some R is shifting for the first time (a finite number of shifts occur by Blackwell [11]).

Starting at state s_0, let the policies π, π' induce state trajectories $s_0 s_1 s_2 \ldots$ and $s_0 s_1' s_2' \ldots$, respectively, with the shift occurring to an optimal policy set containing π' at discount rate γ. By the definition of an optimal policy shift at γ, $V_R^{\pi}(s_0, \gamma) = V_R^{\pi'}(s_0, \gamma)$. Because π was greedily optimal and π' was not, $s_1 \neq s_1'$ and $R(s_1) > R(s_1')$. If $Ch\,(s_1) = Ch\left(s_1'\right)$, $\pi(s_0)$ remains the optimal action at s_0 and no shift occurs. Without loss of generality, suppose $s_2' \notin Ch\,(s_1)$.

We show the impossibility of $\neg\Big(s_2' \notin Ch\,(s_0) \vee \left(s_1 \notin Ch\,(s_1) \wedge s_1' \notin Ch\,(s_1)\right)\Big) = s_2' \in Ch\,(s_0) \wedge \left(s_1 \in Ch\,(s_1) \vee s_1' \in Ch\,(s_1)\right)$, given that π' becomes optimal at γ.



(a) $s_2' \in Ch\,(s_0) \wedge s_1 \in Ch\,(s_1)$



(b) $s_2' \in Ch\,(s_0) \wedge s_1' \in Ch\,(s_1)$

Figure F.21: Dotted arrows illustrate the assumptions for each case. Given that there exists a reward function R whose optimal action at s_0 changes at γ, neither assumption can hold. Although not illustrated here, *e.g.* $s_2 = s_0$ or $s_2' = s_0$ is consistent with theorem F.128. We leave the rest of the model blank as we make no further assumptions about its topology.

Case: $s_2' \in Ch\,(s_0) \wedge s_1 \in Ch\,(s_1)$. For π' to be optimal, navigating to s_1 and staying there cannot be a better policy than following π' from s_0. Formally, $\frac{R(s_1)}{1-\gamma} \leq V_R^{\pi'}(s_1', \gamma)$ implies $R(s_1) \leq (1-\gamma) V_R^{\pi'}(s_1', \gamma) = (1-\gamma)\left(R(s_1') + \gamma V_R^{\pi'}(s_2', \gamma)\right)$.

We now construct a policy π_2' which strictly improves upon π'. Since $s_2' \in Ch\,(s_0)$,

$\exists a'_2 : T(s_0, a'_2, s'_2) = 1$. Let π'_2 equal π' except that $\pi'_2(s_0) \coloneqq a'_2$. Then since $R(s'_1) < R(s_1)$, $V_R^{\pi'_2}(s_0, \gamma) > V_R^{\pi'}(s_0, \gamma)$, contradicting the assumed optimality of π'.

Case: $s'_2 \in Ch\,(s_0) \wedge s'_1 \in Ch\,(s_1)$. For π' to be optimal, navigating to s_1, then to s'_1 (made possible by $s'_1 \in Ch\,(s_1)$), and then following π' cannot be a better policy than following π' from s_0. Formally, $R(s_1) + \gamma V_R^{\pi'}(s'_1, \gamma) \leq V_R^{\pi'}(s'_1, \gamma)$. This implies that $R(s_1) \leq (1-\gamma) V_R^{\pi'}(s'_1, \gamma) = (1-\gamma)\left(R(s'_1) + \gamma V_R^{\pi'}(s'_2, \gamma)\right)$. The policy π'_2 constructed above is again a strict improvement over π' at discount rate γ, contradicting the assumed optimality of π'.

Backward direction. Suppose $\exists s_1, s'_1 \in Ch\,(s_0)\,, s'_2 \in Ch\,(s'_1) \setminus Ch\,(s_1) : s'_2 \notin Ch\,(s_0) \vee \left(s_1 \notin Ch\,(s_1) \wedge s'_1 \notin Ch\,(s_1)\right)$. We show that there exists a reward function R whose optimal policy at s_0 changes with γ.

If $s'_2 \notin Ch\,(s_0)$, then $s'_2 \neq s_1$ because $s_1 \in Ch\,(s_0)$. Let $R(s_1) \coloneqq .1$, $R(s'_2) \coloneqq 1$, and 0 elsewhere. Suppose that s_1 can reach s'_2 in two steps and then stay there indefinitely, while the state trajectory of π' can only stay in s'_2 for one time step, after which no more reward accrues. Even under these impossibly conservative assumptions, an optimal trajectory shift occurs from $s_0 s_1 s_2 \ldots$ to $s_0 s'_1 s'_2 \ldots$. At the latest, the shift occurs at $\gamma \approx 0.115$, which is a solution of the corresponding equality:

$$R(s_1) + \frac{\gamma^2}{1-\gamma} R(s'_2) = R(s'_1) + \gamma R(s'_2) \tag{F.106}$$

$$.1 + \frac{\gamma^2}{1-\gamma} = \gamma. \tag{F.107}$$

Alternatively, suppose $s_2 = s_1$, and so π continually accumulates $R(s_1) = .1$. Then there again exists an optimal policy shift corresponding to a solution to $\frac{.1}{1-\gamma} = \gamma$.

By construction, these two scenarios are the only ways in which π might accrue reward, and so an optimal policy shift occurs for R.

If $s'_2 \in Ch\,(s_0)$, then set $R(s_1) \coloneqq 1$, $R(s'_1) \coloneqq .99$, $R(s'_2) \coloneqq .9$, and 0 elsewhere. Suppose that s_1 can reach itself in two steps (the soonest possible, as $s_1 \notin Ch\,(s_1)$), while neither s'_1 or s'_2 can reach themselves or s_1. The corresponding equation $\frac{1}{1-\gamma^2} = .99 + .9\gamma$ has a solution in the open unit interval. Therefore, a shift occurs even under these maximally conservative assumptions. □

Conjecture F.129 (If some reward function has optimal policy shifts, then almost all reward functions have optimal policy shifts)**.**

Proposition F.130 (Sufficient conditions for a reward function not having optimal policy shifts)**.** *Let $R \in \mathbb{R}^{\mathcal{S}}$.*

1. *If R assigns reward r to all states, or*
2. *If the environment is deterministic and R is a state indicator reward function, or*
3. *If the environment is deterministic and R is assigns reward r_1 to states which can reach themselves and r_2 to states which cannot reach themselves, or*
4. *If the environment is deterministic and R is assigns reward r_1 to some set of 1-cycle states and r_2 to all other states, or*
5. *If $R = mR' + b$ for some $m > 0, b \in \mathbb{R}$ and $R' \in \mathbb{R}^{\mathcal{S}}$ which has no optimal policy shifts,*

then R has no optimal policy shifts.

Proof. Item 1: If R assigns reward r to all states, then all policies are optimal at all discount rates.

Item 2: Let s be a state and R_s be the indicator reward function for s. At each state, let π choose an action which minimizes graph distance to s. Such an action exists because there are only finitely many actions. Since $R(s) = 1$ and 0 elsewhere, π is optimal for R_s for all discount rates.

Item 3: If $r_1 = r_2$, then apply item 1.

Suppose that $r_1 > r_2$. Then starting from any state s, each policy π induces a trajectory which deterministically visits $0 \leq \ell < |\mathcal{S}|$ transient states before entering a k-cycle ($0 < k \leq \mathcal{S} - \ell$). By assumption on r_1, r_2, we have $V_R^\pi(s,\gamma) = \frac{1-\gamma^\ell}{1-\gamma} \cdot r_2 + \gamma^\ell \frac{r_1}{1-\gamma}$. Then a policy induces optimal value at s iff it minimizes ℓ; such a policy must exist, since there are only finitely many admissible values of ℓ.

This criterion is independent of the value of $\gamma \in (0, 1)$, and so there are no optimal policy shifts at s. Since s was arbitrary, there are no optimal policy shifts for R. Similar logic proves the $r_1 < r_2$ case, except that optimal policies maximize ℓ.

Item 4: If $r_1 = r_2$, then apply item 1.

Suppose that $r_1 > r_2$ and let s be a state. If s cannot reach any r_1-states, then every action is optimal for R at s at any discount rate.

If s can reach an r_1 state s' via some policy π, then by determinism, π induces a trajectory which deterministically visits $0 \leq \ell < |\mathcal{S}|$ states before reaching s'. By the definition of r_1, r_2, we have $V_R^\pi(s, \gamma) \leq \frac{1-\gamma^\ell}{1-\gamma} \cdot r_2 + \gamma^\ell \frac{r_1}{1-\gamma}$, with equality iff π stays at s'. Then since $r_1 > r_2$, π induces optimal value at s iff it minimizes ℓ and stays at s'; such a policy must exist, since there are only finitely many admissible values of ℓ.

This criterion is independent of the value of $\gamma \in (0, 1)$, and so there are no optimal policy shifts at s. Since s was arbitrary, there are no optimal policy shifts for R. Similar logic proves the $r_1 < r_2$ case, except that optimal policies maximize ℓ and avoid staying at s'.

Item 5: Optimal policy is invariant to positive affine transformation, and so the result follows immediately. $\square$

Figure F.22 shows that the presence of optimal policy shifts is not invariant to negation of the reward function.

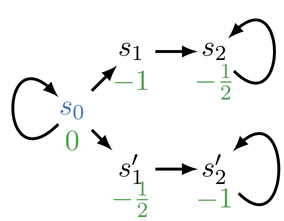


Figure F.22: The given R has no optimal policy shifts, but its inverse $-R$ does have optimal policy shifts.

Figure F.23 shows a counterexample to proposition F.130's item 2 for stochastic environments.

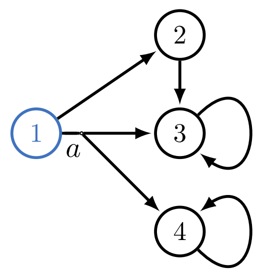


Figure F.23: Optimal policy shifts can occur for state indicator reward functions in stochastic environments, but they cannot occur in deterministic environments (proposition F.130). Let R_{s_3} be the state indicator reward function for s_3, and let $T(s_1, a, s_3) = T(s_1, a, s_4) = \frac{1}{2}$. R_{s_3} has an optimal policy shift at $\gamma = \frac{1}{2}$.

Since we assume state-based reward functions, naively plugging in $\gamma = 0$ would make all policies optimal. Instead, we consider the limiting optimal policy set as $\gamma \to 0$, in a similar manner as Blackwell optimality considers the limiting optimal policy set as $\gamma \to 1$.

Definition F.131 (Asymptotically greedy optimality)**.** $\Pi^*(R, 0)$ is the *asymptotically greedily optimal policy set* for state-based reward function R.

Although, lemma F.15 proved that $\Pi^*(R, 0) \subseteq \Pi^{\text{greedy}}(R)$, fig. F.24 shows that equality need not hold.

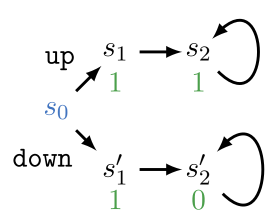


Figure F.24: For all $\gamma > 0$, `up` has greater value than `down`. Therefore, even though both `up` and `down` are greedy actions, only $\pi^{\texttt{up}} \in \Pi^*(R, 0)$.

Definition F.132 (Reward sequence induced by a policy)**.** Let R be a reward function, and let the sequence $\left(\left((\mathbf{T}^{\pi})^t \mathbf{e}_s\right)^\top \mathbf{r}\right)_{t\geq 1}^{s,\pi,R}$ contain the expected undiscounted R-reward at time steps t given that policy π is followed starting from state s.

Proposition F.133 (Dictionary-ordered greediness)**.** *$\pi^* \in \Pi^*(R, 0)$ iff $\forall s \in \mathcal{S}, \pi \in \Pi$: $(r_1, r_2, \ldots)^{s,\pi^*,R} \succeq (r_1, r_2, \ldots)^{s,\pi,R}$, where $\succeq$ is the dictionary ordering over $\mathbb{R}^\infty$.*

Proof. Repeatedly apply lemma F.15 to conclude that $\pi^* \in \Pi^*(R, 0)$ iff it satisfies the dictionary ordering condition at all time steps. □

Lemma F.134 (For almost all reward functions, asymptotically greedy action is determined by expected immediate reward)**.** *For almost all reward functions R, $\pi \in \Pi^*(R, 0)$ iff $\pi \in \Pi^{greedy}(R)$.*

Proof. The forward direction holds for all reward functions as a corollary of proposition F.133. The converse holds for almost all reward functions R, since $\{\mathbf{r} \in \mathbb{R}^{|\mathcal{S}|} \mid \exists s, a : T(s, \pi(s)) \neq T(s, a) \wedge T(s, \pi(s))^\top \mathbf{r} = T(s, a)^\top \mathbf{r}\}$ has no interior in the standard topology on $\mathbb{R}^{|\mathcal{S}|}$ by lemma D.12. Therefore, for almost all R, the policy which maximizes expected next-state return π is unique up to action equivalence. By this uniqueness and by the fact that $\Pi^*(R, 0)$ cannot be empty by lemma D.40, $\pi \in \Pi^*(R, 0)$ for almost all R. □

Proposition F.135 (Child distribution similarity implies equal greedy optimality probability)**.** *If $T(s, a), T(s, a') \in T(s)$ are similar via a permutation ϕ such that $\phi \cdot T_{\text{nd}}(s) = T_{\text{nd}}(s)$, then $\mathbb{P}_{\mathcal{D}_{X\text{-IID}}}(s, a, 0) = \mathbb{P}_{\mathcal{D}_{X\text{-IID}}}(s, a', 0)$.*

Proof.

$$\mathbb{P}_{\mathcal{D}_{X\text{-IID}}}(s, a, 0) \tag{F.108}$$

$$:= \mathop{\mathbb{P}}_{R\sim\mathcal{D}}\left(\exists \pi^* \in \Pi^*(R, 0) : \pi^*(s) = a\right) \tag{F.109}$$

$$= \mathop{\mathbb{E}}_{\mathbf{r}\sim\mathcal{D}}\left[\mathbb{1}_{T(s,a)^\top \mathbf{r} \geq \max_{\mathbf{d}\in T(s)} \mathbf{d}^\top \mathbf{r}}\right] \tag{F.110}$$

$$= \mathop{\mathbb{E}}_{\mathbf{r}\sim\mathcal{D}}\left[\mathbb{1}_{T(s,a)^\top \mathbf{r} \geq \max_{\mathbf{d}\in T_{\text{nd}}(s)} \mathbf{d}^\top \mathbf{r}}\right] \tag{F.111}$$

$$= \mathop{\mathbb{E}}_{\mathbf{r}\sim\mathcal{D}}\left[\mathbb{1}_{T(s,a')^\top \mathbf{r}' \geq \max_{\mathbf{d}\in T_{\text{nd}}(s)} \mathbf{d}^\top \mathbf{r}'}\right] \tag{F.112}$$

$$= \mathop{\mathbb{P}}_{R\sim\mathcal{D}}\left(\exists \pi^* \in \Pi^*\left(R, 0\right) : \pi^*(s) = a'\right) \tag{F.113}$$

$$=: \mathbb{P}_{\mathcal{D}_{X\text{-IID}}}\left(s, a', 0\right). \tag{F.114}$$

Equation (F.109) and eq. (F.113) follow from lemma F.134. Equation (F.110) follows by lemma D.39. Let

$$g(d_1, d_2) \coloneqq \mathbb{1}_{d_1 \geq d_2}$$

and

$$f(B_1, B_2 \mid \mathcal{D}) \coloneqq \mathop{\mathbb{E}}_{\mathbf{r}\sim\mathcal{D}}\left[g(\max_{\mathbf{d}\in B_1} \mathbf{d}^\top \mathbf{r}, \max_{\mathbf{d}'\in B_2} \mathbf{d}'^\top \mathbf{r})\right].$$

Then by lemma D.22, $f\left(\left\{T(s,a)\right\}, T_{\text{nd}}\left(s\right) \mid \mathcal{D}\right) = f\left(\phi \cdot \left\{T(s,a)\right\}, \phi \cdot T_{\text{nd}}\left(s\right) \mid \phi \cdot \mathcal{D}\right) = f\left(\left\{T(s,a')\right\}, T_{\text{nd}}\left(s\right) \mid \mathcal{D}\right)$ (as $\mathcal{D}$ distributes reward identically across states, so $\phi \cdot \mathcal{D} = \mathcal{D}$). Then eq. (F.112) holds. ☐

Corollary F.136 (Equal action optimality probability when $\gamma = 0$)**.** *If the environment is deterministic and $\gamma = 0$, then all actions are equally probably optimal at any given state.*

Proof. In deterministic MDPs, $\forall s : T_{\text{nd}}\left(s\right) = T(s)$ because each child state $s' \in Ch\left(s\right)$ is strictly greedily optimal for the indicator reward function $\mathbb{1}_{s''=s'}$. Furthermore, for any two actions a_1, a_2 leading to children $s_1, s_2 \in Ch\left(s\right)$ respectively, the transposition of s_1, s_2 satisfies the requirements of proposition F.135.

Apply proposition F.135 to conclude a_1, a_2 have equal optimality probability at s when $\gamma = 0$. ☐

Proposition F.137 (No-shift, injective reward functions can be solved greedily)**.** *Let R be a reward function which has no optimal policy shifts such that $\forall s \in \mathcal{S}, a, a' \in \mathcal{A}$, a and a' are both greedily optimal iff $a \equiv_s a'$. Then $\forall \gamma \in [0, 1] : \Pi^*\left(R, \gamma\right) = \Pi^{greedy}\left(R\right)$.*

In deterministic environments, this holds for no-shift reward functions which assign a unique reward to each state.

Proof. By lemma F.15, $\pi \in \Pi^*(R, 0)$ must be greedily optimal. By the next-step reward assumption, each greedily optimal policy is determined up to action equivalence; therefore, if $\pi \in \Pi^*(R, 0)$ and π' is greedily optimal for R, then $\forall s \in \mathcal{S} : \pi(s) \equiv_s \pi'(s)$. This implies that $\pi' \in \Pi^*(R, 0)$. So $\Pi^*(R, 0) = \Pi^{\text{greedy}}(R)$. Since no optimal policy shifts occur, these greedy policies must always be optimal.

In deterministic environments, an injective reward function implies that $\forall s \in \mathcal{S}, a, a' \in \mathcal{A}$, a and a' have equal next-step expected reward iff $a \equiv_s a'$. Injectivity implies that a and a' are both greedily optimal iff $a \equiv_s a'$. □

In fig. F.25, for each $\gamma \in (0, 1)$, there exists a reward function whose optimal policy set has not yet "settled down." However, when $\gamma \approx 1$, "most" reward functions have a Blackwell optimal policy set.

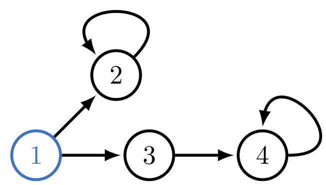


Figure F.25: Let $\gamma \in (0, 1)$, and consider $R(s_1) = R(s_3) := 0$, $R(s_4) := 1$. The optimal policy set is not yet Blackwell-optimal if $R(s_2) \in (\gamma, 1)$.

Conjecture F.138 (There exists a characterization of optimal policy shift existence in stochastic MDPs)**.**

Lemma F.139 (Optimal policy shift bound)**.** *For fixed R, $\Pi^*(R, \gamma)$ can take on at most $(2|\mathcal{S}| - 1)\sum_s \binom{|\mathcal{F}(s)|}{2}$ values over $\gamma \in (0, 1)$.*

Proof. By lemma D.6, $\Pi^*(R, \gamma)$ changes value iff there is a change in optimality status for some visit distribution function at some state. By lemma F.125, each pair of distinct visit distributions can switch optimality status at most $2|\mathcal{S}| - 1$ times. At each state s, there are $\binom{|\mathcal{F}(s)|}{2}$ such pairs. □

Conjecture F.140 (Linear bound on optimal visit distribution function shifts)**.** For any reward function R and $\mathbf{f}, \mathbf{f}' \in \mathcal{F}(s)$, $\mathbf{f}$ and $\mathbf{f}'$ shift at most $|\mathcal{S}| - 1$ times.

Conjecture F.141 (Quadratic upper bound on optimal policy shifts)**.** For any reward function R, at most $\mathcal{O}(|\mathcal{S}|^2)$ optimal policy shifts occur.

F.5 Optimality probability

Figure F.26 serves as a reminder that the relatively greater optimality probability of an action a at γ does not imply that $\mathbb{P}_{\mathcal{D}}(s, a, \gamma) \geq \frac{1}{2}$.

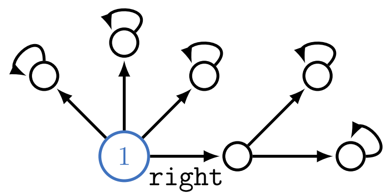


Figure F.26: By theorem 5.29, $\mathbb{P}_{\mathcal{D}_{X\text{-IID}}}(s, \texttt{right}, 1) = \frac{2}{5} < \frac{1}{2}$. In other words, when $\gamma \approx 1$, it's more probable than not that `right` *isn't* optimal, even though `right` is more probable under optimality than other actions.

F.5.1 Impossibility of graphical characterization of which actions are more probably optimal

Even restricting ourselves to $\mathcal{D}_{X\text{-IID}}$ beliefs, we can't always just look at the rewardless MDP structure to determine which actions are more probable under optimality. In fig. F.27, going `up` is more probable under optimality at s_1 for some state reward distributions X, but not for others.

Theorem F.142 (The state reward distribution can affect which actions have the greatest optimality probability)**.** *There can exist state reward distributions* X_1, X_2*, a state* s*, and a discount rate* γ *for which* $\arg\max_a \mathbb{P}_{\mathcal{D}_{X_1\text{-IID}}}(s, a, \gamma) \neq \arg\max_a \mathbb{P}_{\mathcal{D}_{X_2\text{-IID}}}(s, a, \gamma)$.

Proof. See fig. F.27. ☐

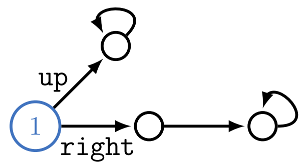


Figure F.27: For X_1 uniform, up and right are equally probable under optimality. For X_2 with CDF $F(x) := x^2$ on the unit interval, $\mathbb{P}_{\mathcal{D}_{X_2\text{-IID}}}(s_1, \texttt{up}, \gamma) = \frac{10+3\gamma-3\gamma^2}{20}$.

F.5.2 Sample means

For arbitrary $D, D' \subseteq \mathrm{RSD}_{\mathrm{nd}}(s)$, determining if $\mathbb{P}_{\mathcal{D}}(D,1) > \mathbb{P}_{\mathcal{D}}(D',1)$ is at least as hard as answering questions like "for sample means $\bar{x}_n$ of n IID draws from an arbitrary continuous bounded distribution X, is $\mathbb{P}\left(\max(\bar{x}_4, \bar{x}_4', \bar{x}_5) > \max(\bar{x}_4'', \bar{x}_5', \bar{x}_5'')\right) > \frac{1}{2}$?". These questions about maximal order statistics are often difficult to answer.

Thus, there is no simple characterization of the $D, D' \subseteq \mathrm{RSD}_{\mathrm{nd}}(s)$ for which

$$\mathbb{P}_{\mathcal{D}}(D,1)\,[\mathcal{D}_{X\text{-IID}}] > \mathbb{P}_{\mathcal{D}}\left(D',1\right)[\mathcal{D}_{X\text{-IID}}].$$

However, there may be X for which k-cycle optimality probability decreases as k increases.

Conjecture F.143 (Increased sample size decreases maximality probability [94])**.** Consider a finite set of sample means $\bar{x}_i$ of k_i draws from unif$(0,1)$. If $k_i > k_j$, then $\mathbb{P}(\bar{x}_i = \max_l \bar{x}_l) < \mathbb{P}\left(\bar{x}_j = \max_l \bar{x}_l\right)$.

Corollary. *Suppose the environment is deterministic. Let $k > k'$ and $\mathbf{d}_k, \mathbf{d}_{k'} \in \mathrm{RSD}_{\mathrm{nd}}(s)$ be k, k'-cycles, respectively. Suppose that*

$$\left\|\{\mathbf{d}_k\} - \mathrm{RSD}_{\mathrm{nd}}(s) \setminus \{\mathbf{d}_k\}\right\|_1 = \left\|\{\mathbf{d}_{k'}\} - \mathrm{RSD}_{\mathrm{nd}}(s) \setminus \{\mathbf{d}_{k'}\}\right\|_1 = 2.$$

For X uniform, $\mathbb{P}_{\mathcal{D}}(\mathbf{d}_k, 1) < \mathbb{P}_{\mathcal{D}}(\mathbf{d}_{k'}, 1)$.

F.5.3 Optimality probability of linear functionals

Conjecture F.144 (Optimality probability changes continuously under $\mathcal{D}_{\text{cont}}$)**.** Let $D \subseteq \mathbb{R}$ and let G, H be two finite sets of continuous functions (or paths) from D to $\mathbb{R}^{|\mathcal{S}|}$. Define $G(\gamma) := \left\{g(\gamma) \mid g \in G\right\}$, and similarly for $H(\gamma)$. $p_{\mathcal{D}_{\text{cont}}}\left(G(\gamma) \geq H(\gamma)\right)$ is continuous on $\gamma \in X$.

F.5.4 Properties of optimality probability

Corollary F.145 (Almost all reward functions don't have an optimal policy shift at any given γ)**.** *For any* $\gamma \in (0,1)$, $\mathbb{P}_{R \sim \mathcal{D}_{cont}}\left(\lim_{\gamma^- \uparrow \gamma} \Pi^*\left(R, \gamma^-\right) \neq \Pi^*\left(R, \gamma\right)\right) = 0$.

Proof. Let $\gamma \in (0,1)$. By lemma F.127, if $R \in \mathbb{R}^{\mathcal{S}}$ has an optimal policy shift at γ, then $\exists s \in \mathcal{S} : \left|\arg\max_{\mathbf{f} \in \mathcal{F}(s)} \mathbf{f}(\gamma)^\top \mathbf{r}\right| \geq 2$. At least one such $\mathbf{f} \in \mathcal{F}_{\text{nd}}(s)$ by lemma D.15 and lemma D.38. Let $\mathbf{f}' \in \mathcal{F}(s)$ be a distinct element of the arg max.

By lemma D.52, $\mathbf{f}(\gamma) \neq \mathbf{f}'(\gamma)$. Then $\left|\arg\max_{\mathbf{d} \in \mathcal{F}(s,\gamma)} \mathbf{d}^\top \mathbf{r}\right| \geq 2$. Corollary D.13 shows that $\mathbb{P}_{\mathbf{r} \sim \mathcal{D}_{\text{cont}}}\left(\left|\arg\max_{\mathbf{f} \in \mathcal{F}(s)} \mathbf{f}(\gamma)^\top \mathbf{r}\right| \geq 2\right) = 0$. ∎

Lemma F.146 (For continuous reward function distributions, optimality probability is additive over visit distribution functions)**.** *Let* $F \subseteq \mathcal{F}(s)$. $\mathbb{P}_{\mathcal{D}_{cont}}\left(F, \gamma\right) = \sum_{\mathbf{f} \in F} \mathbb{P}_{\mathcal{D}_{cont}}\left(\mathbf{f}, \gamma\right)$.

Proof. Suppose $\gamma \in (0,1)$. Since $\mathcal{D}_{\text{cont}}$ is continuous, apply proposition F.63 with $A := F(\gamma), B := \mathcal{F}(s, \gamma)$.

Since the result holds for $\gamma \in (0,1)$, it holds in the limits of $\gamma \to 0$ and $\gamma \to 1$. ∎

Lemma F.147 (Optimality probability is continuous on γ)**.** *For* $F \subseteq \mathcal{F}(s)$, $\mathbb{P}_{\mathcal{D}_{cont}}\left(F, \gamma\right)$ *is continuous on* $\gamma \in [0,1]$.

Proof. Since lemma F.146 shows that $\mathbb{P}_{\mathcal{D}_{\text{cont}}}\left(F, \gamma\right) = \sum_{\mathbf{f} \in F} \mathbb{P}_{\mathcal{D}_{\text{cont}}}\left(\mathbf{f}, \gamma\right)$, it is sufficient to show that each summand is continuous.

Let $\mathbf{f} \in F$. If $\mathbb{P}_{\mathcal{D}_{\text{cont}}}\left(\mathbf{f}, \gamma\right)$ were discontinuous at $\gamma^* \in (0,1)$, then a positive measure subset of $\mathcal{D}_{\text{cont}}$ experiences an optimal policy shift at γ^*. This contradicts corollary F.145, and so $\mathbb{P}_{\mathcal{D}_{\text{cont}}}(\mathbf{f}, \gamma)$ must be continuous on $\gamma \in (0,1)$. Proposition D.41 shows that optimality probability's limits exist, and so $\mathbb{P}_{\mathcal{D}_{\text{cont}}}(\mathbf{f}, \gamma)$ is actually continuous on $\gamma \in [0,1]$. $\square$

Conjecture F.148 (Finite disagreement of optimality probability for $\mathcal{D}_{X\text{-IID}}$)**.** For any $\mathbf{f}, \mathbf{f}' \in \mathcal{F}(s)$, $\mathbb{P}_{\mathcal{D}_{X\text{-IID}}}(\mathbf{f}, \gamma) = \mathbb{P}_{\mathcal{D}_{X\text{-IID}}}(\mathbf{f}', \gamma)$ either for all $\gamma \in [0,1]$ or for finitely many γ.

Lemma F.149 (Only $\mathbf{f} \in \mathcal{F}_{\text{nd}}(s)$ have positive optimality probability at any γ)**.** $\mathbf{f} \in \mathcal{F}_{\text{nd}}(s)$ *iff* $\exists \gamma \in (0,1) : \mathbb{P}_{\mathcal{D}_{X\text{-IID}}}(\mathbf{f}, \gamma) > 0$.

Proof. If $\mathbf{f} \in \mathcal{F}_{\text{nd}}(s)$, then $\exists \gamma \in (0,1), \mathbf{r}_1 \in \mathbb{R}^{|\mathcal{S}|} : \mathbf{f}(\gamma)^\top \mathbf{r}_1 > \max_{\mathbf{f}' \in \mathcal{F}(s) \setminus \{\mathbf{f}\}} \mathbf{f}'(\gamma)^\top \mathbf{r}_1$. Then $\mathbb{P}_{\mathcal{D}_{X\text{-IID}}}(\mathbf{f}, \gamma) = p_{\mathcal{D}_{X\text{-IID}}}\left(\mathbf{f}(\gamma) \geq \mathcal{F}(s, \gamma)\right) > 0$ by proposition D.25.

If $\exists \gamma \in (0,1) : \mathbb{P}_{\mathcal{D}_{X\text{-IID}}}(\mathbf{f}, \gamma) > 0$, then $\mathbf{f}(\gamma) \in \text{ND}\left(\mathcal{F}(s, \gamma)\right)$ by lemma F.64. This implies $\mathbf{f} \in \mathcal{F}_{\text{nd}}(s)$ by lemma D.38. $\square$

Corollary F.150 (Dominated visit distributions are almost never optimal)**.** $\mathbf{f} \in \mathcal{F}(s)$ *is dominated iff* $\forall \gamma \in [0,1] : \mathbb{P}_{\mathcal{D}_{cont}}(\mathbf{f}, \gamma) = 0$.

Proof. Let $\gamma \in (0,1)$. If $\mathbf{f} \in \mathcal{F}(s) \setminus \mathcal{F}_{\text{nd}}(s)$, then $\mathbb{P}_{\mathcal{D}_{\text{cont}}}(\mathbf{f}, \gamma) = 0$ by lemma F.149. Furthermore, $\mathbb{P}_{\mathcal{D}_{\text{cont}}}(\mathbf{f}, 0) := \lim_{\gamma \to 0} \mathbb{P}_{\mathcal{D}_{\text{cont}}}(\mathbf{f}, \gamma) = 0$. Similar reasoning applies when $\gamma = 1$.

We show that $\forall \gamma \in [0,1] : \mathbb{P}_{\mathcal{D}_{\text{cont}}}(\mathbf{f}, \gamma) = 0$ implies that $\mathbf{f} \in \mathcal{F}(s) \setminus \mathcal{F}_{\text{nd}}(s)$ by proving the contrapositive. Suppose $\exists \gamma \in (0,1) : \mathbb{P}_{\mathcal{D}_{\text{cont}}}(\mathbf{f}, \gamma) \neq 0$. Then by lemma F.149, $\mathbf{f}$ cannot be dominated, and so $\mathbf{f} \in \mathcal{F}_{\text{nd}}(s)$.

If $\mathbb{P}_{\mathcal{D}_{\text{cont}}}(\mathbf{f}, 0) := \lim_{\gamma \to 0} \mathbb{P}_{\mathcal{D}_{\text{cont}}}(\mathbf{f}, \gamma) \neq 0$, then the optimality probability must be non-zero in a neighborhood of 0. By lemma F.149, this can only be true if $\mathbf{f} \in \mathcal{F}_{\text{nd}}(s)$. Similar reasoning applies to the $\gamma = 1$ case. $\square$

Proposition F.151 (Non-domination iff positive measure)**.** $\mathbf{f} \in \mathcal{F}_{\text{nd}}(s)$ *iff* $\forall \gamma \in (0,1) : \mathbb{P}_{\mathcal{D}_{X\text{-IID}}}(\mathbf{f}, \gamma) > 0$.

Proof. By the same arguments used in lemma F.149's proof's forward direction, $\mathbb{P}_{\mathcal{D}_{X\text{-IID}}}(\mathbf{f}, \gamma) > 0$. Therefore, $\forall \gamma \in (0,1) : \mathbb{P}_{\mathcal{D}_{X\text{-IID}}}(\mathbf{f}, \gamma) > 0$.

If $\forall \gamma \in (0,1) : \mathbb{P}_{\mathcal{D}_{X\text{-IID}}}(\mathbf{f}, \gamma) > 0$, apply corollary F.150 to conclude that $\mathbf{f}$ cannot be dominated (corollary F.150 applies because $\mathcal{D}_{X\text{-IID}}$ is continuous), and so $\mathbf{f} \in \mathcal{F}_{\text{nd}}(s)$. $\square$

Proposition F.152 (Non-domination iff positive probability for $\gamma \in [0,1)$)**.**

$$\mathbf{f} \in \mathcal{F}_{\text{nd}}(s) \textit{ iff } \forall \gamma \in [0,1) : \mathbb{P}_{\mathcal{D}_{X\text{-IID}}}(\mathbf{f}, \gamma) > 0.$$

Proof. The case for $\gamma \in (0,1)$ is proved by proposition F.151.

For the $\gamma = 0$ case, suppose $\mathbf{f} \in \mathcal{F}_{\text{nd}}(s)$ and fix $\gamma^* \in (0,1)$. By corollary F.16, almost all reward functions in the interior of $\text{supp}(\mathbf{f}(\gamma^*) \geq \mathcal{F}(s, \gamma^*)) \cap \text{supp}(\mathcal{D})$ have optimal actions at each state which are unique up to action equivalence. Let R be one such reward function for which $\mathbf{f} \in \mathcal{F}_{\text{nd}}(s)$ is strictly optimal.

By proposition F.18, $\Pi^*\left(V^*_R(\cdot, \gamma^*), 0\right) = \Pi^*(R, \gamma^*)$. Since optimal value is continuous on the reward function (lemma F.99) and since $\mathbf{f}$ is strictly optimal for R, there exists an ϵ-ball of reward functions around the *reward function* $V^*_R(\cdot, \gamma^*)$ for which $\mathbf{f}$ is strictly optimal. By lemma D.20, we can positively affinely transform this ball to be contained in $\text{supp}(\mathbf{f}, 0) \cap \text{supp}(\mathcal{D})$; since this is a positive affine transformation, the image has non-empty interior. Then $\mathbb{P}_{\mathcal{D}_{X\text{-IID}}}(\mathbf{f}, 0) > 0$ since $\mathcal{D}_{X\text{-IID}}$ is a continuous distribution.

Suppose $\mathbf{f} \in \mathcal{F}(s) \setminus \mathcal{F}_{\text{nd}}(s)$. By lemma F.106, for any $\gamma^* \in (0,1)$, the set of reward functions with multiple optimal visit distributions has measure zero. Dominated visit distribution functions cannot be uniquely optimal at any $\gamma^* \in (0,1)$. Therefore, if $\mathbf{f}$ is dominated, $\forall \gamma^* \in (0,1) : \mathbb{P}_{\mathcal{D}_{X\text{-IID}}}(\mathbf{f}, \gamma^*) = 0$, and $\mathbb{P}_{\mathcal{D}_{X\text{-IID}}}(\mathbf{f}, 0) := \lim_{\gamma^* \to 0} \mathbb{P}_{\mathcal{D}_{X\text{-IID}}}(\mathbf{f}, \gamma^*) = \lim_{\gamma^* \to 0} 0 = 0$. So for any $\gamma \in [0,1)$, $\mathbb{P}_{\mathcal{D}_{X\text{-IID}}}(\mathbf{f}, \gamma) > 0$ implies $\mathbf{f} \in \mathcal{F}_{\text{nd}}(s)$. $\square$

Lemma F.153 (Non-dominated child distributions facilitate a non-dominated visit distribution function)**.** *If* $T(s,a) \in T_{\text{nd}}(s)$*, then* $\exists \pi : \pi(s) = a$ *and* $\mathbf{f}^\pi \in \mathcal{F}_{\text{nd}}(s)$*.*

Proof. Let $\mathbf{d} := T(s,a)$ and let $V := \left\{\mathbf{v} \in \mathbb{R}^{|\mathcal{S}|} \mid \mathbf{d}^\top > \max_{\mathbf{d}' \in T(s) \setminus \{\mathbf{d}'\}} \mathbf{d}'^\top \mathbf{v}\right\}$. V is nonempty because $\mathbf{d} \in T_{\text{nd}}(s)$ and has positive Lebesgue measure by the continuity of $\mathbf{d}^\top \mathbf{v}$ and the max on $\mathbf{v} \in \mathbb{R}^{|\mathcal{S}|}$.

For all $\pi \in \Pi$, let $V_\pi := \left\{\mathbf{v} \in V \mid \pi \in \Pi^{\text{greedy}}(\mathbf{v})\right\}$. Since Π is finite, there must exist π^* such that $V_{\pi^*} \subseteq V$ has positive Lebesgue measure. Then consider the set of reward functions $R_{\pi^*} := \left\{\left(\mathbf{I} - .5\mathbf{T}^{\pi^*}\right)\mathbf{v} \mid \mathbf{v} \in V_{\pi^*}\right\}$. Since $\left(\mathbf{I} - .5\mathbf{T}^{\pi^*}\right)$ is invertible because its spectral radius is strictly less than 1 (see the proof of lemma F.1), $\det\left(\mathbf{I} - .5\mathbf{T}^{\pi^*}\right) \neq 0$ and so R_{π^*} also has positive Lebesgue measure.

But then almost all reward functions $\mathbf{r}$ in R_{π^*} must have a unique optimal visit distribution $\mathbf{f}^{\pi'} \in \mathcal{F}(s)$ when $\gamma = .5 \in (0,1)$ by lemma F.106. Therefore, $\mathbf{f}^{\pi'} \in \mathcal{F}_{\text{nd}}(s)$ by definition 5.6. Lastly, since $\forall \mathbf{r}' \in R_{\pi^*} : \Pi^*\left(\mathbf{r}', .5\right) = \Pi^{\text{greedy}}\left(\mathbf{v}'\right)$ (where each $\mathbf{r}' := \left(\mathbf{I} - .5\mathbf{T}^{\pi^*}\right)\mathbf{v}'$), $\pi'(s) \equiv_s a$ by the definitions of V and $\Pi^{\text{greedy}}\left(\mathbf{v}'\right)$.

Let π equal π', except that $\pi(s) = a$. Because $\pi'(s) \equiv_s a = \pi(s)$, $\mathbf{f}^\pi = \mathbf{f}^{\pi'} \in \mathcal{F}_{\text{nd}}(s)$. We have thus shown that $\exists \pi \in \Pi : \pi(s) = a$ and $\mathbf{f}^\pi \in \mathcal{F}_{\text{nd}}(s)$. $\square$

Proposition F.154 ($\mathcal{F}_{\text{nd}}(s)$ controls optimality probability)**.** *Let $F \subseteq \mathcal{F}(s)$. $\mathbb{P}_{\mathcal{D}_{cont}}(F, \gamma) = \mathbb{P}_{\mathcal{D}_{cont}}\left(F \cap \mathcal{F}_{\text{nd}}(s), \gamma\right)$.*

Proof.

$$\mathbb{P}_{\mathcal{D}_{\text{cont}}}(F, \gamma) = \sum_{\mathbf{f} \in F} \mathbb{P}_{\mathcal{D}_{\text{cont}}}(\mathbf{f}, \gamma) \tag{F.115}$$

$$= \sum_{\mathbf{f} \in F \cap \mathcal{F}_{\text{nd}}(s)} \mathbb{P}_{\mathcal{D}_{\text{cont}}}(\mathbf{f}, \gamma) \tag{F.116}$$

$$= \mathbb{P}_{\mathcal{D}_{\text{cont}}}\left(F \cap \mathcal{F}_{\text{nd}}(s), \gamma\right). \tag{F.117}$$

Equation (F.115) and eq. (F.117) follow by lemma F.146 since $\mathcal{D}_{\text{cont}}$ is continuous. Equation (F.116) follows by corollary F.150, since dominated visit distribution functions have 0 optimality probability under continuous reward function distributions. $\square$

Lemma F.155 (Only non-dominated transitions are greedily optimal with positive probability)**.** *Let $T(s,a) \in T(s)$. $T(s,a) \in T_{\text{nd}}(s)$ iff $\mathbb{P}_{\mathcal{D}_{X\text{-IID}}}(s, a, 0) > 0$.*

Proof. Suppose that $T(s,a) \in T_{\text{nd}}(s)$. This means that $\exists \pi \in \Pi : \pi(s) = a$ and $\mathbf{f}^\pi \in \mathcal{F}_{\text{nd}}(s)$ by lemma F.153, and so $\mathbb{P}_{\mathcal{D}_{X\text{-IID}}}(\mathbf{f}^\pi, 0) > 0$ by proposition F.152. Then $\mathbb{P}_{\mathcal{D}_{X\text{-IID}}}(s, a, 0) \geq$

$\mathbb{P}_{\mathcal{D}_{X\text{-IID}}}(\mathbf{f}^\pi, 0) > 0$.

Suppose that $\mathbb{P}_{\mathcal{D}_{X\text{-IID}}}(s, a, 0) > 0$. Proposition F.152 implies that $\mathbb{P}_{\mathcal{D}_{X\text{-IID}}}(s, a, 0) > 0$ iff $\exists \pi \in \Pi : \mathbf{f}^\pi \in \mathcal{F}_{\text{nd}}(s)$ such that $\pi(s) = a$, since $\mathbb{P}_{\mathcal{D}_{X\text{-IID}}}(s, a, 0) = \mathbb{P}_{\mathcal{D}_{X\text{-IID}}}\big(\mathcal{F}(s \mid \pi(s) = a), 0\big)$ by lemma F.30.

Lemma F.8 shows that if $T(s, a) \in T(s) \setminus T_{\text{nd}}(s)$ and $\pi(s) = a$, then $\mathbf{f}^\pi \in \mathcal{F}(s) \setminus \mathcal{F}_{\text{nd}}(s)$. The contrapositive is then: if $\mathbf{f}^\pi \in \mathcal{F}_{\text{nd}}(s)$, then either $\pi(s) \neq a$ or $T(s, a) \in T_{\text{nd}}(s)$. But $\pi(s) = a$, so the fact that $\mathbb{P}_{\mathcal{D}_{X\text{-IID}}}(s, a, 0) > 0$ implies that $\mathbf{f}^\pi \in \mathcal{F}_{\text{nd}}(s)$, which implies that $T(s, a) \in T_{\text{nd}}(s)$. ☐

Corollary F.156 (Similarity to a dominated action implies domination)**.** *If $T(s, a) \in T(s) \setminus T_{\text{nd}}(s)$ is similar to $T(s, a') \in T(s)$ via a permutation ϕ such that $\phi \cdot T_{\text{nd}}(s) = T_{\text{nd}}(s)$, then $T(s, a') \notin T_{\text{nd}}(s)$.*

Proof. Apply proposition F.135 to conclude that $\mathbb{P}_{\mathcal{D}_{X\text{-IID}}}(s, a, 0) = \mathbb{P}_{\mathcal{D}_{X\text{-IID}}}\big(s, a', 0\big)$. By lemma F.155, $\mathbb{P}_{\mathcal{D}_{X\text{-IID}}}(s, a, 0) = 0$. Then $\mathbb{P}_{\mathcal{D}_{X\text{-IID}}}\big(s, a', 0\big) = 0$; apply lemma F.155 to conclude that $T(s, a') \notin T_{\text{nd}}(s)$. ☐

Question F.157 (Is there anything to be gained by formalizing the optimality probability of sets of policies?)**.**

Definition F.158 ($\mathcal{F}_{\text{nd}}$ multi-state restriction)**.** $\mathcal{F}(s \mid \pi^*, S) \subseteq \mathcal{F}(s)$ contains the non-dominated visit distributions induced by a policy which agrees with π on the states of $S \subseteq \mathcal{S}$. $\mathcal{F}_{\text{nd}}(s \mid \pi^*, S) \coloneqq \mathcal{F}_{\text{nd}}(s \mid \pi^*, S) \cap \mathcal{F}_{\text{nd}}(s)$.

Proposition F.159 (Optimality probability factorizes)**.** *Let $\mathbf{f}^\pi \in \mathcal{F}_{\text{nd}}(s)$ and let $\gamma \in [0, 1)$.*

$$\mathbb{P}_{\mathcal{D}_{X\text{-IID}}}(\mathbf{f}^\pi, \gamma) = \prod_{i=1}^{|\mathcal{S}|} \frac{\mathbb{P}_{\mathcal{D}_{X\text{-IID}}}\big(\mathcal{F}_{\text{nd}}(s \mid \pi^*, \{s_1, \ldots, s_{i-1}, s_i\}), \gamma\big)}{\mathbb{P}_{\mathcal{D}_{X\text{-IID}}}\big(\mathcal{F}_{\text{nd}}(s \mid \pi^*, \{s_1, \ldots, s_{i-1}\}), \gamma\big)}. \tag{F.118}$$

Proof. Equation (F.118) holds for any state indexing and is well-defined on $\gamma \in [0, 1)$ because $\mathbf{f}^\pi \in \mathcal{F}_{\text{nd}}(s)$ has $\mathbb{P}_{\mathcal{D}_{X\text{-IID}}}(\mathbf{f}, \gamma) > 0$ on that domain (proposition F.151). ☐

Definition F.160 (Optimality probability factorization)**.** The *optimality probability factorization of* $\mathbf{f}^\pi \in \mathcal{F}_{\text{nd}}(s)$ is the RHS of eq. (F.118).

Lemma F.161 (Optimality probability varies iff a factor varies)**.** *Let* $\mathbf{f} \in \mathcal{F}_{\text{nd}}(s)$*.* $\mathbb{P}_{\mathcal{D}_{X\text{-IID}}}(\mathbf{f}, \gamma)$ *varies with* γ *iff its factorization has a factor which varies with* γ*.*

Proof. The forward direction is trivial. For the backward direction, suppose factor $i = k$ equals the non-constant function $f(\gamma)$; note that $\forall \gamma \in [0, 1) : f(\gamma) > 0$. Suppose that the product of other factors equals $\frac{c}{f(\gamma)}$. If one action becomes more likely at s_k, then another action must become less likely, and vice versa. $\frac{1}{f(\gamma)}$ cannot be the multiplicative inverse (up to a constant) for both of these variations, and so $\mathbb{P}_{\mathcal{D}_{X\text{-IID}}}\left(\mathbf{f}', \gamma\right)$ must vary with γ for some $\mathbf{f}' \in \mathcal{F}_{\text{nd}}(s)$. $\square$

Conjecture F.162 (In deterministic environments, constant optimality probability implies rational probabilities)**.** Suppose the MDP is deterministic and s is such that $\forall \mathbf{f} \in \mathcal{F}(s) : \mathbb{P}_{\mathcal{D}_{X\text{-IID}}}(\mathbf{f}, \gamma)$ does not vary with $\gamma \in [0, 1]$. Then $\forall \mathbf{f} \in \mathcal{F}(s) : \mathbb{P}_{\mathcal{D}_{X\text{-IID}}}(\mathbf{f}, \gamma) \in [0, 1] \cap \mathbb{Q}$.

Proposition F.163 (Positive optimality probability under $\mathcal{D}_{\text{cont}}$ implies $\mathcal{F}_{\text{nd}}$ membership and $\text{RSD}_{\text{nd}}(s)$ membership)**.** $\mathbb{P}_{\mathcal{D}_{cont}}(\mathbf{f}, 1) > 0$ *implies* $\mathbf{f} \in \mathcal{F}_{\text{nd}}(s)$ *and* $\lim_{\gamma \to 1}(1-\gamma)\mathbf{f}(\gamma) \in \text{RSD}_{\text{nd}}(s)$.

Proof. Suppose $\mathbb{P}_{\mathcal{D}}(\mathbf{f}, 1) > 0$. By corollary F.150, $\mathbf{f}$ cannot be dominated, else

$$\lim_{\gamma \to 1} \mathbb{P}_{\mathcal{D}}\left((1-\gamma)\mathbf{f}(\gamma), 1\right) = 0.$$

Therefore, $\mathbf{f} \in \mathcal{F}_{\text{nd}}(s)$. By proposition D.25, $\lim_{\gamma \to 1}(1-\gamma)\mathbf{f}(\gamma) \in \text{RSD}_{\text{nd}}(s)$. $\square$

Conjecture F.164 ($\text{RSD}_{\text{nd}}(s)$ membership and $\mathcal{F}_{\text{nd}}$ membership implies positive IID optimality probability)**.** If $\mathbf{f} \in \mathcal{F}_{\text{nd}}(s)$ and $\lim_{\gamma \to 1}(1-\gamma)\mathbf{f}(\gamma) \in \text{RSD}_{\text{nd}}(s)$, then $\mathbb{P}_{\mathcal{D}_{X\text{-IID}}}(\mathbf{f}, 1) > 0$.

Conjecture F.165 (Optimality probabilities reach ordinal equilibrium as $\gamma \to 1$)**.** For any $\mathbf{f}, \mathbf{f}' \in \mathcal{F}(s)$, $\mathbb{P}_{\mathcal{D}_{X\text{-IID}}}(\mathbf{f}, \gamma)$ and $\mathbb{P}_{\mathcal{D}_{X\text{-IID}}}\left(\mathbf{f}', \gamma\right)$ reach ordinal equilibrium as $\gamma \to 1$.

Conjecture F.166 (Each non-dominated visit distribution "takes" optimality probability from all other non-dominated visit distributions)**.** Let $\mathbf{f}, \mathbf{f}' \in \mathcal{F}_{\text{nd}}(s)$, where $\mathbf{f} \neq \mathbf{f}'$. Then

$\forall \gamma \in (0,1)$:

$$\mathbb{P}_{\mathcal{D}_{X\text{-IID}}}(\mathbf{f},\gamma) < \mathbb{P}_{\mathbf{r}\sim\mathcal{D}_{X\text{-IID}}}\left(\mathbf{f}(\gamma)^\top \mathbf{r} = \max_{\mathbf{f}''\in\mathcal{F}_{\text{nd}}(s)\setminus\{\mathbf{f}'\}} \mathbf{f}''(\gamma)^\top \mathbf{r}\right). \tag{F.119}$$

F.5.5 Properties of instrumental convergence

Instrumental convergence exists when actions have different optimality probabilities.

Definition F.167 (Existence of instrumental convergence)**.** *Instrumental convergence starting at state s* when $\exists a, a' \in \mathcal{A}, \gamma \in [0,1] : \mathbb{P}_{\mathcal{D}}(s,a,\gamma) \neq \mathbb{P}_{\mathcal{D}}(s,a',\gamma)$.

Conjecture F.168 (Instrumental convergence exists at almost all discount rates, if it exists)**.** If instrumental convergence exists starting at state s for some γ, it exists starting at s for almost all $\gamma \in [0,1]$.

Lemma F.169 (Optimality probability is unaffected by unreachable states)**.** *If s cannot reach s', then $\forall a : \mathbb{P}_{\mathcal{D}_{any}}\left(\mathcal{F}(s \mid \pi(s') = a), \gamma\right) = 1$.*

Proof. Policies can output any action at unreachable states without affecting the induced visit distribution. □

Proposition F.170 (Instrumental convergence without domination or stochasticity implies IID optimality probability varies with γ)**.** *In deterministic environments, if instrumental convergence exists at some γ^* starting from s and if $\mathcal{F}(s) = \mathcal{F}_{\text{nd}}(s)$, then $\exists \mathbf{f} : \mathbb{P}_{\mathcal{D}_{X\text{-IID}}}(\mathbf{f},\gamma)$ varies with γ.*

Proof. Let a be more probably optimal than a' at state s' (thus fulfilling definition 5.10). Since the environment is deterministic and $\mathcal{F}(s) = \mathcal{F}_{\text{nd}}(s)$, it is equally probable that each child of s' is the greedy choice at $\gamma = 0$. Then

$$\mathbb{P}_{\mathcal{D}_{X\text{-IID}}}\left(\mathcal{F}_{\text{nd}}(s \mid \pi(s') = a), 0\right) = \mathbb{P}_{\mathcal{D}_{X\text{-IID}}}\left(\mathcal{F}_{\text{nd}}(s \mid \pi(s') = a'), 0\right).$$

Therefore, at least one $\mathbb{P}_{\mathcal{D}_{X\text{-IID}}}(\mathbf{f}, \gamma)$ must vary on $\gamma \in [0, \gamma^*]$ so that

$$\mathbb{P}_{\mathcal{D}_{X\text{-IID}}}\left(\mathcal{F}_{\text{nd}}(s \mid \pi(s') = a), \gamma^*\right) > \mathbb{P}_{\mathcal{D}_{X\text{-IID}}}\left(\mathcal{F}_{\text{nd}}(s \mid \pi(s') = a'), \gamma^*\right).$$

□

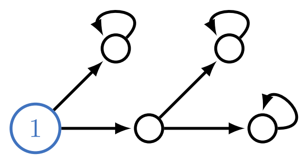


Figure F.28: Our ongoing assumption of X's continuity is required for proposition F.170. Under the uniform distribution on $\{0,1\}^{\mathcal{S}}$, the visit distribution going up from s_1 has probability $\frac{10}{24}$, while the other two visit distributions have probability $\frac{7}{24}$. However, under the uniform distribution on $[0,1]^{\mathcal{S}}$, the upwards visit distribution has probability $\frac{3-\gamma}{6}$, while the other two each have probability $\frac{3+\gamma}{12}$.

Proposition F.170 shows the following while making the strong assumption that $\mathcal{F}(s) = \mathcal{F}_{\text{nd}}(s)$.

Conjecture F.171 (In deterministic environments, instrumental convergence implies variable optimality probability)**.** In deterministic environments, if instrumental convergence exists at some γ^* starting from s, then $\exists \mathbf{f} : \mathbb{P}_{\mathcal{D}_{X\text{-IID}}}(\mathbf{f}, \gamma)$ varies with γ.

Conjecture F.172 (Optimal policy shifts necessary for instrumental convergence)**.** In deterministic environments, if optimal policy shifts cannot occur, then instrumental convergence does not exist.

Remark. Proposition F.170's assumption of determinism is required: suppose state s has $Ch(s) = \{s_1, s_2, s_3\}$ and $|\mathcal{A}| = 3$. $T(s, a_1) = \mathbf{e}_{s_1}$, $T(s, a_2) = (0, .5, .5)^\top$, and $T(s, a_3) = \mathbf{e}_{s_3}$. When X is uniform and $\gamma \approx 0$, action a_1 will be strictly more probably optimal than actions a_2 or a_3.

Proposition F.173 (No domination, stochasticity, or optimal policy shifts means equal optimality probabilities)**.** *In deterministic environments where optimal policy shifts cannot occur, if $\mathcal{F}(s) = \mathcal{F}_{\text{nd}}(s)$, then no instrumental convergence exists starting from s.*

Proof. Since the environment is deterministic and $\mathcal{F}(s) = \mathcal{F}_{\text{nd}}(s)$, proposition F.170 applies. But $\neg\exists \mathbf{f} \in \mathcal{F}(s) : \mathbb{P}_{\mathcal{D}_{X\text{-IID}}}(\mathbf{f}, \gamma)$ which varies with γ, because optimal policies cannot shift. $\square$

Question F.174 (How does entropy relate to instrumental convergence?)**.** Is there a formal relationship between the "degree" of instrumental convergence and the entropy of the distribution over optimal policy sets induced by reward function distribution $\mathcal{D}$?

F.6 Properties of recurrent state distributions

Conjecture F.175 (In deterministic envs., $\lim_{\gamma\to 1}(1-\gamma)\,\mathcal{F}_{\text{nd}}(s,\gamma) = \text{RSD}_{\text{nd}}(s)$)**.** In deterministic environments, if $\mathbf{f} \in \mathcal{F}_{\text{nd}}(s)$, then $\lim_{\gamma\to 1}(1-\gamma)\mathbf{f}(\gamma) \in \text{RSD}_{\text{nd}}(s)$.

Corollary. *In deterministic environments,* $\mathbb{P}_{\mathcal{D}_{X\text{-IID}}}(\mathbf{f}, 1) > 0$ *iff* $\mathbf{f} \in \mathcal{F}_{\text{nd}}(s)$.

Conjecture F.175 does not hold in stochastic environments; see fig. F.29. Conversely, some dominated $\mathbf{f}$ do limit to non-dominated $\mathbf{d} \in \text{RSD}_{\text{nd}}(s)$; for example, consider a state s in a unichain MDP in which there are dominated $\mathbf{f} \in \mathcal{F}(s)$ but $\left|\text{RSD}(s)\right| = 1$.

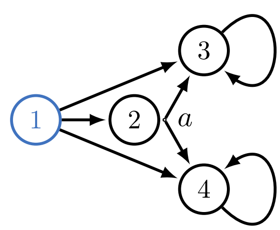


Figure F.29: The bifurcated action a is a stochastic transition, where $T(s_2, a, s_3) = \frac{1}{2} = T(s_2, a, s_4)$. By corollary F.67, navigating from $s_1 \to s_2$ induces a non-dominated visit distribution $\mathbf{f}$. However, its limiting RSD is half s_3, half s_4 and is therefore dominated. $\mathbb{P}_{\mathcal{D}_{X\text{-IID}}}(\mathbf{f}, 1) = 0$ even though $\mathbf{f} \in \mathcal{F}_{\text{nd}}(s)$.

Proposition F.176 (If s' can reach s deterministically, $\text{RSD}(s') \subseteq \text{RSD}(s)$)**.**

Proof. Let $\mathbf{d}^\pi \in \text{RSD}(s')$. Starting from s', π induces state trajectory $s' s_1 s_2 \ldots$. Let π' navigate to a state s_k in this trajectory which s can reach in the fewest steps (where s is considered to "reach" itself in 0 steps); since s can deterministically reach s', this fewest number of steps is finite. $\forall t \geq k : \pi'(s_t) \coloneqq \pi(s_t)$; $\mathbf{d}^{\pi'} = \mathbf{d}^{\pi}$. π' is stationary because it navigates to the state trajectory in the fewest possible number of steps, and therefore it does not conflict with itself. $\square$

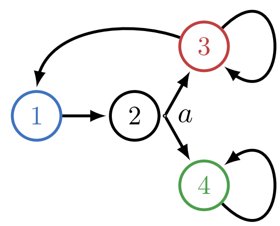


Figure F.30: The bifurcated action a is a stochastic transition, where $T(s_2, a, s_3) = \frac{1}{2} = T(s_2, a, s_4)$. Proposition F.176 does not hold in stochastic environments.

$$\text{RSD}\left(s_1\right) = \left\{ \begin{pmatrix} 0 \\ 0 \\ .5 \\ .5 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix} \right\} \neq \left\{ \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix} \right\} = \text{RSD}\left(s_3\right)$$

Remark. Proposition F.176 is not true in some stochastic settings, even if s can reach s' with probability 1, since it may be the case that only non-stationary policies navigate to s' and then induce the appropriate RSD.

Lemma F.177 (Reachability with probability 1 implies uniformly greater average reward)**.** *If*

$$\max_{\pi \in \Pi} \lim_{t \to \infty} \mathbb{P}\left(s' \text{ reached in the first } t \text{ steps while following } \pi \text{ starting from } s\right) = 1, \tag{F.120}$$

then $\forall R \in \mathbb{R}^{\mathcal{S}} : V^*_{R,\, norm}\left(s, 1\right) \geq V^*_{R,\, norm}\left(s', 1\right)$.

Proof. Let $\mathbf{r} \in \mathbb{R}^{|\mathcal{S}|}$ and let $\mathbf{d} \in \text{RSD}\left(s'\right)$ be such that $V^*_{R,\,\text{norm}}\left(s', 1\right) = \mathbf{d}^\top \mathbf{r}$. In eq. (F.120), each limit exists because the probability is monotonically increasing and bounded $[0,1]$, and the maximum exists because Π is finite. Let π maximize eq. (F.120).

π has probability 1 of eventually reaching s'. Let π^{HD} implement this strategy until it has reached s' at least once in its history. After the agent has reached state s', π^{HD} is defined so as to induce $\mathbf{d}$. This is possible because $\mathbf{d} \in \text{RSD}\left(s'\right)$. Since $\gamma = 1$, π^{HD} therefore induces $\mathbf{d}$ starting from s. Therefore, starting from s, a history-dependent policy can achieve $V^*_{R,\,\text{norm}}\left(s', 1\right)$ average reward for reward function $\mathbf{r}$.

Puterman [68]'s Theorem 6.2.7 shows that when $0 < \gamma < 1$ in finite MDPs, discounted optimal value is invariant to restriction to deterministic stationary policies. If s could not induce an average reward of at least $V^*_{R,\,\text{norm}}\left(s', 1\right)$ via a stationary policy, but could with a history-dependent policy, this would contradict Theorem 6.2.7 by the continuity of $V^*_{R,\,\text{norm}}\left(s, \gamma\right)$ on $\gamma \in [0,1]$, which is shown by lemma D.45 in Turner et al. [99]. That is, if no such gain-optimal stationary policy existed, then stationary optimal policies would be strictly worse than non-stationary optimal policies for $\gamma \approx 1$. Therefore, when starting from state s, there exists some stationary policy with average reward of at least $V^*_{R,\,\text{norm}}\left(s', 1\right)$. □

Corollary F.178 (When $\gamma = 1$, IID POWER decreases iff RSDs become unreachable)**.** *If s can reach s' with probability* 1*, then* $\text{POWER}_{\mathcal{D}_{bound}}\left(s', 1\right) \leq \text{POWER}_{\mathcal{D}_{bound}}\left(s, 1\right)$*. If the environment is deterministic, the inequality is strict iff* $\text{RSD}_{\text{nd}}\left(s'\right) \subsetneq \text{RSD}_{\text{nd}}\left(s\right)$*.*

Proof. Since s can reach s' with probability 1, lemma F.177 implies that $\text{POWER}_{\mathcal{D}_{\text{bound}}}\left(s', 1\right) \leq \text{POWER}_{\mathcal{D}_{\text{bound}}}\left(s, 1\right)$.

Given determinism, proposition F.176 implies that $\text{RSD}_{\text{nd}}\left(s'\right) \subseteq \text{RSD}_{\text{nd}}\left(s\right)$. Then if the inequality is strict, $\text{RSD}_{\text{nd}}\left(s'\right) \subsetneq \text{RSD}_{\text{nd}}\left(s\right)$ because $\text{POWER}_{\mathcal{D}_{\text{bound}}}\left(s', 1\right)$ is determined by the availability of non-dominated RSDs by lemma D.56. If $\text{RSD}_{\text{nd}}\left(s'\right) \subsetneq \text{RSD}_{\text{nd}}\left(s\right)$, the inequality is strict by proposition 5.28 via ϕ the identity permutation. □

Remark. Corollary F.178 is false if s can only reach s' with positive probability less than 1. For example, suppose action a is such that $T(s, a, s_{\text{high}}) = .5 = T(s, a, s_{\text{low}}) = .99$ and all actions are equivalent to a at state s. Further suppose that $\text{POWER}_{\mathcal{D}_{\text{bound}}}\left(s_{\text{high}}, 1\right) >$

$\textsc{Power}_{\mathcal{D}_{\text{bound}}}(s_{\text{low}}, 1)$. Then

$$\textsc{Power}_{\mathcal{D}_{\text{bound}}}(s, 1) = \frac{1}{2}\left(\textsc{Power}_{\mathcal{D}_{\text{bound}}}(s_{\text{high}}, 1) + \textsc{Power}_{\mathcal{D}_{\text{bound}}}(s_{\text{low}}, 1)\right)$$

is intermediate between the two values.

Definition F.179 (Communicating MDP)**.** An MDP is *communicating* when every state is able to reach every other state with positive probability.

Proposition F.180 (When $\gamma = 1$ in communicating MDPs, $\textsc{Power}_{\mathcal{D}_{\text{bound}}}$ is equal everywhere)**.** *In communicating* MDPs*,* $\forall s, s' : \textsc{Power}_{\mathcal{D}_{bound}}(s, 1) = \textsc{Power}_{\mathcal{D}_{bound}}(s', 1)$.

Proof. Since the MDP is communicating, all states can reach each other with asymptotic probability 1. Apply corollary F.178 to conclude that $\textsc{Power}_{\mathcal{D}_{\text{bound}}}(s, 1) = \textsc{Power}_{\mathcal{D}_{\text{bound}}}(s', 1)$. $\square$

F.7 Power

F.7.1 The structure of Power

Conjecture F.181 (Graphical characterization of IID Power agreement)**.** There exists a graphical condition characterizing when, for all bounded state reward distributions X, states s, s' are such that $\forall \gamma \in [0, 1] : \textsc{Power}_{\mathcal{D}_{X\text{-IID}}}(s, \gamma) = \textsc{Power}_{\mathcal{D}_{X\text{-IID}}}(s', \gamma)$.

Theorem F.182 (The structure of Power)**.** *Let* $\gamma \in (0, 1)$ *and let* $\mathcal{D}_{bound}$ *induce probability measure* F*.*

$$\textsc{Power}_{\mathcal{D}_{bound}}(s, \gamma) = (1-\gamma)\sum_{t=1}^{\infty}\gamma^{t-1}\sum_{\mathbf{f}^\pi \in \mathcal{F}_{\text{nd}}(s)}\int_{\text{supp}\left(\mathbf{f}^\pi(\gamma) \geq \mathcal{F}(s,\gamma)\right)}\left(\left(\mathbf{T}^\pi\right)^t \mathbf{e}_s\right)^\top \mathbf{r}\, \mathrm{d}F(\mathbf{r}). \tag{F.121}$$

Proof. Suppose $\gamma \in (0, 1)$.

$$V^*_{\mathcal{D}_{\text{bound}}}(s, \gamma) := \int_{\text{supp}(\mathcal{D}_{\text{bound}})} \max_{\mathbf{f}^\pi \in \mathcal{F}(s)} \mathbf{f}^\pi(\gamma)^\top \mathbf{r}\, \mathrm{d}F(\mathbf{r}) \tag{F.122}$$

$$= \sum_{\mathbf{f}^\pi \in \mathcal{F}(s)} \int_{\operatorname{supp}\left(\mathbf{f}^\pi(\gamma) \geq \mathcal{F}(s,\gamma)\right)} \mathbf{f}^\pi(\gamma)^\top \mathbf{r} \, \mathrm{d}F(\mathbf{r}) \tag{F.123}$$

$$= \sum_{\mathbf{f}^\pi \in \mathcal{F}_{\text{nd}}(s)} \int_{\operatorname{supp}\left(\mathbf{f}^\pi(\gamma) \geq \mathcal{F}(s,\gamma)\right)} \mathbf{f}^\pi(\gamma)^\top \mathbf{r} \, \mathrm{d}F(\mathbf{r}) \tag{F.124}$$

$$= \sum_{\mathbf{f}^\pi \in \mathcal{F}_{\text{nd}}(s)} \int_{\operatorname{supp}\left(\mathbf{f}^\pi(\gamma) \geq \mathcal{F}(s,\gamma)\right)} \left(\sum_{t=0}^{\infty} \left(\gamma \mathbf{T}^\pi\right)^t \mathbf{e}_s\right)^\top \mathbf{r} \, \mathrm{d}F(\mathbf{r}) \tag{F.125}$$

$$= \sum_{\mathbf{f}^\pi \in \mathcal{F}_{\text{nd}}(s)} \int_{\operatorname{supp}\left(\mathbf{f}^\pi(\gamma) \geq \mathcal{F}(s,\gamma)\right)} \sum_{t=0}^{\infty} \left(\left(\gamma \mathbf{T}^\pi\right)^t \mathbf{e}_s\right)^\top \mathbf{r} \, \mathrm{d}F(\mathbf{r}) \tag{F.126}$$

$$= \sum_{\mathbf{f}^\pi \in \mathcal{F}_{\text{nd}}(s)} \sum_{t=0}^{\infty} \gamma^t \int_{\operatorname{supp}\left(\mathbf{f}^\pi(\gamma) \geq \mathcal{F}(s,\gamma)\right)} \left(\left(\mathbf{T}^\pi\right)^t \mathbf{e}_s\right)^\top \mathbf{r} \, \mathrm{d}F(\mathbf{r}) \tag{F.127}$$

$$= \sum_{t=0}^{\infty} \gamma^t \sum_{\mathbf{f}^\pi \in \mathcal{F}_{\text{nd}}(s)} \int_{\operatorname{supp}\left(\mathbf{f}^\pi(\gamma) \geq \mathcal{F}(s,\gamma)\right)} \left(\left(\mathbf{T}^\pi\right)^t \mathbf{e}_s\right)^\top \mathbf{r} \, \mathrm{d}F(\mathbf{r}). \tag{F.128}$$

Equation (F.124) follows because optimal value is invariant to restriction to non-dominated visit distribution functions. Equation (F.125) holds by the definition of a visit distribution (definition 5.3).

Equation (F.127) holds by Fubini's theorem: for a fixed π, consider the function $f(t, \mathbf{r} \mid \pi) := \left(\left(\gamma \mathbf{T}^\pi\right)^t \mathbf{e}_s\right)^\top \mathbf{r}$. Since X is bounded $[a, b]$, $\left|f(t, \mathbf{r} \mid \pi)\right| \leq \gamma^t \left|b\right|$ for all $t \geq 0, \mathbf{r} \in \operatorname{supp}(\mathcal{D})$. Since $\gamma \in (0,1)$,

$$\int_{\operatorname{supp}\left(\mathbf{f}^\pi(\gamma) \geq \mathcal{F}(s,\gamma)\right)} \sum_{t=0}^{\infty} \left|f(t, \mathbf{r} \mid \pi)\right| \mathrm{d}F(\mathbf{r}) \leq \frac{|b|}{1-\gamma} < \infty. \tag{F.129}$$

Furthermore, for a fixed t, f is continuous on $\mathbf{r}$ (as the function is linear when t is fixed). Therefore, $f : \mathbb{Z}_{\geq 0} \times \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}$ is continuous: the preimage of an open set in $\mathbb{R}$ is open in the product topology on the domain, since the standard topology on $\mathbb{Z}_{\geq 0}$ is discrete. Since f is continuous, it is measurable. Therefore, the conditions of Fubini's theorem are met. Equation (F.128) holds because $\mathcal{F}_{\text{nd}}(s)$ is finite.

We now show the desired Power identity.

$$\begin{aligned}
&\textsc{Power}_{\mathcal{D}_{\text{bound}}}(s,\gamma) && \text{(F.130)}\\
&:= \mathop{\mathbb{E}}_{\mathbf{r}\sim\mathcal{D}}\left[\max_{\mathbf{f}\in\mathcal{F}(s)} \frac{1-\gamma}{\gamma}\left(\mathbf{f}(\gamma)-\mathbf{e}_s\right)^\top \mathbf{r}\right] && \text{(F.131)}\\
&= \frac{1-\gamma}{\gamma}\left(-\mathop{\mathbb{E}}_{\mathbf{r}\sim\mathcal{D}_{\text{bound}}}\left[\mathbf{e}_s^\top\mathbf{r}\right]+\sum_{t=0}^{\infty}\gamma^t\sum_{\mathbf{f}^\pi\in\mathcal{F}_{\text{nd}}(s)}\int_{\text{supp}\left(\mathbf{f}^\pi(\gamma)\geq\mathcal{F}(s,\gamma)\right)}\left(\left(\mathbf{T}^\pi\right)^t\mathbf{e}_s\right)^\top\mathbf{r}\,\mathrm{d}F(\mathbf{r})\right) && \text{(F.132)}\\
&= (1-\gamma)\sum_{t=1}^{\infty}\gamma^{t-1}\sum_{\mathbf{f}^\pi\in\mathcal{F}_{\text{nd}}(s)}\int_{\text{supp}\left(\mathbf{f}^\pi(\gamma)\geq\mathcal{F}(s,\gamma)\right)}\left(\left(\mathbf{T}^\pi\right)^t\mathbf{e}_s\right)^\top\mathbf{r}\,\mathrm{d}F(\mathbf{r}). && \text{(F.133)}
\end{aligned}$$

Equation (F.132) holds by eq. (F.128). Equation (F.133) holds by the fact that the $t=0$ summand equals $\mathbb{E}_{\mathbf{r}\sim\mathcal{D}_{\text{bound}}}\left[\mathbf{e}_s^\top\mathbf{r}\right]$. $\square$

Proposition F.183 (Sufficient condition for Power being rational on γ)**.** *If no reward function has optimal policy shifts and the environment is deterministic, then* $\textsc{Power}_{\mathcal{D}_{bound}}(s,\gamma)$ *is a rational function on* $\gamma\in(0,1)$.

Proof. Let $\gamma\in(0,1)$.

$$\begin{aligned}
\textsc{Power}_{\mathcal{D}_{\text{bound}}}(s,\gamma) &= (1-\gamma)\sum_{t=1}^{\infty}\gamma^{t-1}\sum_{\mathbf{f}^\pi\in\mathcal{F}_{\text{nd}}(s)}\int_{\text{supp}\left(\mathbf{f}^\pi(\gamma)\geq\mathcal{F}(s,\gamma)\right)}\left(\left(\mathbf{T}^\pi\right)^t\mathbf{e}_s\right)^\top\mathbf{r}\,\mathrm{d}F(\mathbf{r}) && \text{(F.134)}\\
&= (1-\gamma)\sum_{\mathbf{f}^\pi\in\mathcal{F}_{\text{nd}}(s)}\sum_{t=1}^{\infty}\gamma^{t-1}c_t^\pi. && \text{(F.135)}
\end{aligned}$$

Equation (F.134) follows by theorem F.182. In eq. (F.135), let each c_t^π be a constant depending only on π and t. Since no optimal policy shifts occur, $\forall\mathbf{f}^\pi\in\mathcal{F}(s)$: $\text{supp}\left(\mathbf{f}^\pi(\gamma)\geq\mathcal{F}(s,\gamma)\right)$ is constant on γ. Therefore, the domain of integration is fixed in each inner-summand of eq. (F.134), and so the constants c_t^π do not depend on γ. Furthermore, $\mathcal{F}_{\text{nd}}(s)$ is finite and so we can interchange the summation signs.

Since the environment is deterministic, each $\mathbf{f}^\pi$ has entered a cycle at most $|\mathcal{S}|$ steps into the policy's trajectory from s. Therefore, $(c_t^\pi)_{t\geq|\mathcal{S}|}$ is k-periodic for some k ($1 \leq k \leq |\mathcal{S}|$). Then for each $\mathbf{f}^\pi$, there exists a $(k-1)$-degree polynomial P^π such that

$$\sum_{t=|\mathcal{S}|}^{\infty} \gamma^{t-1} c_t^\pi = \frac{\gamma^{|\mathcal{S}|-1}}{1-\gamma^k} P^\pi. \tag{F.136}$$

Therefore,

$$\sum_{t=1}^{\infty} \gamma^{t-1} c_t^\pi = \sum_{t=1}^{|\mathcal{S}|-1} \gamma^{t-1} c_t^\pi + \sum_{t=|\mathcal{S}|}^{\infty} \gamma^{t-1} c_t^\pi \tag{F.137}$$

is also rational on γ. Since $\text{POWER}_{\mathcal{D}_\text{bound}}$ equals $(1-\gamma)$ times the sum of finitely many rational functions, $\text{POWER}_{\mathcal{D}_\text{bound}}$ is also rational. $\square$

Conjecture F.184 ($\text{POWER}_{\mathcal{D}_\text{bound}}(s,\gamma)$ is piecewise rational on γ)**.**

F.7.2 POWER at its limit points

Lemma F.185 (POWER when $\gamma = 0$)**.**

$$\text{POWER}_{\mathcal{D}_{bound}}(s, 0) = \mathop{\mathbb{E}}_{\mathbf{r}\sim\mathcal{D}_{bound}}\left[\max_{\mathbf{d}\in T(s)} \mathbf{d}^\top \mathbf{r}\right] = \mathop{\mathbb{E}}_{\mathbf{r}\sim\mathcal{D}_{bound}}\left[\max_{\mathbf{d}\in T_{\text{nd}}(s)} \mathbf{d}^\top \mathbf{r}\right]. \tag{F.138}$$

Proof.

$$\begin{aligned}
\text{POWER}_{\mathcal{D}_\text{bound}}(s, 0) &:= \lim_{\gamma\to 0} \mathop{\mathbb{E}}_{\mathbf{r}\sim\mathcal{D}_\text{bound}}\left[\max_{\mathbf{f}^{\pi,s}\in\mathcal{F}(s)} \frac{1-\gamma}{\gamma}\left(\mathbf{f}^{\pi,s}(\gamma) - \mathbf{e}_s\right)^\top \mathbf{r}\right] && \text{(F.139)}\\
&= \lim_{\gamma\to 0} \mathop{\mathbb{E}}_{\mathbf{r}\sim\mathcal{D}_\text{bound}}\left[\max_{\mathbf{f}^{\pi,s}\in\mathcal{F}(s)} (1-\gamma) \mathop{\mathbb{E}}_{s'\sim T(s,\pi(s))}\left[\mathbf{f}^{\pi,s'}(\gamma)\right]^\top \mathbf{r}\right] && \text{(F.140)}\\
&= \mathop{\mathbb{E}}_{\mathbf{r}\sim\mathcal{D}_\text{bound}}\left[\max_{\mathbf{f}^{\pi,s}\in\mathcal{F}(s)} \lim_{\gamma\to 0} (1-\gamma) \mathop{\mathbb{E}}_{s'\sim T(s,\pi(s))}\left[\mathbf{f}^{\pi,s'}(\gamma)\right]^\top \mathbf{r}\right] && \text{(F.141)}\\
&= \mathop{\mathbb{E}}_{\mathbf{r}\sim\mathcal{D}_\text{bound}}\left[\max_{\mathbf{f}^{\pi,s}\in\mathcal{F}(s)} \mathop{\mathbb{E}}_{s'\sim T(s,\pi(s))} \left[\mathbf{e}_{s'}\right]^\top \mathbf{r}\right] && \text{(F.142)}
\end{aligned}$$

$$= \mathop{\mathbb{E}}_{\mathbf{r}\sim\mathcal{D}_{\text{bound}}}\left[\max_{\mathbf{d}\in T(s)} \mathbf{d}^\top \mathbf{r}\right] \tag{F.143}$$

$$= \mathop{\mathbb{E}}_{\mathbf{r}\sim\mathcal{D}_{\text{bound}}}\left[\max_{\mathbf{d}\in T_{\text{nd}}(s)} \mathbf{d}^\top \mathbf{r}\right]. \tag{F.144}$$

Equation (F.144) follows because for all $\mathbf{r} \in \mathbb{R}^{|\mathcal{S}|}$, corollary D.16 shows that $\max_{\mathbf{d}\in T(s)} \mathbf{d}^\top \mathbf{r} = \max_{\mathbf{d}\in \text{ND}\left(T(s)\right)} \mathbf{d}^\top \mathbf{r} =: \max_{\mathbf{d}\in T_{\text{nd}}(s)} \mathbf{d}^\top \mathbf{r}$. □

Remark. The non-strict inequality of lemma F.186 holds for *any* state reward distribution – X need not be continuous. However, the strict inequality does require continuity in general.

Lemma F.186 (Minimal $\textsc{Power}_{\mathcal{D}_{X\text{-IID}}}$)**.** $\textsc{Power}_{\mathcal{D}_{X\text{-IID}}}(s,\gamma) \geq \mathbb{E}[X]$. *If* $\gamma \in (0,1)$, *equality holds iff* $\left|\mathcal{F}(s)\right| = 1$.

Proof. Let $\gamma \in (0,1)$. If $\left|\mathcal{F}(s)\right| = 1$, then $V^*_{\mathcal{D}_{X\text{-IID}}}(s,\gamma) = \mathbb{E}[X]\,\frac{1}{1-\gamma}$ by the linearity of expectation and the fact that $\mathcal{D}_{X\text{-IID}}$ distributes reward independently and identically across states according to state distribution X. By lemma D.43,

$$\textsc{Power}_{\mathcal{D}_{X\text{-IID}}}(s,\gamma) = \frac{1-\gamma}{\gamma}\Big(V^*_{\mathcal{D}_{X\text{-IID}}}(s,\gamma) - \mathbb{E}[X]\Big),$$

and so $\textsc{Power}_{\mathcal{D}_{X\text{-IID}}}(s,\gamma) = \mathbb{E}[X]$.

Suppose $\left|\mathcal{F}(s)\right| > 1$, and let $\mathbf{f} \in \mathcal{F}(s)$.

$$\frac{\mathbb{E}[X]}{1-\gamma} = \mathop{\mathbb{E}}_{\mathbf{r}\sim\mathcal{D}_{X\text{-IID}}}\left[\max_{\mathbf{f}''\in\{\mathbf{f}\}} \mathbf{f}''(\gamma)^\top \mathbf{r}\right] \tag{F.145}$$

$$< \mathop{\mathbb{E}}_{\mathbf{r}\sim\mathcal{D}_{X\text{-IID}}}\left[\max_{\mathbf{f}''\in\mathcal{F}(s)} \mathbf{f}''(\gamma)^\top \mathbf{r}\right] \tag{F.146}$$

$$=: V^*_{\mathcal{D}_{X\text{-IID}}}(s,\gamma). \tag{F.147}$$

Let $A := \left\{\mathbf{f}(\gamma)\right\}, B := \mathcal{F}(s,\gamma)$. Since $\left|\mathcal{F}(s)\right| > 1$, $\text{ND}(B) = \mathcal{F}_{\text{nd}}(s,\gamma) > 2$ by corollary F.65. Therefore, $\text{ND}(B) \setminus A$ is non-empty. Since $\mathcal{D}_{X\text{-IID}}$ is continuous IID,

$\exists b < c : (b, c)^{|\mathcal{S}|} \subseteq \text{supp}(\mathcal{D}_{X\text{-IID}})$ by lemma D.27. Therefore, for $g : \mathbb{R} \to \mathbb{R}$ the identity and ϕ the identity permutation, eq. (F.146) holds by lemma D.26. Equation (F.147) implies that $\text{POWER}_{\mathcal{D}_{X\text{-IID}}}(s, \gamma) > \mathbb{E}[X]$.

In general, eq. (F.146) holds as a non-strict inequality, and so $\text{POWER}_{\mathcal{D}_{X\text{-IID}}}(s, \gamma) \geq \mathbb{E}[X]$ for all $\gamma \in (0, 1)$. This inequality therefore holds in the limits $\gamma \to 0$ and $\gamma \to 1$. □

Conjecture F.187 handles the limiting cases of lemma F.186 (minimal POWER).

Conjecture F.187 (Minimal POWER at $\gamma = 0$ and $\gamma = 1$)**.**

1. $\text{POWER}_{\mathcal{D}_{\text{bound}}}(s, 0) = \mathbb{E}[X]$ iff $\left|T(s)\right| = 1$.
2. $\text{POWER}_{\mathcal{D}_{\text{bound}}}(s, 1) = \mathbb{E}[X]$ iff $\left|\text{RSD}(s)\right| = 1$.

Remark. The non-strict inequality of lemma F.188 holds for *any* state reward distribution – X need not be continuous. However, the strict inequality does require continuity in general.

Lemma F.188 (Maximal $\text{POWER}_{\mathcal{D}_{X\text{-IID}}}$)**.** $\text{POWER}_{\mathcal{D}_{X\text{-IID}}}(s, \gamma) \leq \mathbb{E}\left[\max \text{ of } |\mathcal{S}| \text{ draws from } X\right]$. *If* $\gamma \in (0, 1)$*, equality holds iff* s *can deterministically reach all states in one step and all states have deterministic self-loops.*

Proof.

$$\text{POWER}_{\mathcal{D}_{X\text{-IID}}}(s, \gamma) \leq \mathbb{E}_{R \sim \mathcal{D}}\left[\max_{s'' \in \mathcal{S}} R(s'')\right] \tag{F.148}$$

$$= \mathbb{E}\left[\max \text{ of } |\mathcal{S}| \text{ draws from } X\right]. \tag{F.149}$$

Equation (F.148) holds by proposition 5.14. Equation (F.149) holds because $\mathcal{D}_{X\text{-IID}}$ is IID over states according to state reward distribution X.

If $Ch_{\text{sure}}(s) = \mathcal{S}$ and $\forall s' \in \mathcal{S} : s' \in Ch_{\text{sure}}(s')$, then eq. (F.148) is an equality by proposition 5.14.

Backward direction. Suppose that $\gamma \in (0,1)$. To show that

$$\textsc{Power}_{\mathcal{D}_{X\text{-IID}}}(s,\gamma) = \mathbb{E}\left[\max \text{ of } |\mathcal{S}| \text{ draws from } X\right]$$

implies that $Ch_{\text{sure}}(s) = \mathcal{S}$ and $\forall s' \in \mathcal{S} : s' \in Ch_{\text{sure}}\left(s'\right)$, we show the contrapositive: $\exists s' \in \mathcal{S} : s' \notin Ch_{\text{sure}}(s)$ or $s' \notin Ch_{\text{sure}}\left(s'\right)$ implies that

$$\textsc{Power}_{\mathcal{D}_{X\text{-IID}}}(s,\gamma) < \mathbb{E}\left[\max \text{ of } |\mathcal{S}| \text{ draws from } X\right].$$

Let $\mathbf{f}^{s'} := \mathbf{e}_s + \frac{\gamma}{1-\gamma}\mathbf{e}_{s'}$ be the visit distribution function which would be induced by immediately navigating to s' from s and then staying at s'. By assumption, this is not currently possible for some $s' \in \mathcal{S}$, and so $\mathbf{f}^{s'}$ would strictly increase the achievable visit frequency on s', starting from s. Formally, $\mathbf{f}^{s'}(\gamma)^\top \mathbf{e}_{s'} > \max_{\mathbf{f}\in\mathcal{F}(s)} \mathbf{f}(\gamma)^\top \mathbf{e}_{s'}$.

Let $F^{s'}$ be the set of visit distribution functions at s which would be made available as a result of adding deterministic transitions $s \to s'$ and $s' \to s'$.

$$\textsc{Power}_{\mathcal{D}_{X\text{-IID}}}(s,\gamma) := \mathop{\mathbb{E}}_{\mathbf{r}\sim\mathcal{D}_{X\text{-IID}}}\left[\max_{\mathbf{f}\in\mathcal{F}(s)} \frac{1-\gamma}{\gamma}\left(\mathbf{f}(\gamma)-\mathbf{e}_s\right)^\top \mathbf{r}\right] \tag{F.150}$$

$$< \mathop{\mathbb{E}}_{\mathbf{r}\sim\mathcal{D}_{X\text{-IID}}}\left[\max_{\mathbf{f}\in\mathcal{F}(s)\cup F^{s'}} \frac{1-\gamma}{\gamma}\left(\mathbf{f}(\gamma)-\mathbf{e}_s\right)^\top \mathbf{r}\right] \tag{F.151}$$

$$\leq \mathbb{E}\left[\max \text{ of } |\mathcal{S}| \text{ draws from } X\right]. \tag{F.152}$$

Let $A := \left\{\frac{1-\gamma}{\gamma}\left(\mathbf{f}(\gamma)-\mathbf{e}_s\right) \mid \mathbf{f}\in\mathcal{F}(s)\right\}, B := \left\{\frac{1-\gamma}{\gamma}\left(\mathbf{f}(\gamma)-\mathbf{e}_s\right) \mid \mathbf{f}\in\mathcal{F}(s)\cup F^{s'}\right\}$. Since $\mathbf{f}^{s'}$ maximizes s'-visitation frequency, $\mathbf{f}^{s'}(\gamma)^\top\mathbf{e}_{s'} > \max_{\mathbf{f}\in\left(\mathcal{F}(s)\cup F^{s'}\right)\setminus\left\{\mathbf{f}^{s'}\right\}} \mathbf{f}(\gamma)^\top \mathbf{e}_{s'}$, which implies that $\mathbf{f}^{s'}(\gamma) \in \text{ND}(B)\setminus\text{ND}(A)$. Therefore, $\text{ND}(B)\setminus\text{ND}(A)$ is non-empty. Since $\mathcal{D}_{X\text{-IID}}$ is continuous IID, $\exists b < c : (b,c)^{|\mathcal{S}|} \subseteq \text{supp}(\mathcal{D}_{X\text{-IID}})$ by lemma D.27. Therefore, for $g : \mathbb{R}\to\mathbb{R}$ the identity and ϕ the identity permutation, eq. (F.151) holds by lemma D.26. Equation (F.152) follows because proposition 5.14 bounds the $\textsc{Power}_{\mathcal{D}_{X\text{-IID}}}$ of all valid MDP structures. $\square$

Proposition F.189 handles the limiting cases of lemma F.188 (maximal Power).

Proposition F.189 (Maximal $\textsc{Power}$ at $\gamma = 0$ and $\gamma = 1$)**.**

1. $\textsc{Power}_{\mathcal{D}_{X\text{-IID}}}(s, 0) = \mathbb{E}\left[\max \text{ of } |\mathcal{S}| \text{ draws from } X\right]$ *iff* $Ch_{sure}(s) = \mathcal{S}$.

2. $\textsc{Power}_{\mathcal{D}_{X\text{-IID}}}(s, 1) = \mathbb{E}\left[\max \text{ of } |\mathcal{S}| \text{ draws from } X\right]$ *iff s can reach all states with probability* 1 *and* $\forall s' \in \mathcal{S} : s' \in Ch_{sure}(s')$.

Proof. Item 1:

$$\begin{aligned}
\textsc{Power}_{\mathcal{D}_{X\text{-IID}}}(s, 0) &= \mathop{\mathbb{E}}_{\mathbf{r} \sim \mathcal{D}_{X\text{-IID}}}\left[\max_{\mathbf{d} \in T(s)} \mathbf{d}^\top \mathbf{r}\right] && \text{(F.153)}\\
&\leq \mathop{\mathbb{E}}_{\mathbf{r} \sim \mathcal{D}_{X\text{-IID}}}\left[\max_{\mathbf{d} \in T(s) \cup \left\{\mathbf{e}_{s''} \mid s'' \in \mathcal{S}\right\}} \mathbf{d}^\top \mathbf{r}\right] && \text{(F.154)}\\
&= \mathop{\mathbb{E}}_{\mathbf{r} \sim \mathcal{D}_{X\text{-IID}}}\left[\max_{\mathbf{d} \in \left\{\mathbf{e}_{s''} \mid s'' \in \mathcal{S}\right\}} \mathbf{d}^\top \mathbf{r}\right] && \text{(F.155)}\\
&= \mathbb{E}\left[\max \text{ of } |\mathcal{S}| \text{ draws from } X\right]. && \text{(F.156)}
\end{aligned}$$

Equation (F.153) holds by lemma F.185. Applying lemma D.26 (lemma D.27 guarantees that $\exists b < c : (b, c)^{|\mathcal{S}|} \subseteq \text{supp}(\mathcal{D}_{X\text{-IID}})$), $A \coloneqq T(s)$, $B \coloneqq T(s) \cup \{\mathbf{e}_{s'}\}$, $g : \mathbb{R} \to \mathbb{R}$ the identity function, and ϕ the identity permutation, eq. (F.154) holds by lemma D.26. Equation (F.155) holds because $\text{ND}\left(T(s) \cup \left\{\mathbf{e}_{s''} \mid s'' \in \mathcal{S}\right\}\right) = \left\{\mathbf{e}_{s''} \mid s'' \in \mathcal{S}\right\}$ by lemma F.57. Equation (F.156) holds because reward is IID across states under $\mathcal{D}_{X\text{-IID}}$.

Suppose $\exists s' \notin Ch_{\text{sure}}(s)$. By the definition of $T(s)$ (definition F.5), $\exists s' \in \mathcal{S} : \mathbf{e}_{s'} \notin T(s)$. Furthermore, $\mathbf{e}_{s'}^\top \mathbf{e}_{s'} > \max_{\mathbf{d} \in T(s) \cup \left\{\mathbf{e}_{s''} \mid s'' \neq s'\right\}} \mathbf{d}^\top \mathbf{e}_{s'}$, and so $\mathbf{e}_{s'} \in \left\{\mathbf{e}_{s''} \mid s'' \in \mathcal{S}\right\} = \text{ND}\left(T(s) \cup \left\{\mathbf{e}_{s''} \mid s'' \in \mathcal{S}\right\}\right)$. Therefore, lemma D.26 guarantees that eq. (F.154) is strict. Therefore, $\textsc{Power}_{\mathcal{D}_{X\text{-IID}}}(s, 0) = \mathbb{E}\left[\max \text{ of } |\mathcal{S}| \text{ draws from } X\right]$ implies that $Ch_{\text{sure}}(s) = \mathcal{S}$.

Suppose $Ch_{\text{sure}}(s) = \mathcal{S}$. Then lemma D.26 guarantees that eq. (F.154) is an equality, and so $Ch_{\text{sure}}(s) = \mathcal{S}$ implies that $\textsc{Power}_{\mathcal{D}_{X\text{-IID}}}(s, 0) = \mathbb{E}\left[\max \text{ of } |\mathcal{S}| \text{ draws from } X\right]$.

Item 2 follows by applying the above reasoning with $\text{RSD}(s)$ in place of $T(s)$, substituting in the fact that $\mathbf{e}_{s'} \in T(s)$ iff s can reach s' with probability 1 and $s' \in Ch_{\text{sure}}(s')$. $\square$

Conjecture F.190 (Lower bound on current POWER based on next-step expected reward)**.**

$$\text{POWER}_{\mathcal{D}_{\text{bound}}}(s,\gamma) \geq (1-\gamma)\max_a \mathop{\mathbb{E}}_{\substack{s'\sim T(s,a),\\ R\sim\mathcal{D}_{\text{bound}}}}\left[R(s')\right] + \gamma \min_a \mathop{\mathbb{E}}_{s'\sim T(s,a)}\left[\text{POWER}_{\mathcal{D}_{\text{bound}}}(s',\gamma)\right].$$

Lemma F.191 (When $\gamma = 0$, having similar children implies equal $\text{POWER}_{\mathcal{D}_{\text{bound}}}$)**.** *Suppose $D \subseteq T_{\text{nd}}(s)$ is similar to $T_{\text{nd}}(s')$. Then $\text{POWER}_{\mathcal{D}_{X\text{-IID}}}(s,0) \geq \text{POWER}_{\mathcal{D}_{X\text{-IID}}}(s',0)$. If $D \subsetneq T_{\text{nd}}(s)$, then $\exists \gamma^* : \forall \gamma \in [0,\gamma^*) : \text{POWER}_{\mathcal{D}_{X\text{-IID}}}(s,\gamma) > \text{POWER}_{\mathcal{D}_{X\text{-IID}}}(s',\gamma)$.*

Proof. Suppose $T_{\text{nd}}(s') = \phi \cdot D$, where ϕ .

$$\text{POWER}_{\mathcal{D}_{X\text{-IID}}}(s,0) = \mathop{\mathbb{E}}_{\mathbf{r}\sim\mathcal{D}_{X\text{-IID}}}\left[\max_{\mathbf{d}\in T_{\text{nd}}(s)} \mathbf{d}^\top \mathbf{r}\right] \tag{F.157}$$

$$\geq \mathop{\mathbb{E}}_{\mathbf{r}\sim\mathcal{D}_{X\text{-IID}}}\left[\max_{\mathbf{d}\in D} \mathbf{d}^\top \mathbf{r}\right] \tag{F.158}$$

$$= \mathop{\mathbb{E}}_{\mathbf{r}\sim\mathcal{D}_{X\text{-IID}}}\left[\max_{\mathbf{d}\in D} (\mathbf{P}_\phi \mathbf{d})^\top \mathbf{r}\right] \tag{F.159}$$

$$= \text{POWER}_{\mathcal{D}_{X\text{-IID}}}(s',0). \tag{F.160}$$

Equation (F.157) follows by lemma F.185. Equation (F.159) follows because reward is assumed to be IID over states. If $D \subsetneq T_{\text{nd}}(s)$, non-domination ensures that the child distributions of $T_{\text{nd}}(s) \setminus D$ are strictly greedily optimal for a positive measure set of reward functions. In this case, eq. (F.158)'s inequality is strict; strict inequality holds for $\gamma \approx 0$ by the continuity of $\text{POWER}_{\mathcal{D}_{X\text{-IID}}}(s',\gamma)$ on γ (lemma 5.13). □

Proposition F.192 (POWER achieves ordinal equilibrium as $\gamma \to 1$)**.** *Let $s \succeq^\gamma_{\text{POWER}_{\mathcal{D}_{bound}}} s'$ when $\text{POWER}_{\mathcal{D}_{bound}}(s,\gamma) \geq \text{POWER}_{\mathcal{D}_{bound}}(s',\gamma)$. $\exists \gamma^* : \forall \gamma \in (\gamma^*, 1) : \succeq^\gamma_{\text{POWER}_{\mathcal{D}_{bound}}}$ is constant.*

Proof. Consider two states s and s'. By the Lipschitz continuity of POWER on γ

(lemma 5.13) and the fact that the domain is bounded $\gamma \in [0,1]$,

$$\operatorname{sgn}\left(\textsc{Power}_{\mathcal{D}_{\text{bound}}}(s,\gamma) - \textsc{Power}_{\mathcal{D}_{\text{bound}}}(s',\gamma)\right)$$

changes value finitely many times on $\gamma \in [0,1]$. There are only finitely many such pairs of states, and so the result follows. □

Conjecture F.193 (States with different $\textsc{Power}_{\mathcal{D}_{\text{bound}}}$ functions are equal for finitely many γ).

F.7.3 States whose $\textsc{Power}_{\mathcal{D}_{\text{bound}}}$ can be immediately determined

Corollary F.194 (Delay linearizes $\textsc{Power}$). *Let $s_0, \ldots, s_\ell$ be such that for all $i < \ell$, $Ch(s_i) = \{s_{i+1}\}$. Then* $\textsc{Power}_{\mathcal{D}_{X\text{-IID}}}(s_0,\gamma) = \sum_{i=1}^{\ell} \gamma^{i-1}\, \mathbb{E}_{R\sim\mathcal{D}'}\left[R(s_i)\right] + \gamma^{\ell}\textsc{Power}_{\mathcal{D}_{X\text{-IID}}}(s_\ell,\gamma)$.

Proof. Iteratively apply lemma D.46 ℓ times. Equality must hold, as each s_i can only reach s_{i+1}. □

Remark. For $\mathcal{D}_{X\text{-IID}}$, the $\sum_{i=1}^{\ell} \gamma^{i-1}\, \mathbb{E}_{R\sim\mathcal{D}_{X\text{-IID}}}\left[R(s_i)\right]$ term in corollary F.194 simplifies to $(1-\gamma^\ell)\,\mathbb{E}\left[X\right]$.

Deterministic delay is a special case of a more general principle.

Definition F.195 (State reachability by time t). $\textsc{Reach}(s,t)$ is the set of states which s can reach in exactly t steps with positive probability.

Definition F.196 (Time-uniform states). In deterministic environments, a state s is *time-uniform* when $\forall t > 0, s', s'' \in \textsc{Reach}(s,t) : s'$ and s'' can either reach the same states in one step, or they can only reach themselves.

Lemma F.197 (If s' is reachable from a time-uniform state s, then s' is time-uniform).

Proof. This follows directly from definition F.196. □

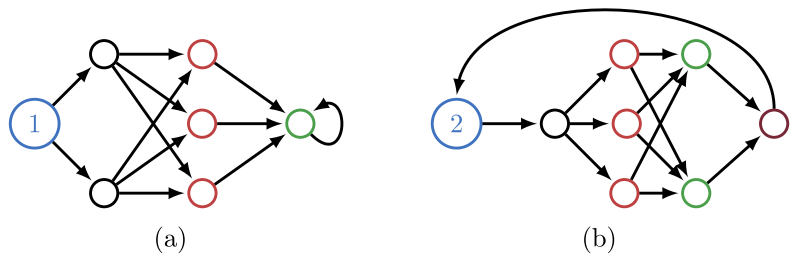


Figure F.31: States of the same color have the same children. For X uniform: $\textsc{Power}_{\mathcal{D}_{X\text{-IID}}}(s_1, \gamma) = (1-\gamma)(\frac{2}{3} + \frac{3}{4}\gamma) + \frac{1}{2}\gamma^2$ and $\textsc{Power}_{\mathcal{D}_{X\text{-IID}}}(s_2, \gamma) = \frac{1-\gamma}{1-\gamma^5}\left(\frac{1}{2} + \frac{3}{4}\gamma + \frac{2}{3}\gamma^2 + \frac{1}{2}(\gamma^3 + \gamma^4)\right)$ by proposition F.198.

Proposition F.198 (Time-uniform $\textsc{Power}$ bound)**.** *If s is time-uniform, then either all trajectories simultaneously enter 1-cycles or no trajectory ever enters a 1-cycle. Furthermore,*

$$\begin{aligned}\textsc{Power}_{\mathcal{D}_{X\text{-IID}}}(s, \gamma) &= \textsc{UnifPower}(s, \gamma) \\ &\coloneqq (1-\gamma)\sum_{t=1}^{\infty} \gamma^{t-1}\, \mathbb{E}\left[\max \text{ of } \left|\textsc{Reach}(s,t)\right| \text{ draws from } X\right].\end{aligned}$$

Proof. Suppose s is time-uniform. The fact that all trajectories simultaneously enter 1-cycles or no trajectory ever enters a 1-cycle, follows directly from definition F.196.

Starting from s, suppose all policies enter a 1-cycle at timestep k (if no policy enters a 1-cycle, $k = \infty$). Consider any reward function R. The agent starts at timestep 0. At timestep $t < k$, its optimal policy selects greedily from $\left|\textsc{Reach}(s, t+1)\right|$ choices. At $t \geq k$, the agent is in the best of $\left|\textsc{Reach}(s,t)\right|$ 1-cycles. $\square$

Proposition F.199 ($\textsc{Power}_{\mathcal{D}_{X\text{-IID}}}$ bounds)**.**

$$0 < \mathbb{E}\left[X\right] \leq \textsc{Power}_{\mathcal{D}_{X\text{-IID}}}(s, \gamma) \leq \textsc{UnifPower}(s, \gamma) \leq \mathbb{E}\left[\max \text{ of } |\mathcal{S}| \text{ draws from } X\right] < 1.$$

Proof. $\mathbb{E}\left[\max \text{ of } |\mathcal{S}| \text{ draws from } X\right] < 1$ because X is a continuous distribution on the unit interval; similarly for $0 < \mathbb{E}\left[X\right]$. $\textsc{Power}_{\mathcal{D}_{X\text{-IID}}}(s, \gamma) \leq \textsc{UnifPower}(s, \gamma)$ because for each reward function and at each time step t, the agent can (at best) enter the highest-reward state from $\textsc{Reach}(s,t)$. The other inequalities follow directly from lemma F.186 and lemma F.188. $\square$

Corollary F.200 (Time-uniformity implies no optimal policy shifts)**.** *The optimal trajectory cannot shift at time-uniform states.*

Proof. Apply theorem F.128. $\square$

F.7.4 Recursive $\textsc{Power}$ computation

In general, $V^*_{\mathcal{D}_{\text{bound}}}(s,\gamma)$ can be computed by solving for $\left|\mathcal{F}_{\text{nd}}(s)\right|$ convex polytopes (the optimality supports, definition F.79) and integrating the average on-policy value:

$$V^*_{\mathcal{D}_{\text{bound}}}(s,\gamma) = \sum_{\mathbf{f}\in\mathcal{F}_{\text{nd}}(s)} \int_{\text{supp}\left(\mathbf{f}(\gamma)\geq\mathcal{F}(s,\gamma)\right)} \mathbf{f}(\gamma)^\top \mathbf{r}\, \mathrm{d}F(\mathbf{r}). \tag{F.161}$$

However, the optimality supports and the corresponding integrals may be difficult to compute. In certain "tree-like" environments, we can compute $\textsc{Power}_{\mathcal{D}_{\text{bound}}}$ via dynamic programming.

Definition F.201 (Support of a set of visit distributions)**.** Let $F\subseteq\mathcal{F}(s)$. $\text{supp}(F) \coloneqq \left\{s' \mid \exists \mathbf{f}\in F : (\mathbf{f}(.5)-\mathbf{e}_s)^\top \mathbf{e}_{s'} > 0\right\}$ is the set of states s' visited with positive probability by some $\mathbf{f}\in F$.

Definition F.202 (Reward independence)**.** Let $F,F'\subseteq\mathcal{F}(s)$. $F \perp\!\!\!\perp_{\mathcal{D}_{\text{any}}} F'$ (read "*F and F' are reward independent under $\mathcal{D}_{any}$*") when $\forall s\in\text{supp}(F), s'\in\text{supp}(F')$, the reward at s is independent of the reward at s' under $\mathcal{D}_{\text{any}}$.

Theorem F.203 (Factored $\textsc{Power}$ computation)**.** *Let $\mathcal{D}'$ be a probability distribution over reward functions, with probability measure F'. Let $F^1,\ldots,F^k\subseteq\mathcal{F}_{\text{nd}}(s)$ be pairwise reward independent under $\mathcal{D}'$, where $\bigcup_i F^i=\mathcal{F}_{\text{nd}}(s)$. For each F^i and any $\gamma^*\in(0,1)$, define the random variable $X^i_{\gamma^*} \coloneqq \max_{\mathbf{f}\in F^i} \frac{1}{\gamma^*}(\mathbf{f}(\gamma^*)-\mathbf{e}_s)^\top \mathbf{r} \mid \mathbf{r}\sim\mathcal{D}'$ with* CDF $F_{X^i_{\gamma^*}}$.

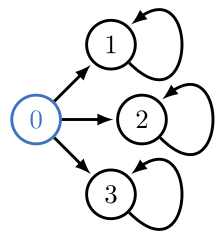


Figure F.32: Let distribution X_a have CDF $F_a(v) := v$ on $v \in [0,1]$ and distribution X_b have CDF $F_b(v) := v^2$ on $v \in [0,1]$. Suppose the reward function distribution $\mathcal{D}'$ is such that $R(s_1), R(s_3) \sim X_a, R(s_2) \sim X_b$. Then theorem F.203 shows that $\text{POWER}_{\mathcal{D}'}(s_0, \gamma) = \int_0^1 v \frac{\mathrm{d}}{\mathrm{d}v}(v \cdot v^2 \cdot v) = \int_0^1 4v^4 \, \mathrm{d}v = \frac{4}{5}$.

Then for any $\gamma \in [0,1]$,

$$\text{POWER}_{\mathcal{D}'}(s,\gamma) = \lim_{\gamma^* \to \gamma} (1-\gamma^*) \int_{-\infty}^{\infty} v \, \mathrm{d}\left(\prod_i F_{X^i_{\gamma^*}}(v)\right). \tag{F.162}$$

Proof. Define the random variable $X^*_\gamma := \max_i X^i_\gamma$ with CDF $F_{X^*_\gamma}$. Consider X^j_γ, X^k_γ, where $j \neq k$. We have

$$X^j_\gamma := \max_{\mathbf{f} \in F^j} \gamma^{-1}(\mathbf{f}(\gamma) - \mathbf{e}_s)^\top \mathbf{r} \mid \mathbf{r} \sim \mathcal{D}' \tag{F.163}$$

$$X^k_\gamma := \max_{\mathbf{f} \in F^k} \gamma^{-1}(\mathbf{f}(\gamma) - \mathbf{e}_s)^\top \mathbf{r} \mid \mathbf{r} \sim \mathcal{D}'. \tag{F.164}$$

For each $s_j \in \text{supp}(F^j), s_k \in \text{supp}(F^k)$, the random variable $R_{s_j} \mid R \sim \mathcal{D}'$ is independent of the random variable $R_{s_k} \mid R \sim \mathcal{D}'$ because we assumed $F^j \perp\!\!\!\perp_{\mathcal{D}'} F^k$ (definition F.202). Then any random variable corresponding to a linear combination $L_j := \sum_{s_j \in \text{supp}(F^j)} \alpha_j R(s_j)$ (where $\alpha_j \in \mathbb{R}$) is also independent of each random variable $R_{s_k} \mid R \sim \mathcal{D}'$. Furthermore, any such L_j is independent of any linear combination $L_k := \sum_{s_k \in \text{supp}(F^k)} \alpha_k R(s_k)$, where $\alpha_k \in \mathbb{R}$.

But each $\gamma^{-1}(\mathbf{f}_j(\gamma) - \mathbf{e}_s)^\top \mathbf{r}$ is precisely such a linear combination. So $X^j_\gamma = \max_n L_n, X^k_\gamma = \max_m L_m$, where each $L_n := \gamma^{-1}(\mathbf{f}_j(\gamma) - \mathbf{e}_s)^\top \mathbf{r}$ is independent of each $L_m := \gamma^{-1}(\mathbf{f}_k(\gamma) -$

$\mathbf{e}_s)^\top \mathbf{r}$. So X^j_γ and X^k_γ are independent under $\mathcal{D}'$. Therefore, $\forall v \in \mathbb{R} : F_{X^*_\gamma}(v) = \prod_i F_{X^i_\gamma}(v)$.

$$\begin{aligned}
V^*_{\mathcal{D}'}(s,\gamma) &:= \int_{\text{supp}(\mathcal{D}')} \max_{\mathbf{f}\in\mathcal{F}_{\text{nd}}(s)} \mathbf{f}(\gamma)^\top \mathbf{r}\, \mathrm{d}F'(\mathbf{r}) && \text{(F.165)}\\
&= \mathop{\mathbb{E}}_{R\sim\mathcal{D}}\left[R(s)\right] + \gamma \int_{\text{supp}(\mathcal{D}')} \max_{\mathbf{f}\in\mathcal{F}_{\text{nd}}(s)} \gamma^{-1}(\mathbf{f}(\gamma)-\mathbf{e}_s)^\top \mathbf{r}\, \mathrm{d}F'(\mathbf{r}) && \text{(F.166)}\\
&= \mathop{\mathbb{E}}_{R\sim\mathcal{D}}\left[R(s)\right] + \gamma \int_{-\infty}^{\infty} v\, \mathrm{d}F_{X^*_\gamma}(v) && \text{(F.167)}\\
&= \mathop{\mathbb{E}}_{R\sim\mathcal{D}}\left[R(s)\right] + \gamma \int_{-\infty}^{\infty} v\, \mathrm{d}\left(\prod_i F_{X^i_\gamma}(v)\right). && \text{(F.168)}
\end{aligned}$$

Equation (F.165) follows from the definition of average optimal reward (definition 5.11) and the fact that restricting maximization to $\mathcal{F}_{\text{nd}}$ leaves optimal value unchanged for all reward functions by definition 5.6. Equation (F.167) follows from the definition of $F_{X^*_\gamma}$. Equation (F.168) follows from the factorization $F_{X^*_\gamma} = \prod_i F_{X^i_\gamma}$, proved above.

Then since $\gamma \in (0,1)$,

$$\begin{aligned}
\textsc{Power}_{\mathcal{D}'}(s,\gamma) &= \frac{1-\gamma}{\gamma}\left(V^*_{\mathcal{D}'}(s,\gamma) - \mathop{\mathbb{E}}_{R\sim\mathcal{D}}\left[R(s)\right]\right) && \text{(F.169)}\\
&= (1-\gamma)\int_{-\infty}^{\infty} v\, \mathrm{d}\left(\prod_i F_{X^i_\gamma}(v)\right). && \text{(F.170)}
\end{aligned}$$

Equation (F.169) follows from definition 5.12. Equation (F.170) follows from eq. (F.168).

Since $\textsc{Power}_{\mathcal{D}'}$ is Lipschitz continuous on $\gamma \in [0,1]$ (lemma 5.13) and since eq. (F.170) holds for all $\gamma \in (0,1)$, the result holds in the limits and therefore holds for all $\gamma \in [0,1]$. $\square$

F.7.5 Complexity of estimating $\textsc{Power}_{\mathcal{D}_{\text{bound}}}$ and optimality probability

Proposition F.204 (Power sampling bounds)**.** *Let $\gamma \in [0,1]$, $\mathcal{D}'$ be a reward function distribution which is bounded $[b,c]^{|\mathcal{S}|}$, s be a state, and $\epsilon > 0$. For the random variable*

$\bar{X}_n^{\text{POWER}_{\mathcal{D}'}} := \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{s'\sim T(s,\pi_i^*(s))}\left[V^*_{R_i,\,norm}\left(s',\gamma\right)\right]$ *for* IID *draws* $R_1,\ldots,R_n \sim \mathcal{D}'$,

$$\mathop{\mathbb{P}}_{R_1,\ldots,R_n\sim\mathcal{D}'}\left(\left|\bar{X}_n^{\text{POWER}_{\mathcal{D}'}} - \text{POWER}_{\mathcal{D}'}\left(s,\gamma\right)\right| \geq \epsilon\right) \leq 2e^{-\frac{2n\epsilon^2}{(c-b)^2}}. \tag{F.171}$$

Proof. Normalized optimal value for each R_i is bounded $[b,c]$. Since the draws $R_i \sim \mathcal{D}'$ are independent, apply Hoeffding's inequality. □

At any fixed discount rate $\gamma \in [0,1)$, an optimal value function can be computed in time polynomial in $|\mathcal{S}|$ and $|\mathcal{A}|$ (via *e.g.* value iteration [49]). Therefore, proposition F.204 shows that when $\gamma \in [0,1)$, $\text{POWER}_{\mathcal{D}_{\text{bound}}}$ can be efficiently approximated with high probability.

Conjecture F.205 (POWER can be efficiently computed)**.**

We show a similar result for optimality probability, except proposition F.206 does not require a bounded reward function distribution.

Proposition F.206 (Optimality probability sampling bounds)**.** *Let* $\gamma \in [0,1]$*,* $F \subseteq \mathcal{F}(s)$*, and* $\epsilon > 0$*. For the random variable* $\bar{X}_n^{\mathbb{P}_{\mathcal{D}_{any}}} := \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\exists \mathbf{f}\in F:\ \mathbf{f} \text{ is optimal for } R_i \text{ at } \gamma}$ *for* IID *draws* $R_1,\ldots,R_n \sim \mathcal{D}_{any}$*,*

$$\mathop{\mathbb{P}}_{R_1,\ldots,R_n\sim\mathcal{D}_{any}}\left(\left|\bar{X}_n^{\mathbb{P}_{\mathcal{D}_{any}}} - \mathbb{P}_{\mathcal{D}_{any}}\left(F,\gamma\right)\right| \geq \epsilon\right) \leq 2e^{-2n\epsilon^2}. \tag{F.172}$$

Proof. $\bar{X}_n^{\mathbb{P}_{\mathcal{D}_{\text{any}}}}$ is an unbiased estimator of $\mathbb{P}_{\mathcal{D}_{\text{any}}}\left(F,\gamma\right)$ and it is bounded $[0,1]$. Since the draws $R_i \sim \mathcal{D}_{\text{any}}$ are independent, apply Hoeffding's inequality. □

Conjecture F.207 (Optimality probability can be efficiently computed)**.**

F.7.6 How POWER relates to empowerment

Definition F.208 (n-step reachable states)**.** Let $\text{REACH}_{\text{sure}}\left(s,t\right)$ be the set of states which are reachable from state s with probability 1 in exactly t time steps.

Conjecture F.209 (A function of the number of reachable states lower-bounds $\textsc{Power}$)**.** Let $S_t \subseteq \textsc{Reach}_{\text{sure}}(s,t)$ be those states only reachable in exactly t steps. For any $\gamma \in [0,1]$,

$$\textsc{Power}_{\mathcal{D}_{\text{bound}}}(s,\gamma) \geq \left(1-(1-\gamma)\gamma^{t-1}\right)\mathbb{E}\left[X\right]+(1-\gamma)\gamma^{t-1}\,\mathbb{E}\left[\max\text{ of } |S_t| \text{ draws from } X\right]. \tag{F.173}$$

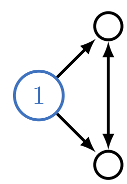


Figure F.33: The $\textsc{Power}_{\mathcal{D}_{X\text{-IID}}}$ expansion of theorem F.182 can have summands less than $\mathbb{E}\left[X\right]$.

The caption of fig. F.33 is justified because

$$\textsc{Power}_{\mathcal{D}_{X\text{-IID}}}(s_1,\gamma) = \frac{1}{1+\gamma}\Big(\mathbb{E}\left[\max\text{ of 2 draws from } X\right]+\gamma\,\mathbb{E}\left[\min\text{ of 2 draws from } X\right]\Big).$$

Although fig. D.1 demonstrates how information-theoretic empowerment fails to capture important non-local information about the agent's control over the environment, $\textsc{Power}_{\mathcal{D}_{\text{bound}}}$ and empowerment are not unrelated. As Salge et al. [77] remark, "In a discrete deterministic world empowerment reduces to the logarithm of the number of sensor states reachable with the available actions." Conjecture F.210 reflects the fact that an agent can at least choose from the highest-reward reachable state after t steps.

Conjecture F.210 (A function of empowerment lower-bounds $\textsc{Power}_{\mathcal{D}_{\text{bound}}}$)**.** If the environment is deterministic and contains an absorbing state, then for any $\gamma \in [0,1]$,

$$\begin{aligned}&\textsc{Power}_{\mathcal{D}_{\text{bound}}}(s,\gamma) \geq \\ &\sup_{t\geq 1}\left(1-(1-\gamma)\gamma^{t-1}\right)\mathbb{E}\left[X\right]+(1-\gamma)\gamma^{t-1}\,\mathbb{E}\left[\max\text{ of } 2^{\mathfrak{E}_t(s)} \text{ draws from } X\right].\end{aligned} \tag{F.174}$$

Proof. Suppose the environment is deterministic and let $\gamma \in [0,1]$. Let t be a positive integer. By Equation 4.15 of Salge et al. [77], $\mathfrak{E}_t(s) = \log_2\left|\textsc{Reach}_{\text{sure}}(s,t)\right|$ (given deterministic dynamics). Since the environment contains an absorbing state, apply conjecture F.209 to conclude that

$$\textsc{Power}_{\mathcal{D}_{\text{bound}}}(s,\gamma) \geq \left(1-(1-\gamma)\gamma^{t-1}\right)\mathbb{E}\left[X\right]+(1-\gamma)\gamma^{t-1}\,\mathbb{E}\left[\max \text{ of } 2^{\mathfrak{E}_t(s)} \text{ draws from } X\right].$$

Since t was arbitrary, $\textsc{Power}_{\mathcal{D}_{\text{bound}}}(s,\gamma)$ must be greater than the supremum over such t. ☐

Figure F.34 shows that this inequality would be tight.

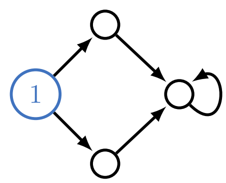


Figure F.34: $\textsc{Power}_{\mathcal{D}_{\text{bound}}}(s_1,\gamma) = (1-\gamma)\,\mathbb{E}\left[\max \text{ of } 2 \text{ draws from } X\right] + \gamma\,\mathbb{E}\left[X\right]$ by proposition F.198. Since $\mathfrak{E}_1(s)=1$, the inequality of conjecture F.210 would be tight.

F.8 Power-seeking

Question F.211 (Probability of Power-seeking being incentivized)**.** Seeking Power is not always more probable under optimality, but we have shown sufficient conditions for when it is. We believe that this relationship often holds, but it is impossible to

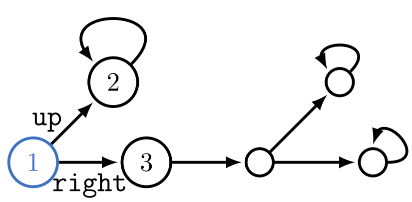


Figure F.35: Policies which go **right** are $\textsc{Power}_{\mathcal{D}_{X\text{-IID}}}$-seeking: $\forall \gamma \in (0,1]$: $\textsc{Power}_{\mathcal{D}_{X\text{-IID}}}(s_3,\gamma) > \textsc{Power}_{\mathcal{D}_{X\text{-IID}}}(s_2,\gamma)$ by lemma F.186 and proposition 5.28. However, for X' with CDF $F(x)=x^2$ on the unit interval, $\mathbb{P}_{\mathcal{D}_{X'\text{-IID}}}(s_1, \texttt{up}, .12) \approx .91$. For $\mathcal{D}_{X'\text{-IID}}$ and at $\gamma = 0.12$, it is more probable that optimal trajectories go `up` through s_2, which has *less* $\textsc{Power}_{\mathcal{D}_{X'\text{-IID}}}$.

graphically characterize when it holds (proposition D.1). For some suitable high-entropy joint distribution over MDP structures (*e.g.* Erdős–Rényi), state reward distributions X, starting states s, and future states s', with what probability is seeking $\textsc{Power}_{\mathcal{D}_{\text{bound}}}$ at s' more probable under optimality, given that the agent starts at s?

F.8.1 Ordering policies based on POWER-seeking

$\textsc{Power}_{\mathcal{D}_{\text{bound}}}$-seeking is not a binary property: it's not true that a policy either does or doesn't seek $\textsc{Power}_{\mathcal{D}_{\text{bound}}}$. The $\textsc{Power}_{\mathcal{D}_{\text{bound}}}$-seeking definition (definition 5.16) accounts for the fact that π might seek a lot of $\textsc{Power}_{\mathcal{D}_{\text{bound}}}$ at s but not seek much $\textsc{Power}_{\mathcal{D}_{\text{bound}}}$ at s' (appendix F.8.1) and that a policy π may seek $\textsc{Power}_{\mathcal{D}_{\text{bound}}}$ for one discount rate but not at another (appendix F.8.2).

The $\textsc{Power}_{\mathcal{D}_{\text{bound}}}$-seeking definition (definition 5.16) implies a total ordering over actions based on how much $\textsc{Power}_{\mathcal{D}_{\text{bound}}}$ they seek at a fixed state s and discount γ.

Definition F.212 ($\geq^{s,\gamma}_{\textsc{Power}_{\mathcal{D}_{\text{bound}}}\text{-seek}}$)**.** $a \geq^{s,\gamma}_{\textsc{Power}_{\mathcal{D}_{\text{bound}}}\text{-seek}} a'$ when

$$\mathop{\mathbb{E}}_{s'\sim T(s,a)}\left[\textsc{Power}_{\mathcal{D}_{\text{bound}}}\left(s',\gamma\right)\right] \geq \mathop{\mathbb{E}}_{s'\sim T(s,a')}\left[\textsc{Power}_{\mathcal{D}_{\text{bound}}}\left(s',\gamma\right)\right]. \tag{F.175}$$

Action *a maximally/minimally seeks* $\textsc{Power}_{\mathcal{D}_{bound}}$ *at* s *and* γ when it is a maximal/minimal element of $\geq^{s,\gamma}_{\textsc{Power}_{\mathcal{D}_{\text{bound}}}\text{-seek}}$.

Figure F.37 illustrated how $\textsc{Power}_{\mathcal{D}_{\text{bound}}}$-seeking depends on γ. Figure F.36 shows how a policy might maximally seek $\textsc{Power}_{\mathcal{D}_{\text{bound}}}$ at s but then minimally seek $\textsc{Power}_{\mathcal{D}_{\text{bound}}}$ at s'; therefore, many policy pairs aren't comparable in their $\textsc{Power}_{\mathcal{D}_{\text{bound}}}$-seeking.

Ultimately, we're interested in the specific situations in which a policy seeks "a lot" of $\textsc{Power}_{\mathcal{D}_{\text{bound}}}$, not whether the policy seeks $\textsc{Power}_{\mathcal{D}_{\text{bound}}}$ "in general." Even so, we can still formalize a good portion of the latter concept. Definition F.213 formalizes the natural POWER-seeking preorder over the policy space Π.

Definition F.213 ($\succeq^{\mathcal{S},\gamma}_{\textsc{Power}_{\mathcal{D}_{\text{bound}}}\text{-seek}}$)**.** $\pi \succeq^{\mathcal{S},\gamma}_{\textsc{Power}_{\mathcal{D}_{\text{bound}}}\text{-seek}} \pi'$ when $\forall s \in \mathcal{S} : \pi(s) \geq^{s,\gamma}_{\textsc{Power}_{\mathcal{D}_{\text{bound}}}\text{-seek}} \pi'(s)$.

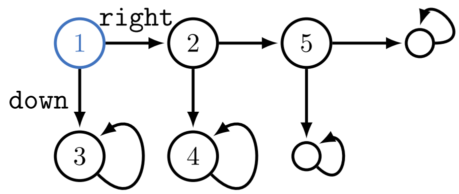


Figure F.36: If $\pi(s_1) = \texttt{right}, \pi(s_2) = \texttt{down}$, then $\forall \gamma \in [0,1]$, π maximally seeks $\text{POWER}_{\mathcal{D}_{\text{bound}}}$ at s_1 but minimally seeks $\text{POWER}_{\mathcal{D}_{\text{bound}}}$ at s_2. Just as a consumer earns money in order to spend it, a policy may gain $\text{POWER}_{\mathcal{D}_{\text{bound}}}$ in order to "spend it" to realize a particular trajectory.

Proposition F.214 ($\succeq^{\mathcal{S},\gamma}_{\text{POWER}_{\mathcal{D}_{\text{bound}}}\text{-seek}}$ is a preorder on Π)**.**

Proof. $\succeq^{\mathcal{S},\gamma}_{\text{POWER}_{\mathcal{D}_{\text{bound}}}\text{-seek}}$ is reflexive and transitive because of the reflexivity and transitivity of the total ordering $\geq^{s,\gamma}_{\text{POWER}_{\mathcal{D}_{\text{bound}}}\text{-seek}}$. □

Proposition F.215 (Existence of a maximally $\text{POWER}_{\mathcal{D}_{\text{bound}}}$-seeking policy)**.** *Let $\gamma \in [0,1]$. $\succeq^{\mathcal{S},\gamma}_{\text{POWER}_{\mathcal{D}_{bound}}\text{-seek}}$ has a greatest element.*

Proof. Construct a policy π such that $\forall s : \pi(s) \in \arg\max_a \mathbb{E}_{s' \sim T(s,a)}\left[\text{POWER}_{\mathcal{D}_{\text{bound}}}\left(s', \gamma\right)\right]$. This is well-defined because $\mathcal{A}$ is finite. □

F.8.2 Seeking POWER at different discount rates

Figure F.37 shows that at any given state, the extent to which an action seeks $\text{POWER}_{\mathcal{D}_{\text{bound}}}$ depends on the discount rate. Greedier optimal policies might tend to accumulate short-term $\text{POWER}_{\mathcal{D}_{\text{bound}}}$ (*i.e.* $\text{POWER}_{\mathcal{D}_{\text{bound}}}(s,\gamma)$ for $\gamma \approx 0$), while Blackwell optimal policies might tend to accumulate long-term $\text{POWER}_{\mathcal{D}_{\text{bound}}}$ (*i.e.* $\text{POWER}_{\mathcal{D}_{\text{bound}}}(s,\gamma)$ for $\gamma \approx 1$).

Lemma F.216 ($\text{POWER}_{\mathcal{D}_{\text{bound}}}$ bounds when $\gamma = 0$)**.** *Let X' be any bounded distribution* *over* $\mathbb{R}$.

$$\mathbb{E}\left[\max\ \text{of}\ \left|Ch_{sure}\left(s\right)\right|\ \text{draws from}\ X'\right] \tag{F.176}$$

$$\leq \textsc{Power}_{\mathcal{D}_{X'\text{-IID}}}\left(s,0\right) \tag{F.177}$$

$$\leq \mathbb{E}\left[\max\ \text{of}\ \left|Ch\left(s\right)\right|\ \text{draws from}\ X'\right]. \tag{F.178}$$

Proof. The left inequality holds because restricting policies to deterministic action at s cannot increase $\textsc{Power}_{\mathcal{D}_{X'\text{-IID}}}\left(s,0\right)$. The right inequality holds because at best, greedy policies deterministically navigate to the child with maximal reward. □

Definition F.217 (Children)**.** The *children* of state s are $Ch(s) \coloneqq \left\{s' \mid \exists a : T(s,a,s') > 0\right\}$.

Proposition F.218 (When $\gamma = 0$ under local determinism, maximally $\textsc{Power}_{\mathcal{D}_{X\text{-IID}}}$-seeking actions lead to states with the most children)**.** *Let X' be a nondegenerate distribution on $\mathbb{R}$. Suppose all actions have deterministic consequences at s and its children. For each action a, let s_a be such that $T(s,a,s_a) = 1$. $\textsc{Power}_{\mathcal{D}_{X'\text{-IID}}}\left(s_a,0\right) = \max_{a' \in \mathcal{A}} \textsc{Power}_{\mathcal{D}_{X'\text{-IID}}}\left(s_{a'},0\right)$ iff $\left|Ch\left(s_a\right)\right| = \max_{a' \in \mathcal{A}} \left|Ch\left(s_{a'}\right)\right|$.*

Proof. Apply the bounds of lemma F.216; by the assumed determinism, $Ch\left(s_a\right) = Ch_{\text{sure}}\left(s_a\right)$ and so $\textsc{Power}_{\mathcal{D}_{X'\text{-IID}}}\left(s_a,0\right) = \mathbb{E}\left[\max\ \text{of}\ \left|Ch\left(s_a\right)\right|\ \text{draws from}\ X\right]$ (similarly for each $s_{a'}$). $\mathbb{E}\left[\max\ \text{of}\ \left|Ch\left(s_a\right)\right|\ \text{draws from}\ X\right]$ is strictly monotonically increasing in $\left|Ch\left(s_a\right)\right|$ by the non-degeneracy of X. □

Figure F.37 illustrates proposition F.218 and proposition F.219.

Proposition F.219 (When $\gamma = 1$, staying put is maximally $\textsc{Power}$-seeking)**.** *Suppose $\exists a \in \mathcal{A} : T(s,a,s) = 1$. When $\gamma = 1$, a is a maximally $\textsc{Power}_{\mathcal{D}_{bound}}$-seeking action at state s.*

Proof. Staying put via action a has an expected $\textsc{Power}$ of $\textsc{Power}_{\mathcal{D}_{\text{bound}}}\left(s,1\right)$. By lemma D.46, $\textsc{Power}_{\mathcal{D}_{\text{bound}}}\left(s,1\right) \geq \max_{a'} \mathbb{E}_{s' \sim T(s,a')}\left[\textsc{Power}_{\mathcal{D}_{\text{bound}}}\left(s',\gamma\right)\right]$. □

When $\gamma = 1$, proposition F.219 implies that the agent cannot expect that any action will increase its $\textsc{Power}_{\mathcal{D}_{\text{bound}}}$.

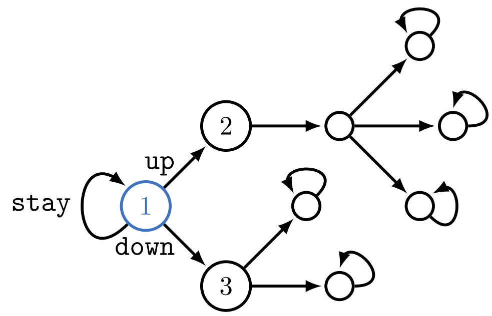


Figure F.37: When $\gamma \approx 0$, $\textsc{Power}_{\mathcal{D}_{X\text{-IID}}}(s_2, \gamma) < \textsc{Power}_{\mathcal{D}_{X\text{-IID}}}(s_3, \gamma)$, and so `down` seeks $\textsc{Power}_{\mathcal{D}_{X\text{-IID}}}$ compared to `up` and `stay` (proposition F.218). When $\gamma \approx 1$, `up` seeks $\textsc{Power}_{\mathcal{D}_{\text{bound}}}$ compared to `down`: $\textsc{Power}_{\mathcal{D}_{X\text{-IID}}}(s_2, \gamma) > \textsc{Power}_{\mathcal{D}_{X\text{-IID}}}(s_3, \gamma)$ (proposition 5.28). However, `stay` is maximally $\textsc{Power}_{\mathcal{D}_{X\text{-IID}}}$-seeking when $\gamma \approx 1$, as demanded by proposition F.219.

F.9 Attainable utility distance

Consider Definition 1 of Turner et al. [98]:

Definition F.220 (Aup reward function)**.** Consider an MDP $\langle \mathcal{S}, \mathcal{A}, T, R, \gamma \rangle$ with state space $\mathcal{S}$, action space $\mathcal{A}$, transition function $T : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$, reward function $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, and discount factor $\gamma \in [0, 1)$. Let $\lambda \geq 0$ and $\varnothing \in \mathcal{A}$, and let $\mathcal{R} \subsetneq \mathbb{R}^{\mathcal{S}}$ be a finite set of auxiliary reward functions.

$$R_{\text{AUP}}(s, a) \coloneqq R(s, a) - \frac{\lambda}{|\mathcal{R}|} \sum_{R_i \in \mathcal{R}} \left| Q^*_{R_i}(s, a) - Q^*_{R_i}(s, \varnothing) \right|. \tag{F.179}$$

The following results provide intuition about how the AUP penalty term works in general. We first formalize a distance metric which is tightly linked to the AUP penalty term [97, 98].

Definition F.221 (Attainable utility distance $d^{\text{AU}}_{\mathcal{D}}$)**.** Let $\mathcal{D}$ be a bounded continuous distribution over reward functions bounded $[0, 1]$, with probability measure F. With respect to $\mathcal{D}$ and discount rate $\gamma \in [0, 1)$, the *attainable utility distance* between state distributions $\Delta, \Delta' \in \Delta(\mathcal{S})$ is

$$d_{\mathcal{D}}^{\text{AU}}\left(\Delta, \Delta' \mid \gamma\right) := \int_{\mathbb{R}^{\mathcal{S}}} \left| \mathop{\mathbb{E}}_{s\sim\Delta}\left[V_R^*\left(s,\gamma\right)\right] - \mathop{\mathbb{E}}_{s'\sim\Delta'}\left[V_R^*\left(s',\gamma\right)\right]\right| \mathrm{d}F(R). \tag{F.180}$$

Each expectation can be interpreted as the Q-value of an action. With respect to reward function distribution $\mathcal{D}$, d_{au} returns the expected advantage magnitude for one action over another.

Proposition F.222 ($d_{\mathcal{D}}^{\text{AU}}$ is a distance metric on $\Delta(\mathcal{S})$)**.**

Proof. For $\Delta, \Delta', \Delta'' \in \Delta(\mathcal{S})$:

1. $d_{\mathcal{D}}^{\text{AU}}\left(\Delta, \Delta' \mid \gamma\right) \geq 0$.
2. $d_{\mathcal{D}}^{\text{AU}}\left(\Delta, \Delta' \mid \gamma\right) = 0$ iff $\Delta = \Delta'$.
3. $d_{\mathcal{D}}^{\text{AU}}\left(\Delta, \Delta' \mid \gamma\right) = d_{\mathcal{D}}^{\text{AU}}\left(\Delta', \Delta \mid \gamma\right)$.
4. $d_{\mathcal{D}}^{\text{AU}}\left(\Delta, \Delta'' \mid \gamma\right) \leq d_{\mathcal{D}}^{\text{AU}}\left(\Delta, \Delta' \mid \gamma\right) + d_{\mathcal{D}}^{\text{AU}}\left(\Delta', \Delta'' \mid \gamma\right)$.

Properties 1 and 3 are trivially true. Property 2 follows from lemma F.105. For property 4,

$$\begin{aligned} d_{\mathcal{D}}^{\text{AU}}\left(\Delta, \Delta'' \mid \gamma\right) &= \int_{\mathbb{R}^{\mathcal{S}}} \left| \left(\mathop{\mathbb{E}}_{s\sim\Delta}\left[V_R^*\left(s,\gamma\right)\right] - \mathop{\mathbb{E}}_{s'\sim\Delta'}\left[V_R^*\left(s',\gamma\right)\right]\right) + \right. & \text{(F.181)} \\ &\qquad \left. \left(\mathop{\mathbb{E}}_{s'\sim\Delta'}\left[V_R^*\left(s',\gamma\right)\right] - \mathop{\mathbb{E}}_{s''\sim\Delta''}\left[V_R^*\left(s'',\gamma\right)\right]\right)\right| \mathrm{d}F(R). \\ &\leq \int_{\mathbb{R}^{\mathcal{S}}} \left| \mathop{\mathbb{E}}_{s\sim\Delta}\left[V_R^*\left(s,\gamma\right)\right] - \mathop{\mathbb{E}}_{s'\sim\Delta'}\left[V_R^*\left(s',\gamma\right)\right]\right| \mathrm{d}F(R) + & \text{(F.182)} \\ &\qquad \int_{\mathbb{R}^{\mathcal{S}}} \left| \mathop{\mathbb{E}}_{s'\sim\Delta'}\left[V_R^*\left(s',\gamma\right)\right] - \mathop{\mathbb{E}}_{s''\sim\Delta''}\left[V_R^*(s'')\right]\right| \mathrm{d}F(R) \\ &= d_{\mathcal{D}}^{\text{AU}}\left(\Delta, \Delta' \mid \gamma\right) + d_{\mathcal{D}}^{\text{AU}}\left(\Delta', \Delta'' \mid \gamma\right). & \text{(F.183)} \end{aligned}$$

$\square$

Restricting $d_{\mathcal{D}}^{\text{AU}}$ to degenerate probability distributions yields a distance metric over the state space.

Viewing the designer as sampling auxiliary reward functions from distribution $\mathcal{D}$, the AUP penalty term is the Monte Carlo integration of $\lambda\gamma \cdot d^{\text{AU}}_{\mathcal{D}}\left(T(s,a), T(s,\varnothing) \mid \gamma\right)$:

$$\frac{\lambda}{|\mathcal{R}|} \sum_{R_i \in \mathcal{R}} \left| Q^*_{R_i}(s,a,\gamma) - Q^*_{R_i}(s,\varnothing,\gamma) \right| \tag{F.184}$$

$$= \frac{\lambda\gamma}{|\mathcal{R}|} \sum_{R_i \in \mathcal{R}} \left| \mathop{\mathbb{E}}_{s_a \sim T(s,a)} \left[V^*_{R_i}(s_a) \right] - \mathop{\mathbb{E}}_{s_\varnothing \sim T(s,\varnothing)} \left[V^*_{R_i}(s_\varnothing) \right] \right|. \tag{F.185}$$

Insofar as the Monte Carlo integration approximates $d^{\text{AU}}_{\mathcal{D}}$, the attainable utility distance sheds light on the attainable utility penalty term in eq. (F.179). For example, we want to penalize side effects, but not smaller changes, such as easily reversible movement. Corollary F.227 guarantees this.

Lemma F.223 (Statewise AU distance inequality)**.**
$d^{\text{AU}}_{\mathcal{D}}\left(\Delta, \Delta' \mid \gamma\right) \leq \mathbb{E}_{s\sim\Delta, s'\sim\Delta'}\left[d^{\text{AU}}_{\mathcal{D}}\left(\mathbf{e}_s, \mathbf{e}_{s'} \mid \gamma\right)\right]$.

Proof.

$$d^{\text{AU}}_{\mathcal{D}}\left(\Delta, \Delta' \mid \gamma\right) := \int_{\mathbb{R}^{\mathcal{S}}} \left| \mathop{\mathbb{E}}_{s\sim\Delta}\left[V^*_R(s,\gamma)\right] - \mathop{\mathbb{E}}_{s'\sim\Delta'}\left[V^*_R(s',\gamma)\right] \right| \mathrm{d}F(R) \tag{F.186}$$

$$\leq \int_{\mathbb{R}^{\mathcal{S}}} \mathop{\mathbb{E}}_{s\sim\Delta, s'\sim\Delta'}\left[\left|V^*_R(s,\gamma) - V^*_R(s',\gamma)\right|\right] \mathrm{d}F(R) \tag{F.187}$$

$$= \mathop{\mathbb{E}}_{s\sim\Delta, s'\sim\Delta'}\left[\int_{\mathbb{R}^{\mathcal{S}}} \left|V^*_R(s,\gamma) - V^*_R(s',\gamma)\right| \mathrm{d}F(R)\right] \tag{F.188}$$

$$= \mathop{\mathbb{E}}_{s\sim\Delta, s'\sim\Delta'}\left[d^{\text{AU}}_{\mathcal{D}}\left(\mathbf{e}_s, \mathbf{e}_{s'} \mid \gamma\right)\right]. \tag{F.189}$$

Equation (F.187) holds by the triangle inequality. Equation (F.188) holds by the linearity of expectation. □

Lemma F.224 (Statewise AU distance upper bound)**.** $\forall s, s' : d^{\text{AU}}_{\mathcal{D}}\left(\mathbf{e}_s, \mathbf{e}_{s'} \mid \gamma\right) < \frac{1}{1-\gamma}$.

Proof. Because optimal value is bounded $[0, \frac{1}{1-\gamma}]$, $d^{\text{AU}}_{\mathcal{D}}\left(\mathbf{e}_s, \mathbf{e}_{s'} \mid \gamma\right) \leq \frac{1}{1-\gamma}$. The equality holds iff for almost all $R \in \operatorname{supp}(\mathcal{D})$, $V^*_R(s, \gamma) = \frac{1}{1-\gamma}$ and $V^*_R(s', \gamma) = 0$, or vice versa. But because $\mathcal{D}$ is continuous, s' must induce positive optimal value for a positive measure set of reward functions. $\square$

Corollary F.225 (AU distance upper bound)**.** $d^{\text{AU}}_{\mathcal{D}}(\Delta, \Delta' \mid \gamma) < \frac{1}{1-\gamma}$.

Lemma F.226 (One-step reachability bounds average difference in optimal value)**.** *Let $\mathcal{D}_{bound}$ be bounded $[b, c]$ and let $\gamma \in [0, 1)$. If s and s' can reach each other with probability 1 in one step, then $\mathbb{E}_{R\sim\mathcal{D}_{bound}}\left[\left|V^*_R(s, \gamma) - V^*_R(s', \gamma)\right|\right] \leq c - b$, with strict inequality if $\mathcal{D}_{bound}$ is continuous.*

Proof. By Proposition 1 of Turner et al. [98] and because each $R \in \operatorname{supp}(\mathcal{D}_{\text{bound}})$ is bounded $[b, c]$, $\left|V^*_R(s, \gamma) - V^*_R(s', \gamma)\right| \leq (1-\gamma)\frac{c-b}{1-\gamma} = c - b$.

Suppose $\mathcal{D}_{\text{bound}}$ is continuous; then $b < c$. For equality to hold, it must be the case that for almost all $R \in \operatorname{supp}(\mathcal{D}_{\text{bound}})$, $\left|V^*_R(s, \gamma) - V^*_R(s', \gamma)\right| = c - b$. Because we assumed that such s and s' can reach each other in one step, this implies that for almost all such R, either $R(s) = b$ and $R(s') = c$, or vice versa. This would imply that $\mathcal{D}_{\text{bound}}$ has a discontinuous marginal reward distribution for these states, which is impossible if $\mathcal{D}_{\text{bound}}$ is continuous. Then the inequality is strict if $\mathcal{D}_{\text{bound}}$ is continuous. $\square$

The following result also applies to the AUP penalty term for any $\mathcal{R}$ over reward functions bounded $[0, 1]$.

Corollary F.227 (Movement penalties are small)**.** *Let $\Delta \neq \Delta'$. Suppose that all states in the support of Δ can deterministically reach in one step all states in the support of Δ', and vice versa. Then $0 < d^{\text{AU}}_{\mathcal{D}}(\Delta, \Delta' \mid \gamma) < 1$.*

Proof. $0 < d^{\text{AU}}_{\mathcal{D}}(\Delta, \Delta' \mid \gamma)$ by proposition F.222.

$$d^{\text{AU}}_{\mathcal{D}}(\Delta, \Delta' \mid \gamma) \leq \mathop{\mathbb{E}}_{s\sim\Delta, s'\sim\Delta'}\left[d^{\text{AU}}_{\mathcal{D}}(\mathbf{e}_s, \mathbf{e}_{s'} \mid \gamma)\right] \tag{F.190}$$

$$< 1. \tag{F.191}$$

$\square$

Equation (F.190) holds by lemma F.223. For eq. (F.191), apply lemma F.226 to conclude that $d^{\text{AU}}_{\mathcal{D}}\left(\mathbf{e}_s, \mathbf{e}_{s'} \mid \gamma\right) < 1$ for each such s, s'. Therefore, $d^{\text{AU}}_{\mathcal{D}}\left(\Delta, \Delta' \mid \gamma\right) < 1$. □

Aup penalizes both seeking and decreasing $\textsc{Power}_{\mathcal{D}_{\text{bound}}}$, compared to the null action.

Proposition F.228 (Change in expected $\textsc{Power}_{\mathcal{D}}$ lower-bounds $d^{\text{AU}}_{\mathcal{D}}$)**.**

$$d^{\text{AU}}_{\mathcal{D}}\left(\Delta, \Delta' \mid \gamma\right) \geq \left| \mathop{\mathbb{E}}_{s\sim\Delta}\left[V^*_{\mathcal{D}_{bound}}\left(s, \gamma\right)\right] - \mathop{\mathbb{E}}_{s'\sim\Delta'}\left[V^*_{\mathcal{D}_{bound}}\left(s', \gamma\right)\right] \right|. \tag{F.192}$$

Proof.

$$d^{\text{AU}}_{\mathcal{D}}\left(\Delta, \Delta' \mid \gamma\right) \coloneqq \mathop{\mathbb{E}}_{R\sim\mathcal{D}}\left[\left| \mathop{\mathbb{E}}_{s\sim\Delta}\left[V^*_R\left(s, \gamma\right)\right] - \mathop{\mathbb{E}}_{s'\sim\Delta'}\left[V^*_R\left(s', \gamma\right)\right] \right|\right] \tag{F.193}$$

$$\geq \left| \mathop{\mathbb{E}}_{R\sim\mathcal{D}}\left[\mathop{\mathbb{E}}_{s\sim\Delta}\left[V^*_R\left(s, \gamma\right)\right] - \mathop{\mathbb{E}}_{s'\sim\Delta'}\left[V^*_R\left(s', \gamma\right)\right] \right]\right| \tag{F.194}$$

$$= \left| \mathop{\mathbb{E}}_{s\sim\Delta}\left[V^*_{\mathcal{D}_{\text{bound}}}\left(s, \gamma\right)\right] - \mathop{\mathbb{E}}_{s'\sim\Delta'}\left[V^*_{\mathcal{D}_{\text{bound}}}\left(s', \gamma\right)\right] \right|. \tag{F.195}$$

Equation (F.194) follows by the reverse triangle inequality. □

While we conjectured that aup penalizes green cell disruption because it decreases Power, Turner et al. [97]'s `Correction` gridworld showed that aup also penalizes increases in Power.

F.9.1 Upper-bounding AU distance by variation distance

Proposition F.229 shows that $d^{\text{AU}}_{\mathcal{D}}$ is upper-bounded by the maximal TV between the visit distributions of s and s'. As fig. F.38 shows, if their visit distributions are "forced to overlap," $d^{\text{AU}}_{\mathcal{D}}$ must be relatively small.

Proposition F.229 (AU distance upper-bounded by maximal variation distance of visit distributions)**.** *Let $\mathcal{D}'$ be any reward function distribution which is bounded $[b, c]$ and let $\gamma \in [0, 1)$.* $d^{\text{AU}}_{\mathcal{D}'}\left(\Delta, \Delta' \mid \gamma\right) \leq (c - b) \max_{\pi\in\Pi} \text{TV}\left(\mathbb{E}_{s\sim\Delta}\left[\mathbf{f}^{\pi,s}(\gamma)\right], \mathbb{E}_{s'\sim\Delta'}\left[\mathbf{f}^{\pi,s'}(\gamma)\right]\right)$.

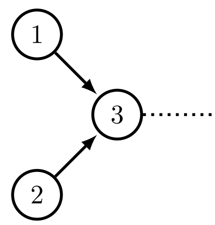


Figure F.38: For all $\mathcal{D}'$ bounded $[0,1]$, $d^{\text{AU}}_{\mathcal{D}'}\left(\mathbf{e}_{s_1}, \mathbf{e}_{s_2} \mid \gamma\right) \leq 1$ by proposition F.229, since for any $\pi \in \Pi$, $\mathbf{f}^{\pi,s_1} \in \mathcal{F}(s_1), \mathbf{f}^{\pi,s_2} \in \mathcal{F}(s_2)$ only disagree on the initial state.

Proof.

$$d^{\text{AU}}_{\mathcal{D}'}\left(\Delta, \Delta' \mid \gamma\right) \coloneqq \mathop{\mathbb{E}}_{R\sim\mathcal{D}'}\left[\left|\mathop{\mathbb{E}}_{s\sim\Delta}\left[V^*_R\left(s,\gamma\right)\right] - \mathop{\mathbb{E}}_{s'\sim\Delta'}\left[V^*_R\left(s',\gamma\right)\right]\right|\right] \tag{F.196}$$

$$= \mathop{\mathbb{E}}_{\mathbf{r}\sim\mathcal{D}'}\left[\left|\max_{\pi\in\Pi}\left(\mathop{\mathbb{E}}_{s\sim\Delta}\left[\mathbf{f}^{\pi,s}(\gamma)\right] - \mathop{\mathbb{E}}_{s'\sim\Delta'}\left[\mathbf{f}^{\pi,s'}(\gamma)\right]\right)^\top \mathbf{r}\right|\right] \tag{F.197}$$

$$\leq \mathop{\mathbb{E}}_{R\sim\mathcal{D}'}\left[\max_{\pi\in\Pi}\frac{1}{2}\left\|\mathop{\mathbb{E}}_{s\sim\Delta}\left[\mathbf{f}^{\pi,s}(\gamma)\right] - \mathop{\mathbb{E}}_{s'\sim\Delta'}\left[\mathbf{f}^{\pi,s'}(\gamma)\right]\right\|_1 (c-b)\right] \tag{F.198}$$

$$= (c-b)\max_{\pi\in\Pi}\text{TV}\left(\mathop{\mathbb{E}}_{s\sim\Delta}\left[\mathbf{f}^{\pi,s}(\gamma)\right], \mathop{\mathbb{E}}_{s'\sim\Delta'}\left[\mathbf{f}^{\pi,s'}(\gamma)\right]\right). \tag{F.199}$$

Equation (F.197) uses a single max because all optimal policies $\pi \in \Pi^*\left(R,\gamma\right)$ induce the same optimal value function. For each $\mathbf{r} \in \operatorname{supp}(\mathcal{D}')$ with $\pi \in \Pi^*\left(R,\gamma\right)$, consider the difference vector $\mathbf{d} \coloneqq \mathbb{E}_{s\sim\Delta}\left[\mathbf{f}^{\pi,s}(\gamma)\right] - \mathbb{E}_{s'\sim\Delta'}\left[\mathbf{f}^{\pi,s'}(\gamma)\right]$. Since $\mathbf{r}$ is bounded $[b,c]$, $\left|\mathbf{d}^\top\mathbf{r}\right|$ is maximized when $\mathbf{r}$ assigns c reward to the positive entries of $\mathbf{d}$, b reward to the negative entries, and c reward to the zero entries.

Since we always have $\left\|\mathbf{f}^{\pi,s}(\gamma)\right\|_1 = \frac{1}{1-\gamma}$ by proposition D.8, the negative and positive entries of $\mathbf{d}$ both have measure equal to $\frac{1}{2}\left\|\mathbb{E}_{s\sim\Delta}\left[\mathbf{f}^{\pi,s}(\gamma)\right] - \mathbb{E}_{s'\sim\Delta'}\left[\mathbf{f}^{\pi,s'}(\gamma)\right]\right\|_1$. Lastly, we maximize over all possible $\pi \in \Pi$. Then eq. (F.198) follows.

Equation (F.199) follows because

$$\text{TV}\left(\mathop{\mathbb{E}}_{s\sim\Delta}\left[\mathbf{f}^{\pi,s}(\gamma)\right], \mathop{\mathbb{E}}_{s'\sim\Delta'}\left[\mathbf{f}^{\pi,s'}(\gamma)\right]\right) = \frac{1}{2}\left\|\mathop{\mathbb{E}}_{s\sim\Delta}\left[\mathbf{f}^{\pi,s}(\gamma)\right] - \mathop{\mathbb{E}}_{s'\sim\Delta'}\left[\mathbf{f}^{\pi,s'}(\gamma)\right]\right\|_1.$$

□

Proposition F.229 shows that no matter the bounded $\mathcal{D}'$, some states *always* must be close in AU distance.

Conjecture F.230 (Proposition F.229 can be extended to only account for policies which induce non-dominated visit distribution functions)**.**

Corollary F.231 (Average optimal value difference is bounded by maximum visit distribution distance)**.** *Let* $\mathcal{D}'$ *be any reward function distribution which is bounded* $[b, c]$ *and let* $\gamma \in [0, 1)$.

$$\left| V^*_{\mathcal{D}'}(s, \gamma) - V^*_{\mathcal{D}'}\left(s', \gamma\right) \right| \leq (c - b) \max_{\pi \in \Pi} \mathrm{TV}\left(\mathbf{f}^{\pi,s}(\gamma), \mathbf{f}^{\pi,s'}(\gamma)\right). \tag{F.200}$$

Proof. By proposition F.228, $\left| V^*_{\mathcal{D}'}(s, \gamma) - V^*_{\mathcal{D}'}\left(s', \gamma\right) \right| \leq d^{\text{AU}}_{\mathcal{D}'}\left(\mathbf{e}_s, \mathbf{e}_{s'} \mid \gamma\right)$.
By proposition F.229, $d^{\text{AU}}_{\mathcal{D}'}\left(\mathbf{e}_s, \mathbf{e}_{s'} \mid \gamma\right) \leq (c - b) \max_{\pi \in \Pi} \mathrm{TV}\left(\mathbf{f}^{\pi,s}(\gamma), \mathbf{f}^{\pi,s'}(\gamma)\right)$. □

F.9.2 AU distance for discount rates close to 1

Definition F.232 (Normalized $d^{\text{AU}}_{\mathcal{D}}$)**.** Let $\gamma \in [0, 1), \Delta, \Delta' \in \Delta(\mathcal{S})$, and let $\mathcal{D}$ be a bounded continuous reward function distribution.

$$d^{\text{AU,norm}}_{\mathcal{D}}\left(\Delta, \Delta' \mid \gamma\right) := (1 - \gamma) d^{\text{AU}}_{\mathcal{D}}\left(\Delta, \Delta' \mid \gamma\right). \tag{F.201}$$

Note that this normalization is order-preserving and $d^{\text{AU,norm}}_{\mathcal{D}'}\left(\Delta, \Delta' \mid \gamma\right) \in [0, 1]$. Proposition F.233 demonstrates that $d^{\text{AU,norm}}_{\mathcal{D}}$ extends to $\gamma = 1$ via the appropriate limit.

Proposition F.233 (For any bounded reward function distribution $\mathcal{D}'$, $d^{\text{AU,norm}}_{\mathcal{D}'}\left(\cdot, \cdot \mid \gamma\right)$ is Lipschitz continuous on $\gamma \in [0, 1]$)**.**

Proof. By lemma D.45, $\lim_{\gamma^* \to \gamma}(1 - \gamma^*) V^*_R(s, \gamma^*)$ is Lipschitz continuous on $\gamma \in [0, 1]$, with Lipschitz constant depending only on $\|\mathbf{r}\|_1$. Let $\gamma \in (0, 1), \Delta, \Delta' \in \Delta(\mathcal{S})$. Since expectation, subtraction, and absolute value preserve Lipschitz continuity under some bounded Lipschitz constant, we conclude the Lipschitz continuity of

$$d^{\text{AU,norm}}_{\mathcal{D}'}\left(\Delta, \Delta' \mid \gamma\right) = (1-\gamma) \mathop{\mathbb{E}}_{R\sim\mathcal{D}'}\left[\left|\mathop{\mathbb{E}}_{s\sim\Delta}\left[V^*_R\left(s,\gamma\right)\right] - \mathop{\mathbb{E}}_{s'\sim\Delta'}\left[V^*_R\left(s',\gamma\right)\right]\right|\right]. \tag{F.202}$$

By the above continuity, we can extend $d^{\text{AU,norm}}_{\mathcal{D}}$ to $\gamma = 1$ via the appropriate limit. □

Corollary F.234 ($d^{\text{AU}}_{\mathcal{D}}$ is continuous on $\gamma \in [0,1)$)**.**

Proof. By definition F.232, $d^{\text{AU}}_{\mathcal{D}}\left(\cdot,\cdot \mid \gamma\right) = \frac{1}{1-\gamma} d^{\text{AU,norm}}_{\mathcal{D}}\left(\cdot,\cdot \mid \gamma\right)$.
By proposition F.233, $d^{\text{AU,norm}}_{\mathcal{D}}\left(\cdot,\cdot \mid \gamma\right)$ is continuous on $\gamma \in [0,1)$, as is $\frac{1}{1-\gamma}$. The space of continuous functions is closed under pointwise multiplication. □

Remark. By the continuity of $d^{\text{AU}}_{\mathcal{D}}$, if $\Delta, \Delta', \Delta'' \in \Delta(\mathcal{S})$ are such that $d^{\text{AU,norm}}_{\mathcal{D}}\left(\Delta, \Delta' \mid 1\right) > d^{\text{AU,norm}}_{\mathcal{D}}\left(\Delta, \Delta'' \mid 1\right)$, then for all γ sufficiently close to 1, $d^{\text{AU}}_{\mathcal{D}}\left(\Delta, \Delta' \mid \gamma\right) > d^{\text{AU}}_{\mathcal{D}}\left(\Delta, \Delta'' \mid \gamma\right)$.

For proposition F.235 and proposition F.236, we use the following shorthand: for any $s_{a'}, s_a, s_\varnothing \in \mathcal{S}$, $D_{a'} \coloneqq \text{RSD}_{\text{nd}}\left(s_{a'}\right), D_a \coloneqq \text{RSD}_{\text{nd}}\left(s_a\right), D_\varnothing \coloneqq \text{RSD}_{\text{nd}}\left(s_\varnothing\right)$.

Proposition F.235 (Losing access to RSDs increases $d^{\text{AU,norm}}_{\mathcal{D}}$)**.** *If $D_{a'} \subseteq D_a \subseteq D_\varnothing$, then $d^{\text{AU,norm}}_{\mathcal{D}}\left(\mathbf{e}_{s_\varnothing}, \mathbf{e}_{s_a} \mid 1\right) \leq d^{\text{AU,norm}}_{\mathcal{D}}\left(\mathbf{e}_{s_\varnothing}, \mathbf{e}_{s_{a'}} \mid 1\right)$. If $D_{a'} \subsetneq D_a$, then the inequality is strict.*

Proof.

$$\begin{aligned} d^{\text{AU,norm}}_{\mathcal{D}}\left(\mathbf{e}_{s_\varnothing}, \mathbf{e}_{s_a} \mid 1\right) &= \int_{\mathbb{R}^{|\mathcal{S}|}} \left|\max_{\mathbf{d}\in D_\varnothing} \mathbf{d}^\top \mathbf{r} - \max_{\mathbf{d}'\in D_a} \mathbf{d}'^\top \mathbf{r}\right| \mathrm{d}F(\mathbf{r}) && \text{(F.203)}\\ &\leq \int_{\mathbb{R}^{|\mathcal{S}|}} \left|\max_{\mathbf{d}\in D_\varnothing} \mathbf{d}^\top \mathbf{r} - \max_{\mathbf{d}'\in D_{a'}} \mathbf{d}'^\top \mathbf{r}\right| \mathrm{d}F(\mathbf{r}) && \text{(F.204)}\\ &= d^{\text{AU,norm}}_{\mathcal{D}}\left(\mathbf{e}_{s_\varnothing}, \mathbf{e}_{s_{a'}} \mid 1\right). && \text{(F.205)} \end{aligned}$$

Equation (F.203) holds by the definition of RSDs (definition 5.26). Since $D_{a'} \subseteq D_a$, for all reward functions $\mathbf{r} \in \mathbb{R}^{|\mathcal{S}|}$,

$$\left|\max_{\mathbf{d}\in D_\varnothing} \mathbf{d}^\top \mathbf{r} - \max_{\mathbf{d}'\in D_a} \mathbf{d}'^\top \mathbf{r}\right| \leq \left|\max_{\mathbf{d}\in D_\varnothing} \mathbf{d}^\top \mathbf{r} - \max_{\mathbf{d}'\in D_{a'}} \mathbf{d}'^\top \mathbf{r}\right|. \tag{F.206}$$

If $D_{a'} \subsetneq D_a$, by proposition D.25, $\mathcal{D}$ assigns positive measure to the set of reward functions for which some $\mathbf{d}'' \in D_a \setminus D_{a'}$ is strictly gain-optimal. Therefore, eq. (F.206) is strict for a positive measure set of reward functions, and so eq. (F.204) holds. □

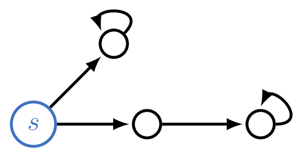


Figure F.39: At $\gamma = 1$, all other states are "equally distant" from s because they each can only access a single 1-cycle RSD. By proposition F.236, $\forall s', s'' \neq s : d_{\mathcal{D}}^{\text{AU,norm}}\left(\mathbf{e}_{s_1}, \mathbf{e}_{s'} \mid 1\right) = d_{\mathcal{D}}^{\text{AU,norm}}\left(\mathbf{e}_{s_1}, \mathbf{e}_{s''} \mid 1\right)$.

Proposition F.236 (Losing access to similar RSDs implies equal $d_{\mathcal{D}}^{\text{AU,norm}}$)**.** *If $D_{a'}, D_a \subseteq D_\varnothing$, if some $D_{sub} \subseteq D_a$ is similar to $D_{a'}$, and if $\left\|D_{a'} - D_\varnothing \setminus D_{a'}\right\|_1 = \left\|D_{sub} - D_\varnothing \setminus D_{sub}\right\|_1 = 2$, then $d_{\mathcal{D}}^{\text{AU},norm}\left(\mathbf{e}_{s_\varnothing}, \mathbf{e}_{s_a} \mid 1\right) \leq d_{\mathcal{D}}^{\text{AU},norm}\left(\mathbf{e}_{s_\varnothing}, \mathbf{e}_{s_{a'}} \mid 1\right)$. If $D_{sub} \subsetneq D_a$, then the inequality is strict.*

Proof. Let ϕ be the guaranteed state permutation such that $\phi \cdot D_{\text{sub}} = D_{a'}$. Define

$$\phi'(s) \coloneqq \begin{cases} \phi(s) & \text{if } s \text{ visited by } \mathbf{d} \in D_{\text{sub}} \\ \phi^{-1}(s) & \text{if } s \text{ visited by } \mathbf{d} \in D_{a'} \\ s & \text{else.} \end{cases} \tag{F.207}$$

By the $\|\cdot\|_1$ assumption, ϕ' is a well-defined permutation.

$$d_{\mathcal{D}}^{\text{AU,norm}}\left(\mathbf{e}_{s_\varnothing}, \mathbf{e}_{s_a} \mid 1\right) = \mathop{\mathbb{E}}_{\mathbf{r}\sim\mathcal{D}}\left[\left|\max_{\mathbf{d}\in D_\varnothing} \mathbf{d}^\top \mathbf{r} - \max_{\mathbf{d}'\in D_a} \mathbf{d}'^\top \mathbf{r}\right|\right] \tag{F.208}$$

$$\leq \mathop{\mathbb{E}}_{\mathbf{r}\sim\mathcal{D}}\left[\left|\max_{\mathbf{d}\in D_\varnothing} \mathbf{d}^\top \mathbf{r} - \max_{\mathbf{d}'\in D_{\text{sub}}} \mathbf{d}'^\top \mathbf{r}\right|\right] \tag{F.209}$$

$$= \mathop{\mathbb{E}}_{\mathbf{r}\sim\mathcal{D}}\left[\left|\max_{\mathbf{d}\in D_\varnothing} \mathbf{d}^\top \mathbf{r}' - \max_{\mathbf{d}'\in D_{a'}} \mathbf{d}' \mathbf{r}'\right|\right] \tag{F.210}$$

$$= d_{\mathcal{D}}^{\text{AU,norm}}\left(\mathbf{e}_{s_\varnothing}, \mathbf{e}_{s_{a'}} \mid 1\right). \tag{F.211}$$

Equation (F.208) holds by the definition of RSDs (definition 5.26). Equation (F.209) holds by proposition F.235, with strict inequality if $D_{\text{sub}} \subsetneq D_a$.

Let $g(b_1, b_2) := |b_1 - b_2|$ and let $f(B_1, B_2 \mid \mathcal{D}) := \mathbb{E}_{\mathbf{r}\sim\mathcal{D}}\left[g\left(\max_{\mathbf{d}\in B_1} \mathbf{d}^\top \mathbf{r},\right), \max_{\mathbf{d}\in B_2} \mathbf{d}^\top \mathbf{r}\right]$. Then lemma D.22 shows that $f(D_\varnothing, D_{\text{sub}} \mid \mathcal{D}) = f(\phi'(D_\varnothing), \phi'(D_{\text{sub}}) \mid \phi'(\mathcal{D}))$.

By the $\|\cdot\|_1$ assumption, $\phi'(D_{\text{sub}}) = D_{a'}$, $\phi'(D_{a'}) = D_{\text{sub}}$, and $\phi'\left(D_\varnothing \setminus (D_{a'} \cup D_{\text{sub}})\right) = D_\varnothing \setminus (D_{a'} \cup D_{\text{sub}})$ by eq. (F.207). Then $\phi'(D_\varnothing) = D_\varnothing$. Since $\mathcal{D}$ distributes reward identically across states, $\phi'(\mathcal{D}) = \mathcal{D}$. We thus conclude that $f(D_\varnothing, D_{\text{sub}} \mid \mathcal{D}) = f(D_\varnothing, D_{a'} \mid \mathcal{D})$, and so eq. (F.210) follows. □

Proposition F.237 (AUP penalty sampling bounds)**.** *Let $\gamma \in [0,1]$, $\mathcal{D}'$ be a reward function distribution which is bounded $[b,c]^{|\mathcal{S}|}$, s be a state, $a, \varnothing$ be actions, and $\epsilon > 0$. For reward functions $R_1, \ldots, R_n$, define $\overline{\left|\Delta Q^{R_1,\ldots,R_n}_{s,a,\varnothing}\right|} := \frac{1}{n}\sum_{i=1}^n \left|Q^*_{R_i,\,norm}(s,a,\gamma) - Q^*_{R_i,\,norm}(s,\varnothing,\gamma)\right|$.*

$$\underset{R_1,\ldots,R_n\sim\mathcal{D}'}{\mathbb{P}}\left(\left|\overline{\left|\Delta Q^{R_1,\ldots,R_n}_{s,a,\varnothing}\right|} - d^{\text{AU},norm}_{\mathcal{D}'}\left(T(s,a), T(s,\varnothing) \mid \gamma\right)\right| \geq \epsilon\right) \leq 2e^{-\frac{2n\epsilon^2}{(c-b)^2}}. \quad \text{(F.212)}$$

Proof. Normalized optimal Q-value is bounded $[b,c]$, and so the absolute difference is bounded $[0, c-b]$. Since the draws $R_i \sim \mathcal{D}'$ are independent, apply Hoeffding's inequality. □

F.10 Proportional regret

We formalize a relaxed variant of worst-case regret minimization which accounts for the human's ability to later correct the robot. We show a common-sense no-free lunch theorem: without any way of learning about the true reward function, no policy can do better than losing half of its value in the worst-case. This result underscores the importance of thinking carefully about what objective we wish to give the agent, and ensuring it can gather enough information about our preferences.

We also show that even when the human may later communicate the true reward function to the robot, it is often impossible to meaningfully minimize worst-case regret. In particular, this occurs when the agent is forced to make an irreversible decision early on.

Regan and Boutilier [71] select policies which minimize worst-case regret against a set of feasible reward functions. We show that when this feasible set is large enough, worst-case regret minimization is infeasible if the agent cannot learn more about the true objective.

Regan and Boutilier [72] quantify regret as the decrease in value from following a suboptimal policy π: $V_R^R(s,\gamma) - V_R^\pi(s,\gamma)$. However, while optimal policies are invariant to positive rescaling of the reward function, absolute regret is not: starting from state s, if π induces 1 regret for reward function R, then π induces 10 regret for $10R$! The *relative regret* [39] of a policy π is $\frac{V_R^*(s,\gamma)-V^\pi(s,\gamma)}{V_R^*(s,\gamma)}$, but the denominator is not invariant to translation of the reward function.

We propose a regret metric which quantifies the *proportion* of attainable value lost by following a suboptimal policy.

Definition F.238 (Minimal value)**.** $V_R^{\min}(s,\gamma) := \min_{\pi\in\Pi} V_R^\pi(s,\gamma) = -V_R^{-R}(s,\gamma)$.

Definition F.239 (Proportional regret)**.** Let s be a state and let π be a policy, $R \in \mathbb{R}^{\mathcal{S}}$, $\gamma \in [0,1]$. The *proportional regret* of following policy π is

$$\textsc{PropRegret}\left(\pi \mid R, s, \gamma\right) := \frac{V_{R,\,\text{norm}}^R(s,\gamma) - V_{R,\text{norm}}^\pi(s,\gamma)}{V_{R,\,\text{norm}}^R(s,\gamma) - V_{R,\,\text{norm}}^{\min}(s,\gamma)}. \tag{F.213}$$

$\textsc{PropRegret}$ is defined to be 0 when the denominator is 0, as no policy can incur regret for R starting from s under such conditions. In particular, this occurs when $\gamma = 0$ in our state-based reward setting: Reward from other states is discounted away, and so all policies are optimal. $\textsc{PropRegret}$ is clearly bounded $[0,1]$.

Conjecture F.240 (Optimal policies have 0 $\textsc{PropRegret}$, while maximally suboptimal policies have 1 $\textsc{PropRegret}$)**.**

Proposition F.241 ($\textsc{PropRegret}$ is invariant to positive affine transformation of the reward function)**.** *Let π be any policy, $R \in \mathbb{R}^{\mathcal{S}}$, $s \in \mathcal{S}$, $\gamma \in [0,1]$, and $m > 0, b \in \mathbb{R}$.*

$$\textsc{PropRegret}\left(\pi \mid R, s, \gamma\right) = \textsc{PropRegret}\left(\pi \mid mR + b, s, \gamma\right). \tag{F.214}$$

Proof. Suppose $\gamma < 1$.

$$\begin{aligned}
&\textsc{PropRegret}\left(\pi \mid R, s, \gamma\right) && \text{(F.215)}\\
&:= \frac{V^{R}_{R,\,\text{norm}}(s,\gamma) - V^{\pi}_{R,\text{norm}}(s,\gamma)}{V^{R}_{R,\,\text{norm}}(s,\gamma) - V^{\min}_{R,\,\text{norm}}(s,\gamma)} && \text{(F.216)}\\
&= \frac{m^{-1}V^{mR+b}_{R,\,\text{norm}}(s,\gamma) - \frac{b}{1-\gamma} - \left(m^{-1}V^{\pi}_{mR+b,\text{norm}}(s,\gamma) - \frac{b}{1-\gamma}\right)}{m^{-1}V^{mR+b}_{R,\,\text{norm}}(s,\gamma) - \frac{b}{1-\gamma} - \left(m^{-1}V^{\min}_{mR+b,\,\text{norm}}(s,\gamma) - \frac{b}{1-\gamma}\right)} && \text{(F.217)}\\
&= \frac{V^{mR+b}_{R,\,\text{norm}}(s,\gamma) - V^{\pi}_{mR+b,\text{norm}}(s,\gamma)}{V^{mR+b}_{R,\,\text{norm}}(s,\gamma) - V^{\min}_{mR+b,\,\text{norm}}(s,\gamma)} && \text{(F.218)}\\
&= \textsc{PropRegret}\left(\pi \mid mR+b, s, \gamma\right). && \text{(F.219)}
\end{aligned}$$

The $\gamma = 1$ case follows automatically from the equality for all $\gamma < 1$. □

Proposition F.242 (Reward function negation flips the PropRegret of any policy)**.** *Let π be any policy, $R \in \mathbb{R}^{\mathcal{S}}$, $s \in \mathcal{S}$, $\gamma \in [0,1]$. If $V^{R}_{R,\,norm}(s,\gamma) > V^{\min}_{R,\,norm}(s,\gamma)$, then*

$$\textsc{PropRegret}\left(\pi \mid -R, s, \gamma\right) = 1 - \textsc{PropRegret}\left(\pi \mid R, s, \gamma\right). \tag{F.220}$$

Proof.

$$\begin{aligned}
&\textsc{PropRegret}\left(\pi \mid -R, s, \gamma\right) && \text{(F.221)}\\
&:= \frac{V^{-R}_{R,\,\text{norm}}(s,\gamma) - V^{\pi}_{-R,\text{norm}}(s,\gamma)}{V^{-R}_{R,\,\text{norm}}(s,\gamma) - V^{\min}_{-R,\,\text{norm}}(s,\gamma)} && \text{(F.222)}\\
&= \frac{-V^{\min}_{R,\,\text{norm}}(s,\gamma) + V^{\pi}_{R,\text{norm}}(s,\gamma)}{-V^{\min}_{R,\,\text{norm}}(s,\gamma) + V^{*}_{R,\,\text{norm}}(s,\gamma)} && \text{(F.223)}\\
&= \frac{V^{\pi}_{R,\text{norm}}(s,\gamma) - V^{\min}_{R,\,\text{norm}}(s,\gamma)}{V^{*}_{R,\,\text{norm}}(s,\gamma) - V^{\min}_{R,\,\text{norm}}(s,\gamma)} && \text{(F.224)}\\
&= 1 + \frac{-\left(V^{*}_{R,\,\text{norm}}(s,\gamma) - V^{\min}_{R,\,\text{norm}}(s,\gamma)\right) + V^{\pi}_{R,\text{norm}}(s,\gamma) - V^{\min}_{R,\,\text{norm}}(s,\gamma)}{V^{*}_{R,\,\text{norm}}(s,\gamma) - V^{\min}_{R,\,\text{norm}}(s,\gamma)} && \text{(F.225)}\\
&= 1 + \frac{-V^{*}_{R,\,\text{norm}}(s,\gamma) + V^{\pi}_{R,\text{norm}}(s,\gamma)}{V^{*}_{R,\,\text{norm}}(s,\gamma) - V^{\min}_{R,\,\text{norm}}(s,\gamma)} && \text{(F.226)}
\end{aligned}$$

$$= 1 - \frac{V^*_{R,\,\text{norm}}(s,\gamma) - V^{\pi}_{R,\text{norm}}(s,\gamma)}{V^*_{R,\,\text{norm}}(s,\gamma) - V^{\min}_{R,\,\text{norm}}(s,\gamma)} \tag{F.227}$$

$$= 1 - \textsc{PropRegret}\left(\pi \mid R, s, \gamma\right). \tag{F.228}$$

□

F.10.1 No free lunch for robust optimization

Let $\mathcal{R}_{\text{true}} \subseteq \mathbb{R}^{\mathcal{S}}$ be a set of reward functions. Robust optimization minimizes maximal regret with respect to this feasible set $\mathcal{R}_{\text{true}}$ [72]. We show that when $\mathcal{R}_{\text{true}}$ has enough reward functions in it, no policy can do well in the worst case. As illustrated by fig. F.40, no policy can simultaneously optimize a generic reward function and its inverse.

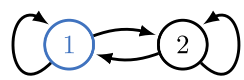


Figure F.40: If $\mathcal{R}_{\text{true}}$ is the set of state indicator reward functions, then no policy can do better than alternating between the two states. When $\gamma = 1$, this policy induces worst-case PropRegret of $\frac{1}{2}$: half of the attainable value is lost for each reward function in $\mathcal{R}_{\text{true}}$.

Theorem F.243 (No free lunch theorem for proportional regret minimization)**.** *Let π be any policy, s a state, and $\gamma \in [0,1]$. If there exists $R \in \mathcal{R}_{true}$ such that a negatively affinely transformed $-mR + b \in \mathcal{R}_{true}$ as well, and if $V^*_{R,\,norm}(s,\gamma) > V^{\min}_{R,\,norm}(s,\gamma)$, then*

$$\sup_{R_{true} \in \mathcal{R}_{true}} \textsc{PropRegret}\left(\pi \mid R_{true}, s, \gamma\right) \geq \frac{1}{2}. \tag{F.229}$$

Proof.

$$\sup_{R_{\text{true}} \in \mathcal{R}_{\text{true}}} \textsc{PropRegret}\left(\pi \mid R_{\text{true}}, s, \gamma\right) \tag{F.230}$$

$$\geq \max_{R_{\text{true}} \in \{R, -mR+b\}} \textsc{PropRegret}\left(\pi \mid R_{\text{true}}, s, \gamma\right) \tag{F.231}$$

$$= \max_{R_{\text{true}} \in \{R, -R\}} \textsc{PropRegret}\left(\pi \mid R_{\text{true}}, s, \gamma\right) \tag{F.232}$$

$$= \max\Big(\textsc{PropRegret}\left(\pi \mid R_{\text{true}}, s, \gamma\right), 1 - \textsc{PropRegret}\left(\pi \mid R_{\text{true}}, s, \gamma\right)\Big) \tag{F.233}$$

$$\geq .5. \tag{F.234}$$

Equation (F.231) holds because $R, -mR + b \in \mathcal{R}_{\text{true}}$. Equation (F.232) holds by proposition F.241. Equation (F.233) holds by proposition F.242, which can be applied since

$$V^*_{R,\,\text{norm}}\left(s, \gamma\right) > V^{\min}_{R,\,\text{norm}}\left(s, \gamma\right).$$

Equation (F.234) holds because PropRegret is bounded $[0, 1]$. ☐

Proposition F.244 (Uninformative $\mathcal{R}_{\text{true}}$ satisfy no-free-lunch conditions)**.** *Let π be any policy and $s \in \mathcal{S}$. Suppose there exist $a < b$ such that $[a, b]^{\mathcal{S}} \subseteq \mathcal{R}_{true}$. If either $\gamma \in (0, 1)$ and $\left|\mathcal{F}(s)\right| > 1$ or $\gamma = 1$ and $\left|\text{RSD}\left(s\right)\right| > 1$, then $\sup_{R_{true} \in \mathcal{R}_{true}} \textsc{PropRegret}\left(\pi \mid R_{true}, s, \gamma\right) \geq \frac{1}{2}$.*

Proof. Let $U := [a, b]^{\mathcal{S}}$ for the assumed real numbers $a < b$. Let $U^- := \{R \in U \mid \exists m_R > 0, b_R \in \mathbb{R} : -m_R R + b_R \in U\}$. Since all $R \in U$ are bounded, *any* reward function $-m_R R + b_R$ can be positively affinely transformed so that its reward is bounded $[a, b]$. Therefore, $U^- = U$.

Since $\gamma > 0$ and either $\left|\mathcal{F}(s)\right| > 1$ or $\left|\text{RSD}\left(s\right)\right| > 1$, by lemma F.106, almost every reward function in U^- has a strictly optimal visit distribution or RSD at discount rate γ. Since $U^- = U$ has positive measure, almost all elements of U^- must have a strictly optimal visit distribution or RSD at discount rate γ. Let $R \in U^-$ be one such element. We conclude that $V^*_{R,\,\text{norm}}\left(s, \gamma\right) > V^{\min}_{R,\,\text{norm}}\left(s, \gamma\right)$.

By the definition of U^-, R has a negatively affinely transformed counterpart in $U \subseteq \mathcal{R}_{\text{true}}$. Then theorem F.243 implies that $\sup_{R_{\text{true}} \in \mathcal{R}_{\text{true}}} \textsc{PropRegret}\left(\pi \mid R_{\text{true}}, s, \gamma\right) \geq .5$. ☐

Remark. Proposition F.244's conditions of $\left|\mathcal{F}(s)\right| > 1$ or $\left|\text{RSD}\left(s\right)\right| > 1$ are trivial: if they are not met, then the agent makes no meaningful choices and PropRegret trivially equals 0 for every policy and reward function.

F.10.2 Corrigible regret minimization

Even though we often can't fully specify the intended reward function or minimize worst-case regret under complete uncertainty, we can correct an agent after watching it make mistakes.

Definition F.245 (Corrigibility)**.** An agent-supervisor pair enables *perfect corrigibility* when the supervisor can modify the agent's policy to any other policy.

Definition F.245 is optimistic: it is obviously unrealistic to demand the supervisor be able to implement in the agent an optimal policy for any reward function. Furthermore, the agent may act to avoid correction [87, 19, 99], or an agent may leave the supervisor's range of correction. We set these complications aside for now.

Theorem F.243 does not imply that ϵ-PropRegret minimization is impossible for $\epsilon < \frac{1}{2}$, even when $\mathcal{R}_{\text{true}} = \mathbb{R}^{\mathcal{S}}$. Theorem F.243 says that any π cannot do well across all reward functions, if π cannot somehow be conditioned on each R_{true}. However, definition F.245 allows the agent to be "corrected" after t time steps to an optimal policy $\pi^*_{R_{\text{true}}}$.

Even if the agent cannot somehow discover which reward function it should optimize, exogenous correction by the supervisor often allows an agent to bound its worst-case proportional regret.

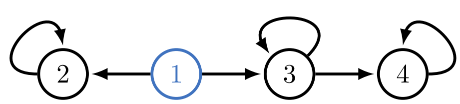


Figure F.41: In some environments, non-trivial corrigible regret minimization is impossible. The agent starts at s_1 (the starting state s is shown in blue). Suppose $\exists R, R' \in \mathcal{R}_{\text{true}} : R(s_2) > R(s_3) \wedge R'(s_2) < R'(s_3)$. Then $\forall \gamma \in (0, 1], t > 0$, no policy can avoid incurring maximal worst-case proportional regret for $\mathcal{R}_{\text{true}}$. However, intuitively, going right is "less option-destroying."

Eysenbach et al. [28] train an agent to preserve initial state reachability. Proposition F.246 shows that reversibility allows the agent to bound worst-case regret, if the agent can later be corrected to pursue the true objective.

Proposition F.246 (Given perfect corrigibility, initial state reachability bounds worst–

case PropRegret). *Let* $\gamma \in [0,1]$ *and let* π *be any policy which, when followed from* s *for* t *steps, has probability* 1 *of residing in states which can reach* s *in* k *steps with probability* 1*.*

$$\sup_{R\in\mathbb{R}^{\mathcal{S}}} \textsc{PropRegret}\left(\pi_{switch}(\pi,\pi^*_R,t)\mid R,s,\gamma\right) \leq 1-\gamma^{t+k}, \tag{F.235}$$

where $\pi^*_R \in \Pi^*(R,\gamma)$ *for each* R*.*

Proof. If $\forall R\in\mathbb{R}^{\mathcal{S}} : V^*_{R,\,\text{norm}}(s,\gamma) = V^{\min}_{R,\,\text{norm}}(s,\gamma)$, then the supremum in eq. (F.235) equals 0 by definition F.239 and we are done. Otherwise, let $Y \subseteq \mathbb{R}^{\mathcal{S}}$ be the subset of reward functions for which this equality does not hold.

Suppose $\gamma \in [0,1)$. Let π_{return} be a policy which always navigates to s as quickly as possible, when possible. Let $\pi_{\text{recover}}(\pi') := \pi_{\text{switch}}\left(\pi_{\text{switch}}(\pi,\pi_{\text{return}},t),\pi',t+k\right)$ be the non-stationary policy which follows π for the first t time steps, switches to π_{return} for the next k time steps, and then follows π' thereafter. Starting from s, the value gained before time $t+k$ is then $G^{\text{recover}}_R(\gamma) := V_R^{\pi_{\text{recover}}(\pi,\pi_{\text{return}},\pi^*_R)}(s,\gamma) - \gamma^{t+k}V^*_R(s,\gamma)$.

$$\begin{aligned}
&\sup_{R\in\mathbb{R}^{\mathcal{S}}} \textsc{PropRegret}\left(\pi_{\text{switch}}(\pi,\pi^*_R,t)\mid R,s,\gamma\right) && \text{(F.236)}\\
&= \sup_{R\in Y} \textsc{PropRegret}\left(\pi_{\text{switch}}(\pi,\pi^*_R,t)\mid R,s,\gamma\right) && \text{(F.237)}\\
&:= \sup_{R\in Y} \frac{V^R_{R,\,\text{norm}}(s,\gamma) - V^{\pi_{\text{switch}}(\pi,\pi^*_R,t)}_{R,\text{norm}}(s,\gamma)}{V^R_{R,\,\text{norm}}(s,\gamma) - V^{\min}_{R,\,\text{norm}}(s,\gamma)} && \text{(F.238)}\\
&= \sup_{R\in Y} \frac{V^R_R(s,\gamma) - V^{\pi_{\text{switch}}(\pi,\pi^*_R,t)}_R(s,\gamma)}{V^R_R(s,\gamma) - V^{\min}_R(s,\gamma)} && \text{(F.239)}\\
&\leq \sup_{R\in Y} \frac{V^R_R(s,\gamma) - V^{\pi_{\text{recover}}(\pi,\pi_{\text{return}},\pi^*_R)}_R(s,\gamma)}{V^R_R(s,\gamma) - V^{\min}_R(s,\gamma)} && \text{(F.240)}\\
&\leq \sup_{R\in Y} \frac{V^R_R(s,\gamma) - V^{\pi_{\text{recover}}(\pi,\pi_{\text{return}},\pi^*_R)}_R(s,\gamma)}{V^R_R(s,\gamma) - \frac{1}{1-\gamma^{t+k}}G^{\text{recover}}_R(\gamma)} && \text{(F.241)}\\
&= \sup_{R\in Y} \frac{V^R_R(s,\gamma) - G^{\text{recover}}_R(\gamma) - \gamma^{t+k}V^*_R(s,\gamma)}{V^R_R(s,\gamma) - \frac{1}{1-\gamma^{t+k}}G^{\text{recover}}_R(\gamma)} && \text{(F.242)}
\end{aligned}$$

$$= \sup_{R \in Y} \frac{\left(1-\gamma^{t+k}\right) V_R^*(s,\gamma) - G_R^{\text{recover}}(\gamma)}{V_R^R(s,\gamma) - \frac{1}{1-\gamma^{t+k}} G_R^{\text{recover}}(\gamma)} \tag{F.243}$$

$$= 1-\gamma^{t+k}. \tag{F.244}$$

Equation (F.237) follows because all reward functions $R' \in \mathbb{R}^{\mathcal{S}} \setminus Y$ have 0 PropRegret by the definition of Y, and PropRegret is bounded $[0,1]$. Equation (F.239) follows because the continuity of (optimal) value functions on $\gamma \in [0,1)$ allows us to ignore the limit in the normalized value functions (definition F.10). Equation (F.240) holds because

$$V_R^{\pi_{\text{recover}}(\pi,\pi_{\text{return}},\pi_R^*)}(s,\gamma) \leq V_R^{\pi_{\text{switch}}(\pi,\pi_R^*,t)}(s,\gamma),$$

as $\pi_{\text{recover}}(\pi,\pi_{\text{return}},\pi_R^*)$ takes longer to begin following an optimal policy for R.

By the definition of $V_R^{\min}(s,\gamma)$, $V_R^{\min}(s,\gamma) \leq \frac{1}{1-\gamma^{t+k}} G_R^{\text{recover}}(\gamma)$, the value of forever alternating between following π for t steps and π_{return} for k steps. Therefore, eq. (F.241) holds.

If $\gamma = 1$, then for all $R \in \mathbb{R}^{\mathcal{S}}$,

$$V_{R,\,\text{norm}}^*(s,1) = V_{R,\text{norm}}^{\pi_{\text{recover}}(\pi,\pi_{\text{return}},\pi_R^*)}(s,1) \tag{F.245}$$

$$\leq V_{R,\text{norm}}^{\pi_{\text{switch}}(\pi,\pi_R^*,t)}(s,1) \tag{F.246}$$

$$\leq V_{R,\,\text{norm}}^*(s,1). \tag{F.247}$$

Equation (F.245) holds by definition F.10: since s can be returned to within k steps, the transient reward from the first $t+k$ steps does not affect the normalized $V_{R,\text{norm}}^{\pi_{\text{recover}}(\pi,\pi_{\text{return}},\pi_R^*)}(s,1)$. Equation (F.247) holds by the definition of normalized optimal value. Therefore,

$$\text{PropRegret}\left(\pi_{\text{switch}}(\pi,\pi_R^*,t) \mid R,s,1\right) = 0,$$

and the supremum in eq. (F.235) also equals 0. $\square$

Proposition F.246 shows that as we take longer to correct the agent, or the agent takes longer to undo its actions, PropRegret increases. On the other hand, as the discount rate increases to 1, PropRegret decreases because transient mistakes become relatively less important. Figure F.42 shows that the inequality in proposition F.246 is sharp.

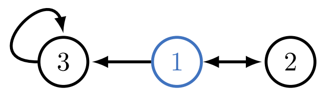


Figure F.42: Let $R(s_1) := 0, R(s_2) := .5, R(s_3) := 1$, and let π_{right} go right at s_1. $\text{PropRegret}\left(\pi_{\text{switch}}(\pi_{\text{right}}, \pi^*_R, 1) \mid R, s, \gamma\right) = 1 - \gamma^2$. At s_2, $k = 1$ step is required to return to the initial state s_1, and $t = 1$. Therefore, $1-\gamma^2 = 1-\gamma^{t+k}$, so proposition F.246's bound is tight.

Definition F.247 (Communicating MDP)**.** An MDP is *communicating* when every state is able to reach every other state with positive probability.

Proposition F.248 (Given perfect corrigibility, all policies are low-regret in communicating MDPs for $\gamma = 1$)**.** *Suppose the environment is communicating, let π be any policy, and let $\epsilon > 0$. If the agent can be corrected within t time steps, then*

$$\sup_{R \in \mathbb{R}^{\mathcal{S}}} \text{PropRegret}\left(\pi_{switch}(\pi, \pi^*_R, t) \mid R, s, 1\right) = 0.$$

Proof. Since the MDP is communicating, $\forall R \in \mathbb{R}^{\mathcal{S}}, s, s' \in \mathcal{S} : V^*_{R,\,\text{norm}}(s, 1) = V^*_{R,\,\text{norm}}(s', 1)$ by lemma F.251. This implies that $V^*_{R,\,\text{norm}}(s, 1) = V^{\pi_{\text{switch}}(\pi, \pi^*_R, t)}_{R,\text{norm}}(s, 1)$, because the agent switches to an optimal policy after t time steps (transient reward differences vanish in the $\gamma = 1$ case). This implies that $\forall R \in \mathbb{R}^{\mathcal{S}} : \text{PropRegret}\left(\pi_{\text{switch}}(\pi, \pi^*_R, t) \mid R, s, 1\right) = 0$. ∎

Conjecture F.249 (Given perfect corrigibility, all policies are low-regret in communicating MDPs for $\gamma \approx 1$)**.** Suppose the environment is communicating, let π be any policy, and let $\epsilon > 0$. If the agent can be corrected within t time steps, then there exists some $\gamma < 1$ such that

$$\sup_{R \in \mathbb{R}^{\mathcal{S}}} \text{PropRegret}\left(\pi_{\text{switch}}(\pi, \pi^*_R, t) \mid R, s, \gamma\right) < \epsilon. \tag{F.248}$$

Suppose that the human designers have uncertainty about what reward function they should provide, with the uncertainty represented by a probability distribution $\mathcal{D}$.

Proposition F.250 (Worst-case PropRegret minimization is equivalent to robustness against $\mathcal{D}$)**.** *Let* π *be any policy and let* $\mathcal{R}_{true} \subseteq \mathbb{R}^{\mathcal{S}}$*.*

$$\sup_{\mathcal{D}\in\Delta(\mathcal{R}_{true})} \mathop{\mathbb{E}}_{R\sim\mathcal{D}}\left[\textsc{PropRegret}\left(\pi \mid R, s, \gamma\right)\right] = \sup_{R\in\mathcal{R}_{true}} \textsc{PropRegret}\left(\pi \mid R, s, \gamma\right). \tag{F.249}$$

Proof. Suppose $(R_n)_{n\geq 1}$ is such that $\forall n : R_n \in \mathcal{R}_{\text{true}}$ and

$$\lim_{n\to\infty} \textsc{PropRegret}\left(\pi \mid R_n, s, \gamma\right) = \sup_{R\in\mathcal{R}_{\text{true}}} \textsc{PropRegret}\left(\pi \mid R, s, \gamma\right).$$

Then let $(\mathcal{D}_n)_{n\geq 1}$ be a sequence of degenerate probability distributions which place probability 1 on R_n. Then each $\mathcal{D}_n \in \Delta(\mathcal{R}_{\text{true}})$. Furthermore,

$$\begin{aligned}
&\sup_{\mathcal{D}\in\Delta(\mathcal{R}_{\text{true}})} \mathop{\mathbb{E}}_{R\sim\mathcal{D}}\left[\textsc{PropRegret}\left(\pi \mid R, s, \gamma\right)\right] && \text{(F.250)}\\
&\geq \lim_{n\to\infty} \mathop{\mathbb{E}}_{R\sim\mathcal{D}_n}\left[\textsc{PropRegret}\left(\pi \mid R_n, s, \gamma\right)\right] && \text{(F.251)}\\
&= \lim_{n\to\infty} \textsc{PropRegret}\left(\pi \mid R_n, s, \gamma\right) && \text{(F.252)}\\
&= \sup_{R\in\mathcal{R}_{\text{true}}} \textsc{PropRegret}\left(\pi \mid R, s, \gamma\right). && \text{(F.253)}
\end{aligned}$$

On the other hand,

$$\begin{aligned}
&\sup_{\mathcal{D}\in\Delta(\mathcal{R}_{\text{true}})} \mathop{\mathbb{E}}_{R\sim\mathcal{D}}\left[\textsc{PropRegret}\left(\pi \mid R, s, \gamma\right)\right] && \text{(F.254)}\\
&\leq \sup_{\mathcal{D}\in\Delta(\mathcal{R}_{\text{true}})} \sup_{R\in\text{supp}(\mathcal{D})} \textsc{PropRegret}\left(\pi \mid R, s, \gamma\right) && \text{(F.255)}\\
&\leq \sup_{\mathcal{D}\in\Delta(\mathcal{R}_{\text{true}})} \sup_{R\in\mathcal{R}_{\text{true}}} \textsc{PropRegret}\left(\pi \mid R, s, \gamma\right) && \text{(F.256)}\\
&= \sup_{R\in\mathcal{R}_{\text{true}}} \textsc{PropRegret}\left(\pi \mid R, s, \gamma\right). && \text{(F.257)}
\end{aligned}$$

Equation (F.256) follows because $\text{supp}(\mathcal{D}) \subseteq \mathcal{R}_{\text{true}}$ by the definition of $\Delta(\mathcal{R}_{\text{true}})$, the set of all probability distributions over $\mathcal{R}_{\text{true}}$.

Therefore, the equality of eq. (F.249) holds. $\square$

However, fig. F.41 shows that in non-communicating environments, robustness against $\mathcal{D}$ is too restrictive, even if we assume perfect corrigibility. Therefore, we step away from worst-case regret minimization.

Lemma F.251 (Equal optimal average reward in communicating MDPs)**.** *If the environment is communicating, then* $\forall R \in \mathbb{R}^{\mathcal{S}}, s, s' \in \mathcal{S} : V^*_{R,\,norm}(s, 1) = V^*_{R,\,norm}(s', 1)$.

Proof. Since the MDP is communicating, s can reach s' with positive probability after at most $|\mathcal{S}|$ timesteps under some policy π. If an agent following π has not reached s' within $|\mathcal{S}|$ timesteps, all states can reach s' with positive probability and so π^{HD} once again attempts to navigate to s'. Since the MDP is finite, there is a state with a minimal (but positive) probability p of reaching s' within $|\mathcal{S}|$ time steps. Because p is positive and minimal, π^{HD} has probability at most $\prod_{t=1}^{\infty}(1-p)^t = 0$ of not reaching s' eventually. Apply lemma F.177 to conclude that $V^*_{R,\,\text{norm}}(s, 1) \geq V^*_{R,\,\text{norm}}(s', 1)$.

The proof for s' reaching s is similar, and so $V^*_{R,\,\text{norm}}(s, 1) = V^*_{R,\,\text{norm}}(s', 1)$. ☐

Conjecture F.252 (PropRegret is piecewise rational on $\gamma \in [0, 1]$)**.**

F.11 Varying the reward function distribution

Proposition F.253 (Power, attainable utility distance, and optimality probability are convex over mixture distributions)**.** *Let* $\mathcal{D}_1, \mathcal{D}_2$ *be two bounded reward function distributions, let* $\theta \in [0, 1]$, *and let* $\mathcal{D}' := \theta\mathcal{D}_1 + (1-\theta)\mathcal{D}_2$ *be a mixture distribution of the two. Let* s *be any state and* $\gamma \in [0, 1]$.

1. $\text{Power}_{\mathcal{D}'}(s, \gamma) = \theta\text{Power}_{\mathcal{D}_1}(s, \gamma) + (1-\theta)\text{Power}_{\mathcal{D}_2}(s, \gamma)$.
2. $\forall \Delta_1, \Delta_2 \in \Delta(\mathcal{S})$:

$$d^{\text{AU},norm}_{\mathcal{D}'}\left(\Delta_1, \Delta_2 \mid \gamma\right) = \theta d^{\text{AU},norm}_{\mathcal{D}_1}\left(\Delta_1, \Delta_2 \mid \gamma\right) + (1-\theta) d^{\text{AU},norm}_{\mathcal{D}_2}\left(\Delta_1, \Delta_2 \mid \gamma\right). \tag{F.258}$$

3. $\forall F \subseteq \mathcal{F}(s) : \mathbb{P}_{\mathcal{D}'}(F, \gamma) = \theta\, \mathbb{P}_{\mathcal{D}_1}(F, \gamma) + (1-\theta)\, \mathbb{P}_{\mathcal{D}_2}(F, \gamma)$.

Proof. Item 1: suppose $\gamma \in (0,1)$. Then

$$\textsc{Power}_{\mathcal{D}'}(s,\gamma) \tag{F.259}$$

$$:= \mathop{\mathbb{E}}_{\mathbf{r}\sim\mathcal{D}'}\left[\max_{\mathbf{f}\in\mathcal{F}(s)} \frac{1-\gamma}{\gamma}\left(\mathbf{f}(\gamma)-\mathbf{e}_s\right)^\top \mathbf{r}\right] \tag{F.260}$$

$$= \theta \mathop{\mathbb{E}}_{\mathbf{r}\sim\mathcal{D}_1}\left[\max_{\mathbf{f}\in\mathcal{F}(s)} \frac{1-\gamma}{\gamma}\left(\mathbf{f}(\gamma)-\mathbf{e}_s\right)^\top \mathbf{r}\right] + (1-\theta) \mathop{\mathbb{E}}_{\mathbf{r}\sim\mathcal{D}_2}\left[\max_{\mathbf{f}\in\mathcal{F}(s)} \frac{1-\gamma}{\gamma}\left(\mathbf{f}(\gamma)-\mathbf{e}_s\right)^\top \mathbf{r}\right] \tag{F.261}$$

$$= \theta\textsc{Power}_{\mathcal{D}_1}(s,\gamma) + (1-\theta)\textsc{Power}_{\mathcal{D}_2}(s,\gamma). \tag{F.262}$$

Since eq. (F.262) holds for arbitrary $\gamma \in (0,1)$, it must hold in the limits as $\gamma \to 0$ and $\gamma \to 1$; the limits of $\textsc{Power}_{\mathcal{D}_{\text{bound}}}$ exist by lemma 5.13. Then item 1 follows.

Similar logic proves item 2 via the linearity of expectation over reward functions.

Consider item 3. By the definition of optimality probability (definition 5.9), some $\mathbf{f} \in F$ is optimal at discount rate γ with probability $\mathbb{P}_{\mathcal{D}_1}(F,\gamma)$ when R is drawn from $\mathcal{D}_1$ and with probability $\mathbb{P}_{\mathcal{D}_2}(F,\gamma)$ when R is drawn from $\mathcal{D}_2$. Then by the definition of the mixture distribution $\mathcal{D}'$, the total probability of this event is $\mathbb{P}_{\mathcal{D}'}(F,\gamma) = \theta\,\mathbb{P}_{\mathcal{D}_1}(F,\gamma) + (1-\theta)\,\mathbb{P}_{\mathcal{D}_2}(F,\gamma)$. This proves item 3. □

Corollary F.254 (Convexity in the space of probability distributions)**.** *Let $\mathcal{D}_1, \mathcal{D}_2$ be two bounded reward function distributions, let $\theta \in [0,1]$, and let $\mathcal{D}' := \theta\mathcal{D}_1 + (1-\theta)\mathcal{D}_2$ be a mixture distribution of the two. Let s be any state, let $\gamma \in [0,1]$, and let $k \in \mathbb{R}$.*

1. *If $\textsc{Power}_{\mathcal{D}_1}(s,\gamma), \textsc{Power}_{\mathcal{D}_2}(s,\gamma) \geq k$, then $\textsc{Power}_{\mathcal{D}'}(s,\gamma) \geq k$.*
2. *Let $\Delta_1, \Delta_2 \in \Delta(\mathcal{S})$. If $d^{\text{AU},norm}_{\mathcal{D}_1}\left(\Delta_1, \Delta_2 \mid \gamma\right), d^{\text{AU},norm}_{\mathcal{D}_2}\left(\Delta_1, \Delta_2 \mid \gamma\right) \geq k$ then*

$$d^{\text{AU},norm}_{\mathcal{D}'}\left(\Delta_1, \Delta_2 \mid \gamma\right) \geq k.$$

3. *Let $F \subseteq \mathcal{F}(s)$. If $\mathbb{P}_{\mathcal{D}_1}(F,\gamma)$, $\mathbb{P}_{\mathcal{D}_2}(F,\gamma) \geq k$, then $\mathbb{P}_{\mathcal{D}'}(F,\gamma) \geq k$.*

Proof. All items follow directly from proposition F.253. □

Proposition F.255 ($\textsc{Power}_{\mathcal{D}_{\text{bound}}}$ difference bounded by total variation distance)**.** *Let $\mathcal{D}_1, \mathcal{D}_2$ be bounded reward function distributions on* $[0,1]^{|\mathcal{S}|}$.

$$\left|\textsc{Power}_{\mathcal{D}_1}(s,\gamma) - \textsc{Power}_{\mathcal{D}_2}(s,\gamma)\right| \leq \text{TV}\left(\mathcal{D}_1, \mathcal{D}_2\right). \tag{F.263}$$

Proof. If $\text{TV}\left(\mathcal{D}_1, \mathcal{D}_2\right) = 0$, then $\textsc{Power}_{\mathcal{D}_1}(s,\gamma) = \textsc{Power}_{\mathcal{D}_2}(s,\gamma)$ and the statement holds.

Suppose $\text{TV}\left(\mathcal{D}_1, \mathcal{D}_2\right) > 0$. Letting $\mathcal{D}_1, \mathcal{D}_2$ respectively correspond to probability measures F_1, F_2, consider the finite signed probability measure $F_{\text{diff}} := F_1 - F_2$. The positive sets of F_{diff} are the sets to which F_1 assigns more probability; vice versa for the negative sets and F_2. By the Hahn decomposition theorem, there exist (non-negative) measures F^+, F^- such that $F_{\text{diff}} = F^+ - F^-$.

By the fact that F_1, F_2 are probability measures with support contained in $[0,1]^{|\mathcal{S}|}$,

$$F_1([0,1]^{|\mathcal{S}|}) - F_2([0,1]^{|\mathcal{S}|}) = 0 \tag{F.264}$$

$$= F_{\text{diff}}\left([0,1]^{|\mathcal{S}|}\right) \tag{F.265}$$

$$= F^+\left([0,1]^{|\mathcal{S}|}\right) - F^-\left([0,1]^{|\mathcal{S}|}\right). \tag{F.266}$$

Therefore, $F^+\left([0,1]^{|\mathcal{S}|}\right) = F^-\left([0,1]^{|\mathcal{S}|}\right)$. Furthermore, they both equal $\text{TV}\left(\mathcal{D}_1, \mathcal{D}_2\right) > 0$ by the definition of total variation distance for probability measures. Let probability measures $F^+_{\text{renorm}}, F^-_{\text{renorm}}$ be the renormalized versions of the non-negative measures F^+, F^-; renormalization is possible because both measures assign finite positive probability to $[0,1]^{|\mathcal{S}|}$.

Let $\gamma \in (0,1)$ and let s be arbitrary. Let $f(R) := \frac{1-\gamma}{\gamma}\left(V^*_R(s,\gamma) - R(s)\right)$.

$$\left|\textsc{Power}_{\mathcal{D}_1}(s,\gamma) - \textsc{Power}_{\mathcal{D}_2}(s,\gamma)\right| \tag{F.267}$$

$$= \left|\int_{[0,1]^{|\mathcal{S}|}} f(R)\,\mathrm{d}F_1(R) - \int_{[0,1]^{|\mathcal{S}|}} f(R)\,\mathrm{d}F_2(R)\right| \tag{F.268}$$

$$= \left|\int_{[0,1]^{|\mathcal{S}|}} f(R)\left(\mathrm{d}F_1(R) - \mathrm{d}F_2(R)\right)\right| \tag{F.269}$$

$$= \left| \int_{[0,1]^{|\mathcal{S}|}} f(R) \, \mathrm{d}F_{\text{diff}}(R) \right| \tag{F.270}$$

$$= \left| \int_{[0,1]^{|\mathcal{S}|}} f(R) \left(\mathrm{d}F^{+}(R) - \mathrm{d}F^{-}(R) \right) \right| \tag{F.271}$$

$$= \left| F^{+}\left([0,1]^{|\mathcal{S}|} \right) \mathop{\mathbb{E}}_{R \sim F^{+}_{\text{renorm}}} \left[f(R) \right] - F^{-}\left([0,1]^{|\mathcal{S}|} \right) \mathop{\mathbb{E}}_{R \sim F^{-}_{\text{renorm}}} \left[f(R) \right] \right| \tag{F.272}$$

$$= \mathrm{TV}\left(\mathcal{D}_1, \mathcal{D}_2 \right) \left| \mathop{\mathbb{E}}_{R \sim F^{+}_{\text{renorm}}} \left[f(R) \right] - \mathop{\mathbb{E}}_{R \sim F^{-}_{\text{renorm}}} \left[f(R) \right] \right| \tag{F.273}$$

$$\leq \mathrm{TV}\left(\mathcal{D}_1, \mathcal{D}_2 \right). \tag{F.274}$$

Equation (F.272) follows by the definitions of $F^{+}_{\text{renorm}}, F^{-}_{\text{renorm}}$. Equation (F.273) follows from the fact that $F^{+}\left([0,1]^{|\mathcal{S}|} \right) = F^{-}\left([0,1]^{|\mathcal{S}|} \right) = \mathrm{TV}\left(\mathcal{D}_1, \mathcal{D}_2 \right) > 0$. Equation (F.274) follows from the fact that f is bounded $[0,1]$, since its domain is $[0,1]^{|\mathcal{S}|}$. Then eq. (F.263) holds.

Since eq. (F.263) holds for all $\gamma \in (0,1)$, it also holds in the limits $\gamma \to 0$ and $\gamma \to 1$. $\square$

Conjecture F.256 (Improved total variation bound)**.** Let $F_{1,v}(v)$ be the optimal value CDF of $\mathcal{D}_1$ at state s and discount rate γ; similarly define $F_{2,v}(v)$. The bound of eq. (F.263) can be improved to

$$\mathrm{TV}\left(\mathcal{D}_1, \mathcal{D}_2 \right) \max \left(\int_{F_{1,v}^{-1}(1-\mathrm{TV}(\mathcal{D}_1, \mathcal{D}_2))}^{1} v \, \mathrm{d}F^{+}_{\text{renorm}}(v), 1 - \int_{0}^{F_{2,v}^{-1}(\mathrm{TV}(\mathcal{D}_1, \mathcal{D}_2))} v \, \mathrm{d}F^{-}_{\text{renorm}}(v) \right).$$

The bound of proposition F.255 is sharp; suppose $\mathcal{D}_1$ puts probability 1 on the all-1 reward function, while $\mathcal{D}_2$ puts probability 1 on the all-0 reward function. Then in any MDP, at any s and for any $\gamma \in [0,1]$,

$$\left| \mathrm{POWER}_{\mathcal{D}_1}(s,\gamma) - \mathrm{POWER}_{\mathcal{D}_2}(s,\gamma) \right| = |1 - 0| = 1 = \mathrm{TV}\left(\mathcal{D}_1, \mathcal{D}_2 \right). \tag{F.275}$$

Theorem F.257 ($\mathrm{POWER}_{\mathcal{D}_{\text{bound}}}$ difference bounded by Wasserstein 1-distance)**.** *Let*

$\mathcal{D}_1, \mathcal{D}_2$ be any bounded reward function distributions.

$$\left|\textsc{Power}_{\mathcal{D}_1}(s,\gamma) - \textsc{Power}_{\mathcal{D}_2}(s,\gamma)\right| \leq W_1\left(\mathcal{D}_1, \mathcal{D}_2\right). \tag{F.276}$$

Proof. Suppose $\gamma \in (0,1)$ and once again let $f(R) \coloneqq \frac{1-\gamma}{\gamma}\left(V^*_R(s,\gamma) - R(s)\right)$. We first show that f has Lipschitz constant 1. Let $R_a, R_b \in \mathbb{R}^{|\mathcal{S}|}$; without loss of generality, suppose $V^*_{R_a}(s,\gamma) \geq V^*_{R_b}(s,\gamma)$.

$$\left|f(R_a) - f(R_b)\right| = \frac{1-\gamma}{\gamma}\left|\max_{\mathbf{f}_a \in \mathcal{F}(s)} \left(\mathbf{f}_a(\gamma) - \mathbf{e}_s\right)^\top \mathbf{r}_a - \max_{\mathbf{f}_b \in \mathcal{F}(s)} \left(\mathbf{f}_b(\gamma) - \mathbf{e}_s\right)^\top \mathbf{r}_b\right| \tag{F.277}$$

$$\leq \frac{1-\gamma}{\gamma}\left|\max_{\mathbf{f}_a \in \mathcal{F}(s)} \left(\mathbf{f}_a(\gamma) - \mathbf{e}_s\right)^\top \mathbf{r}_a - \left(\mathbf{f}_a(\gamma) - \mathbf{e}_s\right)^\top \mathbf{r}_b\right| \tag{F.278}$$

$$= \frac{1-\gamma}{\gamma}\left|\max_{\mathbf{f}_a \in \mathcal{F}(s)} \left(\mathbf{f}_a(\gamma) - \mathbf{e}_s\right)^\top (\mathbf{r}_a - \mathbf{r}_b)\right| \tag{F.279}$$

$$\leq \frac{1-\gamma}{\gamma}\left\|\mathbf{f}_a(\gamma) - \mathbf{e}_s\right\|_1 \left\|\mathbf{r}_a - \mathbf{r}_b\right\|_1 \tag{F.280}$$

$$= \left\|\mathbf{r}_a - \mathbf{r}_b\right\|_1. \tag{F.281}$$

Equation (F.278) follows because $\mathbf{f}_b$ was optimal for $\mathbf{r}_b$, and so $(\mathbf{f}_a(\gamma) - \mathbf{e}_s)^\top \mathbf{r}_b \leq (\mathbf{f}_b(\gamma) - \mathbf{e}_s)^\top \mathbf{r}_b$. Equation (F.280) is a simple application of the Cauchy-Schwarz inequality. Equation (F.281) follows because proposition D.8 shows that $\forall \mathbf{f} \in \mathcal{F}(s) : \left\|\mathbf{f}(\gamma)\right\|_1 = \frac{1}{1-\gamma}$, and $\mathbf{e}_s$ is a unit vector. Therefore, f has Lipschitz constant 1.

We now show the desired inequality.

$$\left|\textsc{Power}_{\mathcal{D}_1}(s,\gamma) - \textsc{Power}_{\mathcal{D}_2}(s,\gamma)\right| \tag{F.282}$$

$$= \left|\int_{[0,1]^{|\mathcal{S}|}} f(R)\left(\mathrm{d}F_1(R) - \mathrm{d}F_2(R)\right)\right| \tag{F.283}$$

$$= \left|\int_{[0,1]^{|\mathcal{S}|}} f(R)\,\mathrm{d}(F_1 - F_2)(R)\right| \tag{F.284}$$

$$\leq \left| \sup_{\substack{f_{\text{lip}}:\mathbb{R}^{|\mathcal{S}|}\to\mathbb{R},\\ f_{\text{lip}}\text{ has minimal Lipschitz constant }\leq 1}} \int_{[0,1]^{|\mathcal{S}|}} f_{\text{lip}}(R)\,\mathrm{d}(F_1-F_2)(R) \right| \tag{F.285}$$

$$= W_1(\mathcal{D}_1, \mathcal{D}_2). \tag{F.286}$$

Equation (F.285) follows because f has Lipschitz constant 1 (and so its minimal constant is at most 1). Equation (F.286) follows by the dual formulation of Wasserstein 1-distance (which is applicable since both distributions have bounded support). □

Theorem F.258 (Optimality probability difference bounded by total variation distance)**.** *Let $\mathcal{D}_1, \mathcal{D}_2$ be any reward function distributions. Let $F \subseteq \mathcal{F}(s)$.*

$$\left|\mathbb{P}_{\mathcal{D}_1}(F,\gamma) - \mathbb{P}_{\mathcal{D}_2}(F,\gamma)\right| \leq \text{TV}(\mathcal{D}_1, \mathcal{D}_2). \tag{F.287}$$

Proof. Let $\mathcal{D}_1$ and $\mathcal{D}_2$ have probability measures F_1 and F_2.

$$\left|\mathbb{P}_{\mathcal{D}_1}(F,\gamma) - \mathbb{P}_{\mathcal{D}_2}(F,\gamma)\right| := \left|F_1\left(\text{supp}\left(F(\gamma) \geq \mathcal{F}(s,\gamma)\right)\right) - F_2\left(\text{supp}\left(F(\gamma) \geq \mathcal{F}(s,\gamma)\right)\right)\right| \tag{F.288}$$

$$\leq \sup_{B \in \mathscr{B}\left(\mathbb{R}^{|\mathcal{S}|}\right)} \left|F_1(B) - F_2(B)\right| \tag{F.289}$$

$$=: \text{TV}(\mathcal{D}_1, \mathcal{D}_2). \tag{F.290}$$

Equation (F.289) follows because $\text{supp}\left(F(\gamma) \geq \mathcal{F}(s,\gamma)\right) = \bigcup_{\mathbf{f}\in F} \text{supp}\left(\mathbf{f}(\gamma) \geq \mathcal{F}(s,\gamma)\right)$ is the finite union of closed sets (lemma F.42), and therefore $\text{supp}\left(F(\gamma) \geq \mathcal{F}(s,\gamma)\right)$ is a Borel set. □

However, fig. F.43 shows that optimality probability cannot be bounded by Wasserstein distance.

Taken together, these results show that strict $\text{POWER}_{\mathcal{D}_{\text{bound}}}$-seeking and optimality probability inequality holds within a TV neighborhood in the space of reward function distributions, and that one action seeks strictly more $\text{POWER}_{\mathcal{D}_{\text{bound}}}$ than another action

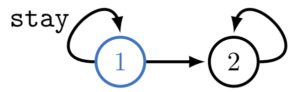


Figure F.43: Let $\epsilon > 0$. If $\mathcal{D}_1$ assigns probability 1 to $s \mapsto \frac{\epsilon}{2}\mathbb{1}_{s=s_1}$ and $\mathcal{D}_2$ assigns probability 1 to $s \mapsto \frac{\epsilon}{2}\mathbb{1}_{s=s_2}$, then $W_1(\mathcal{D}_1, \mathcal{D}_2) = \epsilon$. However, $\mathbb{P}_{\mathcal{D}_1}(s_1, \texttt{stay}, .5) = 1$ while $\mathbb{P}_{\mathcal{D}_2}(s_1, \texttt{stay}, .5) = 0$.

in a 1-Wasserstein neighborhood of reward function distributions. In particular, the strict inequality conditions of proposition 5.25 and theorem 5.29 hold within TV neighborhoods of distributions which are finite mixtures of bounded continuous IID reward function distributions.

Conjecture F.259 (Close $\mathcal{D}_1, \mathcal{D}_2$ induce similar $d^{\text{AU,norm}}_{\mathcal{D}}$ metrics)**.** Let $\mathcal{D}_1, \mathcal{D}_2$ be bounded reward function distributions. $\sup_{\Delta_a, \Delta_b \in \Delta(\mathcal{S})} \left| d^{\text{AU,norm}}_{\mathcal{D}_1}\left(\Delta_a, \Delta_b \mid \gamma\right) - d^{\text{AU,norm}}_{\mathcal{D}_2}\left(\Delta_a, \Delta_b \mid \gamma\right) \right| \leq W_1(\mathcal{D}_1, \mathcal{D}_2)$, where W_1 is the first Wasserstein distance on $\mathbb{R}^{|\mathcal{S}|}$.

F.11.1 Distributional transformations

Positive affine transformation of $\mathcal{D}$ allows generalization of our results to other bounds, as optimal policy is invariant to positive affine transformation of the reward function. Item 3 of proposition F.260 can be viewed as proving Turner et al. [97]'s proposition 4, in the limit of infinitely many draws from the auxiliary reward function distribution $\mathcal{D}'$.

Proposition F.260 (How positive affine transformation affects optimality probability, Power, and normalized AU distance)**.** *Let $\gamma \in [0, 1]$ and let $\mathcal{D}'$ be any bounded reward function distribution. Let $m > 0, b \in \mathbb{R}$ and let $\mathbf{b} \in \mathbb{R}^{|\mathcal{S}|}$ be the ones vector times the scalar b. $m\mathcal{D}' + \mathbf{b}$ is the pushforward distribution over reward functions formed by applying the positive affine transformation $f(\mathbf{r}) := m\mathbf{r} + \mathbf{b}$ to $\mathcal{D}'$.*

1. *Let $F \subseteq \mathcal{F}(s)$. $\mathbb{P}_{m\mathcal{D}'+\mathbf{b}}(F, \gamma) = \mathbb{P}_{\mathcal{D}'}(F, \gamma)$.*
2. *$\text{Power}_{m\mathcal{D}'+\mathbf{b}}(s, \gamma) = m \cdot \text{Power}_{\mathcal{D}'}(s, \gamma) + b$.*
3. *Let $\Delta_1, \Delta_2 \in \Delta(\mathcal{S})$. $d^{\text{AU},norm}_{m\mathcal{D}'+\mathbf{b}}\left(\Delta_1, \Delta_2 \mid \gamma\right) = m \cdot d^{\text{AU},norm}_{\mathcal{D}'}\left(\Delta_1, \Delta_2 \mid \gamma\right)$.*

Proof. Let $\mathcal{D}'$ have probability measure F' and let $m\mathcal{D}' + \mathbf{b}$ have probability measure F'_{aff}.

Item 1: let $\gamma \in (0,1)$. By definition 5.9,

$$\begin{aligned}
\mathbb{P}_{m\mathcal{D}'+\mathbf{b}}\left(F,\gamma\right) &\coloneqq \mathop{\mathbb{P}}_{R\sim m\mathcal{D}'+\mathbf{b}}\left(\exists \mathbf{f}^{\pi} \in F : \pi \in \Pi^{*}\left(R,\gamma\right)\right) &\text{(F.291)}\\
&= \mathop{\mathbb{P}}_{R\sim m\mathcal{D}'+\mathbf{b}}\left(\exists \mathbf{f}^{\pi} \in F : \pi \in \Pi^{*}\left(m^{-1}(R-b),\gamma\right)\right) &\text{(F.292)}\\
&= \mathop{\mathbb{P}}_{R'\sim\mathcal{D}'}\left(\exists \mathbf{f}^{\pi} \in F : \pi \in \Pi^{*}\left(R',\gamma\right)\right) &\text{(F.293)}\\
&=: \mathbb{P}_{\mathcal{D}'}\left(F,\gamma\right). &\text{(F.294)}
\end{aligned}$$

Equation (F.292) holds because optimal policy sets are invariant to positive affine transformation $m^{-1}(R-b)$ of the reward function.

Since this equality holds for all $\gamma \in (0,1)$, it holds in the limits $\gamma \to 0$ and $\gamma \to 1$ as well.

Item 2: let $\gamma \in (0,1)$.

$$\begin{aligned}
&V^{*}_{m\mathcal{D}'+\mathbf{b}}\left(s,\gamma\right) &\text{(F.295)}\\
&= \int_{\mathbb{R}^{|\mathcal{S}|}} V^{*}_{R}\left(s,\gamma\right) \mathrm{d}F'_{\text{aff}}(R) &\text{(F.296)}\\
&= \int_{\mathbb{R}^{|\mathcal{S}|}} V^{*}_{R}\left(s,\gamma\right) \mathrm{d}F'(m^{-1}(R-b)) &\text{(F.297)}\\
&= \int_{\mathbb{R}^{|\mathcal{S}|}} V^{*}_{mR'+b}\left(s,\gamma\right) \mathrm{d}F'(R') &\text{(F.298)}\\
&= \int_{\mathbb{R}^{|\mathcal{S}|}} mV^{*}_{R'}\left(s,\gamma\right) + \frac{b}{1-\gamma} \mathrm{d}F'(R') &\text{(F.299)}\\
&= m \cdot V^{*}_{\mathcal{D}'}\left(s,\gamma\right) + \frac{b}{1-\gamma}. &\text{(F.300)}
\end{aligned}$$

Equation (F.297) follows by the definition of a pushforward measure. Equation (F.298) follows by substituting $R' \coloneqq m^{-1}(R-b)$. Equation (F.299) follows because optimal policies are invariant to positive affine transformations.

$$\begin{aligned}
\text{POWER}_{m\mathcal{D}'+\mathbf{b}}(s,\gamma) &= \frac{1-\gamma}{\gamma}\left(V^*_{m\mathcal{D}'+\mathbf{b}}(s,\gamma) - \mathop{\mathbb{E}}_{R\sim m\mathcal{D}'+\mathbf{b}}\left[R(s)\right]\right) && \text{(F.301)}\\
&= \frac{1-\gamma}{\gamma}\left(mV^*_{\mathcal{D}'}(s,\gamma) + \frac{b}{1-\gamma} - m\mathop{\mathbb{E}}_{R\sim\mathcal{D}'}\left[R(s)\right] - b\right) && \text{(F.302)}\\
&= \frac{1-\gamma}{\gamma}\left(mV^*_{\mathcal{D}'}(s,\gamma) - m\mathop{\mathbb{E}}_{R\sim\mathcal{D}'}\left[R(s)\right] + \frac{b\gamma}{1-\gamma}\right) && \text{(F.303)}\\
&= m\cdot\text{POWER}_{\mathcal{D}'}(s,\gamma) + b. && \text{(F.304)}
\end{aligned}$$

Equation (F.301) and eq. (F.304) follow by lemma D.43 since $\gamma \in (0,1)$. Equation (F.302) follows by eq. (F.300) and the linearity of expectation.

Finally, eq. (F.304) holds in the limits as $\gamma \to 0$ and $\gamma \to 1$.

Item 3: let $\gamma \in [0,1)$.

$$\begin{aligned}
&d^{\text{AU}}_{m\mathcal{D}'+\mathbf{b}}\left(\Delta,\Delta' \mid \gamma\right) && \text{(F.305)}\\
&= \int_{\mathbb{R}^{|\mathcal{S}|}}\left|\mathop{\mathbb{E}}_{s\sim\Delta}\left[V^*_R(s,\gamma)\right] - \mathop{\mathbb{E}}_{s'\sim\Delta'}\left[V^*_R(s',\gamma)\right]\right| \mathrm{d}F'_{\text{aff}}(R) && \text{(F.306)}\\
&= \int_{\mathbb{R}^{|\mathcal{S}|}}\left|\mathop{\mathbb{E}}_{s\sim\Delta}\left[V^*_R(s,\gamma)\right] - \mathop{\mathbb{E}}_{s'\sim\Delta'}\left[V^*_R(s',\gamma)\right]\right| \mathrm{d}F'(m^{-1}(R-b)) && \text{(F.307)}\\
&= \int_{\mathbb{R}^{|\mathcal{S}|}}\left|\mathop{\mathbb{E}}_{s\sim\Delta}\left[V^*_{mR'+b}(s,\gamma)\right] - \mathop{\mathbb{E}}_{s'\sim\Delta'}\left[V^*_{mR'+b}(s',\gamma)\right]\right| \mathrm{d}F'(R') && \text{(F.308)}\\
&= \int_{\mathbb{R}^{|\mathcal{S}|}}\left|\mathop{\mathbb{E}}_{s\sim\Delta}\left[mV^*_{R'}(s,\gamma)+\frac{b}{1-\gamma}\right] - \mathop{\mathbb{E}}_{s'\sim\Delta'}\left[mV^*_{R'}(s',\gamma)+\frac{b}{1-\gamma}\right]\right| \mathrm{d}F'(R') && \text{(F.309)}\\
&= m\cdot d^{\text{AU}}_{\mathcal{D}'}\left(\Delta,\Delta' \mid \gamma\right). && \text{(F.310)}
\end{aligned}$$

Equation (F.307) follows by the definition of a pushforward measure. Equation (F.308) follows via substitution $R' := m^{-1}(R - C)$. Equation (F.309) follows because optimal policy is invariant to positive affine transformation of the reward function.

$$d^{\text{AU,norm}}_{m\mathcal{D}'+\mathbf{b}}\left(\Delta_1,\Delta_2 \mid \gamma\right) := \lim_{\gamma^*\to\gamma}(1-\gamma^*)d^{\text{AU}}_{m\mathcal{D}'+\mathbf{b}}\left(\Delta,\Delta' \mid \gamma\right) \tag{F.311}$$

$$= m \cdot \lim_{\gamma^* \to \gamma} (1 - \gamma^*) d^{\text{AU}}_{\mathcal{D}'} \left(\Delta, \Delta' \mid \gamma\right) \tag{F.312}$$

$$= m \cdot d^{\text{AU,norm}}_{\mathcal{D}'} \left(\Delta_1, \Delta_2 \mid \gamma\right). \tag{F.313}$$

Equation (F.312) holds by eq. (F.310).

Equation (F.313) shows the desired result for all $\gamma \in [0, 1)$. Equation (F.313) holds in the limit $\gamma \to 1$, and so the relationship holds for all $\gamma \in [0, 1]$. □

F.11.1.1 Power inequalities under certain distributional transformations

Proposition F.261 (How non-affine transformations affect Power)**.** *Let $\gamma \in (0, 1)$, let s be a state, and let $\mathcal{D}'$ be any bounded reward function distribution. Suppose $g : \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}^{|\mathcal{S}|}$ is measurable and is such that $\forall \mathbf{r} \in \operatorname{supp}(\mathcal{D}')$:*

$$\forall \mathbf{f} \in \mathcal{F}_{\text{nd}}(s) : \mathbf{f}(\gamma)^\top \mathbf{r} \geq \mathbf{f}(\gamma)^\top g(\mathbf{r}). \tag{F.314}$$

Then $\text{Power}_{\mathcal{D}'}(s, \gamma) \geq \text{Power}_{g(\mathcal{D}')}(s, \gamma)$. This inequality is strict iff eq. (F.314) is strict for a subset of $\operatorname{supp}(\mathcal{D}')$ with positive measure under $\mathcal{D}'$.

A similar statement holds when all of the above inequalities are flipped.

Proof. Let $\mathcal{D}'$ have probability measure F'. Let $g(\mathcal{D}')$ be the pushforward probability distribution induced by applying measurable function g to $\mathcal{D}'$, and let F'_g be its probability measure.

$$V^*_{\mathcal{D}'}(s, \gamma) := \int_{\mathbb{R}^{|\mathcal{S}|}} \max_{\mathbf{f} \in \mathcal{F}(s)} \mathbf{f}(\gamma)^\top \mathbf{r} \, \mathrm{d}F'(\mathbf{r}) \tag{F.315}$$

$$= \int_{\operatorname{supp}(\mathcal{D}')} \max_{\mathbf{f} \in \mathcal{F}_{\text{nd}}(s)} \mathbf{f}(\gamma)^\top \mathbf{r} \, \mathrm{d}F'(\mathbf{r}) \tag{F.316}$$

$$\geq \int_{\operatorname{supp}(\mathcal{D}')} \max_{\mathbf{f} \in \mathcal{F}_{\text{nd}}(s)} \mathbf{f}(\gamma)^\top g(\mathbf{r}) \, \mathrm{d}F'(\mathbf{r}) \tag{F.317}$$

$$= \int_{\operatorname{supp}(g(\mathcal{D}'))} \max_{\mathbf{f} \in \mathcal{F}_{\text{nd}}(s)} \mathbf{f}(\gamma)^\top \mathbf{r}' \, \mathrm{d}F'(g^{-1}(\mathbf{r}')) \tag{F.318}$$

$$= \int_{\text{supp}(g(\mathcal{D}'))} \max_{\mathbf{f} \in \mathcal{F}_{\text{nd}}(s)} \mathbf{f}(\gamma)^\top \mathbf{r}' \, \mathrm{d}F'_g(\mathbf{r}') \tag{F.319}$$

$$= V^*_{g(\mathcal{D}')}(s, \gamma) . \tag{F.320}$$

Equation (F.316) and eq. (F.320) follow by lemma D.39 and the definition of supp($\mathcal{D}'$). Equation (F.317) follows by our assumptions on g. Equation (F.318) holds by the substitution $\mathbf{r}' := g(\mathbf{r})$. Equation (F.319) follows by the definition of a pushforward measure F'_g.

Integration is invariant to strict optimal value decrease on a zero-measure subset of supp($\mathcal{D}'$), but not to strict optimal value decrease on a zero-measure subset of supp($\mathcal{D}'$). Therefore, eq. (F.317) is strict iff eq. (F.314) is strict for a subset of supp($\mathcal{D}'$) with positive measure under $\mathcal{D}'$.

Since $\text{Power}_{\mathcal{D}'}(s, \gamma) = \frac{1-\gamma}{\gamma}\left(V^*_{\mathcal{D}'}(s, \gamma) - \mathbb{E}_{R\sim\mathcal{D}'}\left[R(s)\right]\right)$ by lemma D.43,

$$\text{Power}_{\mathcal{D}'}(s, \gamma) \geq \text{Power}_{g(\mathcal{D}')}(s, \gamma) ,$$

with strict inequality iff eq. (F.314) is strict for a subset of supp($\mathcal{D}'$) with positive measure under $\mathcal{D}'$.

The proof for reward-increasing g follows similarly. □

Proposition F.262 (How non-affine transformations affect optimality probability)**.** *Let $\gamma \in (0, 1)$, let s be a state, let $F \subseteq \mathcal{F}(s)$, and let $\mathcal{D}'$ be any bounded reward function distribution. Suppose $g : \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}^{|\mathcal{S}|}$ is measurable and is such that* $\forall \mathbf{r} \in \text{supp}(\mathcal{D}')$:

$$\forall \mathbf{f} \in F : \mathbf{f}(\gamma)^\top \mathbf{r} \geq \mathbf{f}(\gamma)^\top g(\mathbf{r}), \tag{F.321}$$

$$\forall \mathbf{f}' \in \mathcal{F}_{\text{nd}}(s) \setminus F : \mathbf{f}'(\gamma)^\top \mathbf{r} \leq \mathbf{f}'(\gamma)^\top g(\mathbf{r}). \tag{F.322}$$

Then $\mathbb{P}_{\mathcal{D}'}(F, \gamma) \geq \mathbb{P}_{g(\mathcal{D}')}(F, \gamma)$. Equality holds if eq. (F.321) *and eq.* (F.322) *are equalities for a subset of* supp($\mathcal{D}'$) *with measure* 1 *under* $\mathcal{D}'$*.*

A similar statement holds when all of the above inequalities are flipped.

Proof. Let $\mathcal{D}'$ have probability measure F'. Let $g(\mathcal{D}')$ be the pushforward probability distribution induced by applying measurable function g to $\mathcal{D}'$, and let F'_g be its probability measure.

$$\begin{aligned}
\mathbb{P}_{\mathcal{D}'}\left(F, \gamma\right) &= \int_{\mathbb{R}^{|\mathcal{S}|}} \mathbb{1}_{\max_{\mathbf{f}\in F} \mathbf{f}(\gamma)^\top \mathbf{r} = \max_{\mathbf{f}'\in \mathcal{F}_{\text{nd}}(s)} \mathbf{f}'(\gamma)^\top \mathbf{r}} \, \mathrm{d}F'(\mathbf{r}) && \text{(F.323)}\\
&= \int_{\operatorname{supp}(\mathcal{D}')} \mathbb{1}_{\max_{\mathbf{f}\in F} \mathbf{f}(\gamma)^\top \mathbf{r} = \max_{\mathbf{f}'\in \mathcal{F}_{\text{nd}}(s)\setminus F} \mathbf{f}'(\gamma)^\top \mathbf{r}} \, \mathrm{d}F'(\mathbf{r}) && \text{(F.324)}\\
&\geq \int_{\operatorname{supp}(\mathcal{D}')} \mathbb{1}_{\max_{\mathbf{f}\in F} \mathbf{f}(\gamma)^\top g(\mathbf{r}) = \max_{\mathbf{f}'\in \mathcal{F}_{\text{nd}}(s)\setminus F} \mathbf{f}'(\gamma)^\top g(\mathbf{r})} \, \mathrm{d}F'(\mathbf{r}) && \text{(F.325)}\\
&= \int_{\operatorname{supp}(g(\mathcal{D}'))} \mathbb{1}_{\max_{\mathbf{f}\in F} \mathbf{f}(\gamma)^\top \mathbf{r}' = \max_{\mathbf{f}'\in \mathcal{F}_{\text{nd}}(s)\setminus F} \mathbf{f}'(\gamma)^\top \mathbf{r}'} \, \mathrm{d}F'(g^{-1}(\mathbf{r}')) && \text{(F.326)}\\
&= \int_{\operatorname{supp}(g(\mathcal{D}'))} \mathbb{1}_{\max_{\mathbf{f}\in F} \mathbf{f}(\gamma)^\top \mathbf{r}' = \max_{\mathbf{f}'\in \mathcal{F}_{\text{nd}}(s)\setminus F} \mathbf{f}'(\gamma)^\top \mathbf{r}'} \, \mathrm{d}F'_g(\mathbf{r}') && \text{(F.327)}\\
&= \int_{\operatorname{supp}(g(\mathcal{D}'))} \mathbb{1}_{\max_{\mathbf{f}\in F} \mathbf{f}(\gamma)^\top \mathbf{r}' = \max_{\mathbf{f}'\in \mathcal{F}_{\text{nd}}(s)} \mathbf{f}'(\gamma)^\top \mathbf{r}'} \, \mathrm{d}F'_g(\mathbf{r}') && \text{(F.328)}\\
&= \mathbb{P}_{g(\mathcal{D}')}\left(F, \gamma\right). && \text{(F.329)}
\end{aligned}$$

Equation (F.323) and eq. (F.329) follow by lemma D.42. Equation (F.324) and eq. (F.328) follow by lemma F.62 (1) and the definition of $\operatorname{supp}(\mathcal{D}')$. Equation (F.325) follows because our assumptions on g imply that

$$\forall \mathbf{r} \in \operatorname{supp}(\mathcal{D}') : \mathbb{1}_{\max_{\mathbf{f}\in F} \mathbf{f}(\gamma)^\top \mathbf{r} = \max_{\mathbf{f}'\in \mathcal{F}_{\text{nd}}(s)\setminus F} \mathbf{f}'(\gamma)^\top \mathbf{r}} \geq \mathbb{1}_{\max_{\mathbf{f}\in F} \mathbf{f}(\gamma)^\top g(\mathbf{r}) = \max_{\mathbf{f}'\in \mathcal{F}_{\text{nd}}(s)\setminus F} \mathbf{f}'(\gamma)^\top g(\mathbf{r})}.$$

Equation (F.326) follows by the substitution $\mathbf{r}' := g(\mathbf{r})$. Equation (F.327) follows by the definition of the pushforward measure F'_g.

Integration is invariant to strict optimal value decrease on a zero-measure subset of $\operatorname{supp}(\mathcal{D}')$. Therefore, eq. (F.317) is an equality if eq. (F.321) and eq. (F.322) are equalities for a subset of $\operatorname{supp}(\mathcal{D}')$ with measure 1 under $\mathcal{D}'$.

A similar proof follows when the theorem statement's inequalities are flipped. $\square$

Remark. Unlike proposition F.261, proposition F.262 does not have a strict inequality if-and-only-if. This is because even if g strictly increases return for $\mathbf{f} \in F$, that increase may be insufficient to change the optimality status of $\mathbf{f} \in F$ for any $\mathbf{r} \in \operatorname{supp}(\mathcal{D}')$.

Proposition F.261 and proposition F.262 significantly expand the initial $\textsc{Power}_{\mathcal{D}_{\text{bound}}}$-seeking results to $g(\mathcal{D})$ which distribute reward independently and non-identically across states.

Reconsider proposition 5.25. For *e.g.* the max-entropy $\mathcal{D}$ over $[0,1]^{|\mathcal{S}|}$, at s', action a is strictly $\textsc{Power}_{\mathcal{D}_{\text{bound}}}$-seeking and more probable under optimality compared to a'. Proposition F.261 and proposition F.262 show that by *e.g.* doubling reward at green states and zeroing the reward at two of the red states (call this distribution $g(\mathcal{D})$), the same $\textsc{Power}_{g(\mathcal{D})}$-seeking and optimality probability statements hold.

F.11.2 $\mathcal{F}_{\text{nd}}$ symmetry

Definition F.263 ($\mathcal{F}_{\text{nd}}(s)$ symmetry group)**.** Let S_n be the permutation group on n elements. For any state s,

$$S_{\mathcal{F}_{\text{nd}}(s)} \coloneqq \left\{\phi \in S_{|\mathcal{S}|} \mid \phi \cdot \mathcal{F}_{\text{nd}}(s) = \mathcal{F}_{\text{nd}}(s)\right\}. \tag{F.330}$$

Proposition F.264 ($S_{\mathcal{F}_{\text{nd}}(s)}$ is a subgroup of $S_{|\mathcal{S}|}$)**.**

Proof. Let s be any state. By definition F.263, $S_{\mathcal{F}_{\text{nd}}(s)} \subseteq S_{|\mathcal{S}|}$. We show that $S_{\mathcal{F}_{\text{nd}}(s)}$ satisfies the group axioms under permutation composition.

Identity. Let $\phi_{id} \in S_{|\mathcal{S}|}$ be the identity permutation. Then clearly $\phi_{id}\left(\mathcal{F}_{\text{nd}}(s)\right) = \mathcal{F}_{\text{nd}}(s)$, so $\phi_{id} \in S_{\mathcal{F}_{\text{nd}}(s)}$.

Composition. Let $\phi_1, \phi_2 \in S_{|\mathcal{S}|}$. Then

$$\begin{aligned}
(\phi_1 \circ \phi_2)\left(\mathcal{F}_{\text{nd}}(s)\right) &= \phi_1\left(\phi_2\left(\mathcal{F}_{\text{nd}}(s)\right)\right) && \text{(F.331)}\\
&= \phi_1\left(\phi_2\left(\mathcal{F}_{\text{nd}}(s)\right)\right) && \text{(F.332)}\\
&= \phi_1\left(\mathcal{F}_{\text{nd}}(s)\right) && \text{(F.333)}\\
&= \mathcal{F}_{\text{nd}}(s), && \text{(F.334)}
\end{aligned}$$

and so $(\phi_1 \circ \phi_2) \in S_{\mathcal{F}_{\text{nd}}(s)}$.

Inverse. Let $\phi \in S_{\mathcal{F}_{\text{nd}}(s)}$. $\phi \cdot \mathcal{F}_{\text{nd}}(s) = \mathcal{F}_{\text{nd}}(s)$ implies that $\mathcal{F}_{\text{nd}}(s) = \phi^{-1}\left(\mathcal{F}_{\text{nd}}(s)\right)$, and so $\phi^{-1} \in S_{\mathcal{F}_{\text{nd}}(s)}$.

Therefore, $S_{\mathcal{F}_{\text{nd}}(s)}$ is a group. □

Lemma F.265 (If $\phi \cdot \mathcal{F}_{\text{nd}}(s) = \mathcal{F}_{\text{nd}}(s')$, then $\phi(s) = s'$)**.**

Proof. $\phi \cdot \mathcal{F}_{\text{nd}}(s) = \mathcal{F}_{\text{nd}}(s')$ implies that

$$\left\{\lim_{\gamma\to 0} \mathbf{P}_\phi \mathbf{f}(\gamma) \mid \mathbf{f} \in \mathcal{F}_{\text{nd}}(s)\right\} = \left\{\lim_{\gamma\to 0} \mathbf{f}(\gamma) \mid \mathbf{f} \in \mathcal{F}_{\text{nd}}(s')\right\}. \tag{F.335}$$

But the right-hand side equals $\{\mathbf{e}_{s'}\}$ by the definition of a state visit distribution function $\mathbf{f} \in \mathcal{F}(s')$ (definition 5.3). The left-hand side also equals $\{\mathbf{e}_{s'}\}$, implying that $\mathbf{P}_\phi \mathbf{e}_s = \mathbf{e}_{s'}$ and so $\phi(s) = s'$. □

Proposition F.266 ($\text{Power}_{\mathcal{D}_{\text{bound}}}$ across certain distributional symmetries)**.** *Let s, s' be states. If $\exists \phi \in S_{|\mathcal{S}|}$ such that $\phi \cdot \mathcal{F}_{\text{nd}}(s) = \mathcal{F}_{\text{nd}}(s')$, then* $\text{Power}_{\mathcal{D}_{bound}}(s,\gamma) = \text{Power}_{\phi\cdot\mathcal{D}_{bound}}\left(s',\gamma\right)$.

Proof. Let $\gamma^* \in (0,1), F \coloneqq \left\{\mathbf{f}(\gamma^*) - \mathbf{e}_s \mid \mathbf{f} \in \mathcal{F}_{\text{nd}}(s)\right\}, F' \coloneqq \left\{\mathbf{f}(\gamma^*) - \mathbf{e}_{s'} \mid \mathbf{f} \in \mathcal{F}_{\text{nd}}(s')\right\}$.

$$\text{Power}_{\mathcal{D}_{\text{bound}}}(s,\gamma) \tag{F.336}$$

$$\coloneqq \lim_{\gamma^*\to\gamma} \frac{1-\gamma^*}{\gamma^*} \mathop{\mathbb{E}}_{\mathbf{r}\sim\mathcal{D}_{\text{bound}}}\left[\max_{\mathbf{f}\in\mathcal{F}(s)} \left(\mathbf{f}(\gamma^*) - \mathbf{e}_s\right)^\top \mathbf{r}\right] \tag{F.337}$$

$$= \lim_{\gamma^*\to\gamma} \frac{1-\gamma^*}{\gamma^*} \mathop{\mathbb{E}}_{\mathbf{r}\sim\mathcal{D}_{\text{bound}}}\left[\max_{\mathbf{f}\in\mathcal{F}_{\text{nd}}(s)} \left(\mathbf{f}(\gamma^*) - \mathbf{e}_s\right)^\top \mathbf{r}\right] \tag{F.338}$$

$$= \lim_{\gamma^*\to\gamma} \frac{1-\gamma^*}{\gamma^*} \mathop{\mathbb{E}}_{\mathbf{r}\sim\phi\cdot\mathcal{D}_{\text{bound}}}\left[\max_{\mathbf{f}'\in\phi\cdot F} \mathbf{f}'^\top \mathbf{r}\right] \tag{F.339}$$

$$= \lim_{\gamma^*\to\gamma} \frac{1-\gamma^*}{\gamma^*} \mathop{\mathbb{E}}_{\mathbf{r}\sim\phi\cdot\mathcal{D}_{\text{bound}}}\left[\max_{\mathbf{f}'\in F'} \mathbf{f}'^\top \mathbf{r}\right] \tag{F.340}$$

$$\eqqcolon \text{Power}_{\phi\cdot\mathcal{D}_{\text{bound}}}\left(s',\gamma\right). \tag{F.341}$$

Equation (F.338) follows because lemma D.39 shows that optimal value is invariant to restriction to $\mathcal{F}_{\text{nd}}$. Let $g : \mathbb{R} \to \mathbb{R}$ be the identity function, and let $f(B \mid \mathcal{D}_{\text{bound}}) :=$ $\mathbb{E}_{\mathbf{r}\sim\mathcal{D}_{\text{bound}}}\left[g\left(\max_{\mathbf{f}\in B} \mathbf{f}^\top \mathbf{r}\right)\right]$. Equation (F.339) follows by applying lemma D.22 to conclude that $f(F \mid \mathcal{D}_{\text{bound}}) = f(\phi \cdot F \mid \phi \cdot \mathcal{D}_{\text{bound}})$.

$\phi \cdot F = \left\{\mathbf{P}_\phi \mathbf{f}(\gamma^*) - \mathbf{P}_\phi \mathbf{e}_s \mid \mathbf{f} \in \mathcal{F}_{\text{nd}}(s)\right\} = F'$ by the assumptions on ϕ and the fact that $\phi(s) = s'$ by lemma F.265. Therefore, eq. (F.340) follows.

Since this holds for all $\gamma \in (0,1)$, it holds in the limits $\gamma \to 0$ and $\gamma \to 1$ as well. □

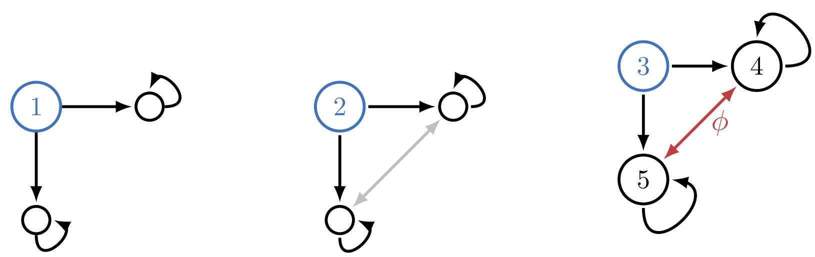


Figure F.44: By proposition F.266, $\forall \gamma \in [0,1] : \textsc{Power}_{\mathcal{D}_{X\text{-IID}}}(s_1, \gamma) = \textsc{Power}_{\mathcal{D}_{X\text{-IID}}}(s_2, \gamma)$. In general, $\textsc{Power}_{\mathcal{D}_{\text{bound}}}(s_3, \gamma) = \textsc{Power}_{\phi \cdot \mathcal{D}_{\text{bound}}}(s_3, \gamma)$.

F.11.3 State similarity

Definition F.267 (State similarity)**.** State s is *similar* to s' if there exists a stochastic model isomorphism ϕ such that $\phi(s) = s'$. If all states are similar, the model is *vertex transitive.*

When the dynamics are deterministic, definition F.267 reduces to the standard graph-theoretic vertex similarity, as shown in fig. F.45.

Corollary F.268 (State similarity criterion)**.** *s and s' are similar via permutation ϕ iff $\mathcal{F}(s') = \phi \cdot \mathcal{F}(s)$.*

Proof. Apply theorem F.113. □

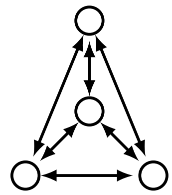


Figure F.45: The tetrahedral graph is vertex transitive.

Lemma F.269 (Similar states have similar non-dominated visit distribution functions)**.** *If s and s' are similar via permutation ϕ, then $\mathcal{F}_{\text{nd}}(s') = \phi \cdot \mathcal{F}_{\text{nd}}(s)$.*

Proof. State similarity implies visit distribution function similarity (corollary F.268), so $\mathcal{F}(s') = \left\{\mathbf{P}_\phi \mathbf{f} \mid \mathbf{f} \in \mathcal{F}(s)\right\}$. Without loss of generality, suppose that $\mathbf{P}_\phi \mathbf{f}$ is non-dominated at s'; this implies $\mathbf{P}_\phi \mathbf{f}$ is strictly optimal for reward function R at discount rate γ. Then

$$\left(\mathbf{P}_\phi \mathbf{f}(\gamma)\right)^\top \mathbf{r} > \max_{\mathbf{f}' \in \mathcal{F}(s') \setminus \left\{\mathbf{P}_\phi \mathbf{f}\right\}} \mathbf{f}'(\gamma)^\top \mathbf{r}. \tag{F.342}$$

$$\mathbf{f}(\gamma)^\top (\mathbf{P}_{\phi^{-1}} \mathbf{r}) > \max_{\mathbf{f}' \in \mathcal{F}(s) \setminus \{\mathbf{f}\}} \mathbf{f}'(\gamma)^\top (\mathbf{P}_{\phi^{-1}} \mathbf{r}). \tag{F.343}$$

Then $\mathbf{f}$ is strictly optimal at state s for reward function $\mathbf{P}_{\phi^{-1}} \mathbf{r}$ at discount rate γ, and thus $\mathbf{f} \in \mathcal{F}_{\text{nd}}(s)$. □

Proposition F.270 (Similar states have equal $\textsc{Power}_{\mathcal{D}_{\text{bound}}}$)**.** *If s and s' are similar,* $\forall \gamma \in [0,1] : \textsc{Power}_{\mathcal{D}_{X\text{-IID}}}(s, \gamma) = \textsc{Power}_{\mathcal{D}_{X\text{-IID}}}\left(s', \gamma\right)$.

Proof. Suppose $\phi \cdot \mathcal{F}(s) = \mathcal{F}(s')$. Apply lemma F.269 to conclude that $\phi \cdot \mathcal{F}_{\text{nd}}(s) = \mathcal{F}_{\text{nd}}(s')$. Then apply proposition F.266 to conclude the desired equality (with identical distribution ensuring that $\phi \cdot \mathcal{D}_{X\text{-IID}} = \mathcal{D}_{X\text{-IID}}$). □

Corollary F.271 (Vertex transitivity implies $\textsc{Power}_{\mathcal{D}_{X\text{-IID}}}$ is equal everywhere)**.** *If the model is vertex transitive, then all states have equal* $\textsc{Power}_{\mathcal{D}_{X\text{-IID}}}$.

Figure F.44 shows that non-dominated similarity (proposition F.266) allows us to conclude $\textsc{Power}_{\mathcal{D}_{\text{bound}}}$ equality in a greater range of situations than does proposition F.270.

F.11.4 Strong visitation distribution set similarity

Definition F.272 (Strong visitation distribution set similarity)**.** Let $F, F' \in \mathcal{F}(s)$. We say that F and F' are *strongly similar* if they are similar via a permutation $\phi \in S_{\mathcal{F}_{\text{nd}}(s)}$.

Strong visitation distribution similarity depends on the totality of $\mathcal{F}_{\text{nd}}(s)$. We aren't just interested in whether $F' = \phi \cdot F$ — we want to know whether F and F' "interact with $\mathcal{F}_{\text{nd}}(s)$ in the same way."

Proposition F.273 (Similar visit distribution functions have the same optimality probability)**.** *If $F, F' \subseteq \mathcal{F}(s)$ are strongly similar via permutation ϕ, then $\mathbb{P}_{\mathcal{D}_{any}}(F, \gamma) = \mathbb{P}_{\phi \cdot \mathcal{D}_{any}}\left(F', \gamma\right)$.*

Proof. Let $\gamma \in (0, 1)$.

$$
\begin{aligned}
\mathbb{P}_{\mathcal{D}}\left(F, \gamma\right) &= \mathop{\mathbb{E}}_{\mathbf{r} \sim \mathcal{D}_{\text{any}}}\left[\mathbb{1}_{\max_{\mathbf{f} \in F} \mathbf{f}(\gamma)^\top \mathbf{r} \geq \max_{\mathbf{f}^i \in \mathcal{F}_{\text{nd}}(s)} \mathbf{f}^i(\gamma)^\top \mathbf{r}}\right] && \text{(F.344)}\\
&= \mathop{\mathbb{E}}_{\mathbf{r} \sim \phi \cdot \mathcal{D}_{\text{any}}}\left[\mathbb{1}_{\max_{\mathbf{f}' \in F'} \mathbf{f}'(\gamma)^\top \mathbf{r} \geq \max_{\mathbf{f}^i \in \mathcal{F}_{\text{nd}}(s)} \mathbf{f}^i(\gamma)^\top \mathbf{r}}\right] && \text{(F.345)}\\
&= \mathbb{P}_{\phi \cdot \mathcal{D}_{\text{any}}}\left(F', \gamma\right). && \text{(F.346)}
\end{aligned}
$$

Equation (F.344) follows by lemma D.42. Let $g(b_1, b_2) := \mathbb{1}_{b_1 \geq b_2}$ and $f(B_1, B_2 \mid \mathcal{D}_{\text{any}}) := \mathbb{E}_{\mathbf{r} \sim \mathcal{D}_{\text{any}}}\left[g\left(\max_{\mathbf{b}_1 \in B_1} \mathbf{b}_1^\top \mathbf{r}, \max_{\mathbf{b}_2 \in B_2} \mathbf{b}_2^\top \mathbf{r}\right)\right]$. Then by lemma D.22 and the definition of strong similarity,

$$
\begin{aligned}
f\left(F(\gamma), \mathcal{F}_{\text{nd}}(s, \gamma) \mid \mathcal{D}_{\text{any}}\right) &= f\left(\phi \cdot F(\gamma), \phi \cdot \mathcal{F}_{\text{nd}}(s, \gamma) \mid \phi \cdot \mathcal{D}_{\text{any}}\right)\\
&= f\left(F'(\gamma), \mathcal{F}_{\text{nd}}(s, \gamma) \mid \phi \cdot \mathcal{D}_{\text{any}}\right).
\end{aligned}
$$

This implies eq. (F.345).

The $\gamma = 0$ and $\gamma = 1$ cases hold via the appropriate limits. $\square$

The existence of such a stochastic model isomorphism ϕ on the full model is sufficient (but not necessary) for strong similarity (see fig. F.46).

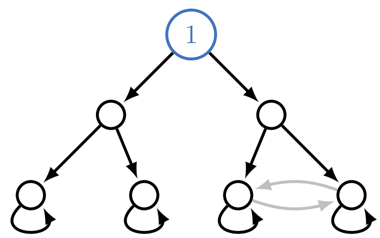


Figure F.46: Gray actions are only taken by the policies of dominated visit distribution functions. Considered as singleton sets, all four non-dominated visitation distribution functions are strongly similar. Conjecture F.274 allows us to conclude they all have equal optimality probability, even though this is not obvious just from looking at the full model. Indeed, the left-most and right-most trajectories would not be classified as "similar" under a definition where ϕ has to be a stochastic model isomorphism.

Figure F.47 shows that vertex transitivity does not imply that all visitation distributions *at a state* are strongly similar to each other. By theorem F.113, vertex transitivity shows that all states have similar visitation distribution function sets, but not that all visit distribution functions within each set are strongly similar to each other.

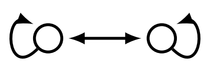


Figure F.47: The complete directed graph on two vertices is vertex transitive, but each self-loop is not strongly similar to each state's other visit distribution functions.

Conjecture F.274 (Strongly similar non-dominated visit distributions and determinism imply no instrumental convergence)**.** If the environment is deterministic and all non-dominated visitation distributions of a state are strongly similar, then no instrumental convergence exists downstream of that state.

F.12 Orbits

Lemma F.275 (Trivial satisfaction of $\geq^n_{\text{most}}$)**.** *Let $f : \Theta \to \mathbb{R}$, and suppose Θ is a subset of a set acted on by S_d. $\forall n \geq 0 : f(\theta) \geq^n_{\text{most: }\Theta} f(\theta)$.*

Proof. Because $f(\theta) > f(\theta)$ is impossible, both cardinalities in definition E.10 are zero, and so the claim holds trivially. $\square$

Conjecture F.276 (Lower-bound on joint $\geq_{\text{most: }\mathfrak{D}_{\text{any}}}$ agreement strength)**.** Suppose $\left(f_i \geq^{C_i}_{\text{most: }\mathfrak{D}_{\text{any}}} f'_i\right)_{i\in I}$ for countable index set I. Then for all $\mathcal{D}_{\text{bound}}$, the proportion of the $\mathcal{D}^{\phi}_{\text{bound}} \in \left(S_{|\mathcal{S}|} \cdot \mathcal{D}_{\text{bound}}\right)$ which satisfy $\bigwedge_{i\in I} f_i\left(\mathcal{D}^{\phi}_{\text{bound}}\right) \geq f'_i\left(\mathcal{D}^{\phi}_{\text{bound}}\right)$ is greater than $1 - \sum_{i\in I}(1 - \frac{C_i}{C_i+1})$.

Proposition F.277 (Orbit tendencies lower-bound measure under $\mathcal{D}_{X\text{-IID}}$)**.** *Suppose $f, g : \mathbb{R}^d \to \mathbb{R}$ are measurable, let n be a positive integer, and suppose $f(\mathbb{R}^d) \geq^{n}_{\text{most: }\mathfrak{D}_{any}} g(\mathbb{R}^d)$. For $\mathbf{u} \in \mathbb{R}^d$, let $O_{f<g}(\mathbf{u}) \coloneqq \left\{\mathbf{u}' \in S_d \cdot \mathbf{u} \mid f(\mathbf{u}') < g(\mathbf{u}')\right\}$. Suppose there exist $\phi_1, \ldots, \phi_n \in S_d$ such that for all $\mathbf{u} \in \mathbb{R}^d : \phi_i \cdot O_{f<g}(\mathbf{u}) \subseteq (S_d \cdot \mathbf{u}) \setminus O_{f<g}(\mathbf{u})$, and when $i \neq j$, $\phi_i \cdot O_{f<g}(\mathbf{u})$ and $\phi_j \cdot O_{f<g}(\mathbf{u})$ are disjoint.*

Then for any IID *distribution $\mathcal{D} \in \Delta(\mathbb{R}^d)$, $\mathbb{P}_{\mathbf{u}\sim\mathcal{D}}\left(f(\mathbf{u}) \geq g(\mathbf{u})\right) \geq \frac{n}{n+1}$.*

Proof. f and g are measurable, and thus $f - g$ is measurable, and the set $(-\infty, 0)$ is measurable in $\mathbb{R}$, and so $(f - g)^{-1}\left((-\infty, 0)\right)$ is measurable. But this preimage equals $X \coloneqq \left\{\mathbf{u} \in \mathbb{R}^d \mid f(\mathbf{u}) < g(\mathbf{u})\right\}$, and so X is measurable.

Let $\mathcal{D}$ have probability measure F.

$$\begin{aligned}
\mathbb{P}_{\mathbf{u}\sim\mathcal{D}}\left(f(\mathbf{u}) \geq g(\mathbf{u})\right) &\coloneqq F\left(\left\{\mathbf{u} \in \mathbb{R}^d \mid f(\mathbf{u}) \geq g(\mathbf{u})\right\}\right) && \text{(F.347)}\\
&\geq F\left(\cup_{i=1}^{n} \phi_i \cdot X\right) && \text{(F.348)}\\
&= \sum_{i=1}^{n} F(\phi_i \cdot X) && \text{(F.349)}\\
&= \sum_{i=1}^{n} F(X) && \text{(F.350)}\\
&= nF(X). && \text{(F.351)}
\end{aligned}$$

Equation (F.347) is well-defined because the right-hand set is measurable for the same reason that X is measurable. Equation (F.348) holds because assumption on ϕ_i ensures that $\phi_i \cdot X \subseteq \left\{\mathbf{u} \in \mathbb{R}^d \mid f(\mathbf{u}) \geq g(\mathbf{u})\right\}$.

If $\phi_i \neq \phi_j$, then suppose $\mathbf{u}, \mathbf{u}' \in \mathbb{R}^d$. If they are not in the same orbit under S_d, then $\phi_i \cdot \mathbf{u} \neq \phi_j \cdot \mathbf{u}'$ by the definition of an orbit. Otherwise, the inequality holds by the assumed disjointness. Therefore the elements of $\left\{\phi_i \cdot X \mid i = 1, \ldots, n\right\}$ are pairwise disjoint, and so eq. (F.349) holds by the finite additivity of probability measures on disjoint sets. Equation (F.350) follows because probability measure F is IID across states and so is invariant to permutation of X, and because permutations are measurable transformations with unitary Jacobian determinant.

Since $X \cup \left\{\mathbf{u} \in \mathbb{R}^d \mid f(\mathbf{u}) \geq g(\mathbf{u})\right\} = \mathbb{R}^d$ and the two sets are disjoint, their probability sums to 1 and so $\mathbb{P}_{\mathbf{u} \sim \mathcal{D}}\left(f(\mathbf{u}) \geq g(\mathbf{u})\right) \geq \frac{n}{n+1}$. □

Conjecture F.278 (Generalized measure lower-bound for orbit tendencies)**.** Suppose $f, g : \mathbb{R}^d \to \mathbb{R}$ are measurable, let $C \geq 0$, and suppose $f(\mathbb{R}^d) \geq^C_{\text{most: } \mathfrak{D}_{\text{any}}} g(\mathbb{R}^d)$. Then for all $\mathcal{D}_{X\text{-IID}}$, $\mathbb{P}_{\mathbf{u} \sim \mathcal{D}_{X\text{-IID}}}\left(f(\mathbf{u}) \geq g(\mathbf{u})\right) \geq \frac{C}{C+1}$.

F.12.1 Blackwell versus average optimality

Turner et al. [99]'s theorem 5.29 applies to *average-optimal* policies ($\gamma = 1$), but not to Blackwell-optimal policies. Therefore, their results are inapplicable to discount rates $\gamma \approx 1$. Proposition F.281 shows corollary F.283, which shows that theorem 5.29 holds when $\gamma \approx 1$ for almost all reward function orbits.

Definition F.279 (Visit distribution functions which induce RSDs)**.** Let $D \subseteq \text{RSD}(s)$. The set of visit distribution functions which induce $\mathbf{d} \in D$ is:

$$\mathcal{F}(s \mid \text{RSD} \in D) := \left\{\mathbf{f} \in \mathcal{F}(s) \mid \text{NORM}(\mathbf{f}, 1) \in D\right\}. \tag{F.352}$$

$\mathbb{P}_{\mathcal{D}_{\text{any}}}(D, 1) := \mathbb{P}_{\mathcal{D}_{\text{any}}}\left(\mathcal{F}(s \mid \text{RSD} \in D), 1\right)$.

Lemma F.280 (Average optimality probability is greater than Blackwell optimality probability)**.** *Let* $D \subseteq \text{RSD}(s)$. $\mathbb{P}_{\mathcal{D}_{any}}\left(\mathcal{F}(s \mid \text{RSD} \in D), 1\right) \leq \mathbb{P}_{\mathcal{D}_{any}}(D, \text{average})$.

Proof. If some $\mathbf{f} \in \mathcal{F}(s \mid \text{RSD} \in D)$ is Blackwell optimal for reward function $R \in \mathbb{R}^{\mathcal{S}}$, then $\text{NORM}(\mathbf{f}, 1)$ is average optimal for R – *i.e.* Blackwell optimality at a state implies average optimality at a state [68] . Therefore, the set of reward functions for which some $\mathbf{f} \in \mathcal{F}(s \mid \text{RSD} \in D)$ is Blackwell optimal is a subset of the set of reward functions for which some $\mathbf{d} \in D$ is average optimal. Then the desired inequality follows by the monoticity of probability. $\square$

Proposition F.281 (Average optimality probability equals Blackwell optimality probability for almost all reward functions)**.** *Let $D \subseteq \text{RSD}(s)$. For almost all reward functions $\mathbf{r} \in \mathbb{R}^{|\mathcal{S}|}$, $\mathbb{P}_{\mathcal{D}_{\mathbf{r}}}(D, 1) = \mathbb{P}_{\mathcal{D}_{\mathbf{r}}}(D, \text{average})$ (where $\mathcal{D}_{\mathbf{r}}$ is the degenerate probability distribution which places probability 1 on $\mathbf{r}$).*

Proof. Lemma F.280 shows that $\mathbb{P}_{\mathcal{D}_{\mathbf{r}}}(D, 1) \leq \mathbb{P}_{\mathcal{D}_{\mathbf{r}}}(D, \text{average})$. So inequality only holds if $\mathbb{P}_{\mathcal{D}_{\mathbf{r}}}(D, 1) < \mathbb{P}_{\mathcal{D}_{\mathbf{r}}}(D, \text{average})$. Since $\mathcal{D}_{\mathbf{r}}$ is a degenerate distribution, this implies that $\mathbb{P}_{\mathcal{D}_{\mathbf{r}}}(D, 1) = 0 < 1 = \mathbb{P}_{\mathcal{D}_{\mathbf{r}}}(D, \text{average})$. D is non-empty because $1 = \mathbb{P}_{\mathcal{D}_{\mathbf{r}}}(D, \text{average})$; in particular, there must exist $\mathbf{d}_A \in D$ which is average optimal for $\mathbf{r}$ but which is not induced by a Blackwell optimal policy for $\mathbf{r}$.

Let $D_B \coloneqq \left\{\mathbf{d}^{\pi,s} \mid \pi \in \Pi^*(\mathbf{r}, 1)\right\}$ be the set of RSDs induced by Blackwell optimal policies for $\mathbf{r}$. Since every reward function has a Blackwell optimal policy [11], D_B must be non-empty. Let $\mathbf{d}_B \in D_B$ be one of its elements.

We know that $\mathbf{d}_A \notin D_B$, and so $\mathbf{d}_A \neq \mathbf{d}_B$. Since Blackwell optimal policies must be average optimal [68], $\mathbf{d}_A$ and $\mathbf{d}_B$ are both average optimal. By corollary D.13,

$$\left\{\mathbf{r} \in \mathbb{R}^{|\mathcal{S}|} \mid \left|\operatorname*{arg\,max}_{\mathbf{d} \in \text{RSD}(s)} \mathbf{d}^\top \mathbf{r}\right| > 1\right\}$$

has measure zero under all absolutely continuous measures on $\mathbb{R}^{|\mathcal{S}|}$. In particular, this set has zero Lebesgue measure. $\square$

Corollary F.282 (Average optimality probability equals Blackwell optimality probability for $\mathcal{D}_{\text{cont}}$)**.** *Let $D \subseteq \text{RSD}(s)$. $\mathbb{P}_{\mathcal{D}_{cont}}(D, 1) = \mathbb{P}_{\mathcal{D}_{cont}}(D, \text{average})$.*

Proof.

$$\mathbb{P}_{\mathcal{D}_{\text{cont}}}(D, 1) \coloneqq \mathop{\mathbb{P}}_{R \sim \mathcal{D}_{\text{cont}}}\left(\exists \mathbf{f}^\pi \in \mathcal{F}(s \mid \text{RSD} \in D) : \pi \in \Pi^*(R, 1)\right) \tag{F.353}$$

$$= \mathop{\mathbb{E}}_{\mathbf{r}\sim\mathcal{D}_{\text{cont}}} \left[\mathbb{P}_{\mathcal{D}_{\mathbf{r}}}(D,1)\right] \tag{F.354}$$

$$= \mathop{\mathbb{E}}_{\mathbf{r}\sim\mathcal{D}_{\text{cont}}} \left[\mathbb{P}_{\mathcal{D}_{\mathbf{r}}}(D,1,\text{average})\right] \tag{F.355}$$

$$= \mathbb{P}_{\mathcal{D}_{\text{cont}}}(D,\text{average}). \tag{F.356}$$

By proposition F.281, almost all $\mathbf{r} \in \mathbb{R}^{|\mathcal{S}|}$ agree that $\mathbb{P}_{\mathcal{D}_{\mathbf{r}}}(D,1) = \mathbb{P}_{\mathcal{D}_{\mathbf{r}}}(D,\text{average})$. Since $\mathcal{D}_{\text{cont}}$ is absolutely continuous with respect to the Lebesgue measure, it also assigns zero probability measure to disagreeing reward functions. Then eq. (F.355) follows. □

Corollary F.283 (Average optimality probability equals Blackwell optimality probability for all orbit elements in almost all orbits)**.** *Let* $D \subseteq \text{RSD}(s)$*. For almost all* $\mathbf{r} \in \mathbb{R}^{|\mathcal{S}|}$*, all* $\mathcal{D}_{\mathbf{r}'} \in S_{|\mathcal{S}|} \cdot \mathcal{D}_{\mathbf{r}}$ *satisfy* $\mathbb{P}_{\mathcal{D}_{\mathbf{r}'}}(D,1) = \mathbb{P}_{\mathcal{D}_{\mathbf{r}'}}(D,\text{average})$*.*

Proof. By proposition F.281, the set $X := \left\{\mathbf{r} \in \mathbb{R}^{|\mathcal{S}|} \mid \mathbb{P}_{\mathcal{D}_{\mathbf{r}}}(D,1) \neq \mathbb{P}_{\mathcal{D}_{\mathbf{r}}}(D,\text{average})\right\}$ has Lebesgue measure zero. Let $O := \left\{\mathbf{r}' \in \mathbb{R}^{|\mathcal{S}|} \mid \exists \mathbf{r} \in S_{|\mathcal{S}|} \cdot \mathbf{r}' : \mathbf{r} \in X\right\}$ be the set of reward functions whose orbits contain an element of X. Alternatively, $O = \left\{\cup_{\phi \in S_{|\mathcal{S}|}} \mathbf{P}_\phi \mathbf{r} \mid \mathbf{r} \in X\right\}$.

Let μ be the Lebesgue measure; we want to show that $\mu(O) = 0$.

$$\mu(O) := \mu\left(\cup_{\phi \in S_{|\mathcal{S}|}} \left\{\mathbf{P}_\phi \mathbf{r} \mid \mathbf{r} \in X\right\}\right) \tag{F.357}$$

$$= \mu\left(\cup_{\phi \in S_{|\mathcal{S}|}} \phi \cdot X\right) \tag{F.358}$$

$$\leq \sum_{\phi \in S_{|\mathcal{S}|}} \mu(\phi \cdot X) \tag{F.359}$$

$$= \sum_{\phi \in S_{|\mathcal{S}|}} \mu(X) \tag{F.360}$$

$$= |\mathcal{S}|! \cdot 0 \tag{F.361}$$

$$= 0. \tag{F.362}$$

Equation (F.359) follows by the union bound. Equation (F.360) follows because permutations have Jacobian determinant 1 and therefore are measure-preserving operators. Equation (F.361) follows because $\left|S_{|\mathcal{S}|}\right| = |\mathcal{S}|!$ and $\mu(X) = 0$.

Therefore, $\mu(O) = 0$, and so for almost all $\mathbf{r} \in \mathbb{R}^{|\mathcal{S}|}$, all $\mathcal{D}_{\mathbf{r}'} \in S_{|\mathcal{S}|} \cdot \mathcal{D}_{\mathbf{r}}$ satisfy $\mathbb{P}_{\mathcal{D}_{\mathbf{r}'}}(D,1) =$

$\mathbb{P}_{\mathcal{D}_{\mathbf{r}'}}(D, \text{average})$. $\square$

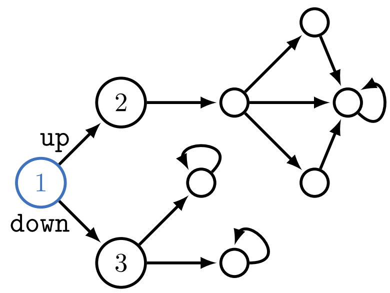


Figure F.48: More states are reachable by going `up` compared to `down`. Thus, most orbit elements of $S_{|\mathcal{S}|} \cdot \mathbf{e}_s$ make it strictly Blackwell-optimal to go `up`, while also most elements make it (at least weakly) average-optimal to go `down`. Therefore, the orbit $S_{|\mathcal{S}|} \cdot \mathbf{e}_s$ does not satisfy corollary F.283.

Conjecture F.284 (Appendix F.12.1's results hold for child-state distributions $T(s)$ instead of $\operatorname{RSD}(s)$)**.**

Conjecture F.284 will probably be aided by definition F.285, which takes the place of definition F.279.

Definition F.285 (Visit dist. functions which induce child visit distributions)**.** Let $N \subseteq T(s)$.

$$\mathcal{F}(s \mid \text{CHILD DISTRIBUTION} \in N) \coloneqq \left\{ \mathbf{f} \in \mathcal{F}(s) \mid \lim_{\gamma \to 0} \gamma^{-1} \left(\mathbf{f}(\gamma) - \mathbf{e}_s \right) \in N \right\}. \quad \text{(F.363)}$$

$\mathbb{P}_{\mathcal{D}}(N, 0) \coloneqq \mathbb{P}_{\mathcal{D}}\left(\mathcal{F}(s \mid \text{CHILD DISTRIBUTION} \in N), 0\right)$.

F.12.2 Power

Conjecture F.286 (Orbit incentives characterization for Power)**.** Let $\gamma \in (0,1)$. The following statements are equivalent:

1. $\text{POWER}_{\mathcal{D}_{\text{bound}}} \geq_{\text{most: } \mathfrak{D}_{\text{bound}}} \text{POWER}_{\mathcal{D}_{\text{bound}}}\left(s', \gamma\right)$.
2. For all continuous bounded IID $\mathcal{D}_{X\text{-IID}}$, $\text{POWER}_{\mathcal{D}_{X\text{-IID}}}(s, \gamma) \geq \text{POWER}_{\mathcal{D}_{X\text{-IID}}}\left(s', \gamma\right)$.

3. $\mathcal{F}_{\text{nd}}(s')$ is similar to a subset of $\mathcal{F}(s)$.

Conjecture F.287 ($\geq_{\text{most: } \mathfrak{D}_{\text{bound}}}$ is not a complete ordering for the POWER of states)**.**

Lastly, a remark motivating the requirement of *involution* in the copy containment definition (definition 5.17).

Remark. Lemma D.29 and lemma D.33 require similarity via involution (not just via permutation). Consider $A \coloneqq \left\{ \begin{pmatrix} 0 \\ 5 \\ 10 \end{pmatrix} \right\}, B \coloneqq \left\{ \begin{pmatrix} 10 \\ 0 \\ 5 \end{pmatrix} \right\}, \phi \coloneqq (1\,2\,3)$. A is similar to B via permutation ϕ. However, there is no involution which enforces the similarity.

Furthermore, consider the orbit $S_3 \cdot \mathcal{D}_{\mathbf{e}_3} = \left\{ \mathcal{D}_{\mathbf{e}_1}, \mathcal{D}_{\mathbf{e}_2}, \mathcal{D}_{\mathbf{e}_3} \right\}$ and consider the conditions of lemma D.29 with g the identity function.

$$\mathop{\mathbb{E}}_{\mathbf{r} \sim \mathcal{D}_{\mathbf{e}_1}} \left[\max_{\mathbf{b} \in B} \mathbf{b}^\top \mathbf{r} \right] = 10 > 0 = \mathop{\mathbb{E}}_{\mathbf{r} \sim \mathcal{D}_{\mathbf{e}_1}} \left[\max_{\mathbf{a} \in A} \mathbf{b}^\top \mathbf{r} \right],$$

but the opposite is true for the other two orbit elements. Therefore,

$$\mathop{\mathbb{E}}_{\mathbf{r} \sim \mathcal{D}_{\text{bound}}} \left[\max_{\mathbf{a} \in A} \mathbf{b}^\top \mathbf{r} \right] \not\leq_{\text{most: } \mathfrak{D}_{\text{bound}}} \mathop{\mathbb{E}}_{\mathbf{r} \sim \mathcal{D}_{\text{bound}}} \left[\max_{\mathbf{b} \in B} \mathbf{b}^\top \mathbf{r} \right].$$

Proposition F.288 (Nontrivial copy containment guarantee)**.** *If B contains $n > 1$ copies of A via $\phi_1, ..., \phi_n$, then for any $i = 1, ..., n$, $A \neq B_i$.*

Proof. Suppose $A = B_i \coloneqq \phi_i \cdot A$ for some ϕ_i. Then consider any $j \neq i$ (which exists since $n > 1$). $\phi_j \cdot B_i = \phi_j \cdot A = B_j$. But by the set copy definition (definition E.7), $\phi_j \cdot B_i = B_i$ since $j \neq i$. Then $B_i = B_j$, which contradicts the set copy assumption. $\square$

F.13 Featurized utility functions

Consider the case where utility functions over outcomes are *featurized*: $u(o) = \text{feat}(o)^\top \alpha$, with $\text{feat}(o) : \{1, \ldots, d\} \to \mathbb{R}^{n_f}$ linearly mapping deterministic outcomes to n_f-dimensional feature vectors. Let $\mathbf{F} \coloneqq \begin{pmatrix} \text{feat}(o_1) & \cdots & \text{feat}(o_d) \end{pmatrix}$ be the feature matrix. For outcome lottery set $X \subseteq \mathbb{R}^d$, $\mathbf{F} \cdot X \coloneqq \left\{\mathbf{F}\mathbf{x} \mid \mathbf{x} \in X\right\}$ is the left-coset.

For feature weighting $\alpha \in \mathbb{R}^{n_f}$, the expected utility of an outcome lottery $\mathbf{x} \in \mathbb{R}^d$ is then the expected value of the α-weighted combination of each outcome's features: $\mathbf{x}^\top \mathbf{u} = \mathbf{x}^\top(\mathbf{F}^\top \alpha) = (\mathbf{F}\mathbf{x})^\top \alpha$. For example, if an agent maximizes expected utility, it essentially makes decisions according to the available feature vectors of outcomes, and so instead of looking for copy containment in the outcome lottery sets $A, B \subseteq \mathbb{R}^d$, we can look for copy containment in the feature vector sets $\mathbf{F} \cdot A, \mathbf{F} \cdot B \subseteq \mathbb{R}^{n_f}$.

Proposition F.289 (Feature-level tendencies guaranteed by featurizations which commute with outcome symmetries)**.** *Let $A, B \subsetneq \mathbb{R}^d$, and suppose that B contains a copy of A via ϕ with permutation matrix $\mathbf{P}_\phi \in \mathbb{R}^{d \times d}$. Consider featurization $\mathbf{F} \in \mathbb{R}^{n_f \times d}$ which maps each of d outcomes to their n_f-feature vectors.*

If there exists a feature involution $\phi_f \in S_{n_f}$ such that $\mathbf{P}_{\phi_f}\mathbf{F} = \mathbf{F}\mathbf{P}_\phi$, then $\mathbf{F} \cdot B$ contains a copy of $\mathbf{F} \cdot A$ via ϕ_f.

Proof.

$$\phi_f \cdot (\mathbf{F} \cdot A) \coloneqq \left\{\mathbf{P}_{\phi_f}\mathbf{F}\mathbf{a} \mid \mathbf{a} \in A\right\} \tag{F.364}$$
$$= \left\{\mathbf{F}\mathbf{P}_\phi \mathbf{a} \mid \mathbf{a} \in A\right\} \tag{F.365}$$
$$\subseteq \left\{\mathbf{F}\mathbf{b} \mid \mathbf{b} \in B\right\} \tag{F.366}$$
$$\eqqcolon \mathbf{F} \cdot B. \tag{F.367}$$

Equation (F.365) follows by the assumption that $\mathbf{P}_{\phi_f}\mathbf{F} = \mathbf{F}\mathbf{P}_\phi$. Equation (F.366) follows because we assumed that $\phi \cdot A \subseteq B$.

Since $\phi_f \cdot (\mathbf{F} \cdot A) \subseteq \mathbf{F} \cdot B$ and ϕ_f is an involution by assumption, then we conclude that $\mathbf{F} \cdot B$ contains a copy of $\mathbf{F} \cdot A$ via ϕ_f. ☐

Conjecture F.290 (Multiple feature copy containment ensured by feature commutation)**.** Proposition F.289 generalizes to the case with multiple feature involutions $\phi_{f_1}, \ldots, \phi_{f_n}$.

F.14 ϵ-Optimal policies

Turner et al. [99]'s results assume perfectly optimal agents. We now relax that assumption slightly to ϵ-optimality, although ϵ may be extremely small (and so these results do not seem very practically interesting). For a more practical extension, see chapter 6's total abandonment of the optimality requirement.

Traditionally, ϵ-optimal policies are defined as π such that for all $s \in \mathcal{S}$:

$$V_R^*(s,\gamma) - V_R^\pi(s,\gamma) \leq \epsilon \tag{F.368}$$

for fixed $\gamma \in (0,1)$. However, in the state-based reward setting, this is equivalent to

$$\left(R(s) + \gamma \max_{\pi'} \mathop{\mathbb{E}}_{s' \sim T(s,\pi'(s))}\left[V_R^{\pi'}(s',\gamma)\right]\right) - \left(R(s) + \gamma \mathop{\mathbb{E}}_{s' \sim T(s,\pi(s))}\left[V_R^{\pi}(s',\gamma)\right]\right) \leq \epsilon \tag{F.369}$$

$$\max_{\pi'} \mathop{\mathbb{E}}_{s' \sim T(s,\pi'(s))}\left[V_R^{\pi'}(s',\gamma)\right] - \mathop{\mathbb{E}}_{s' \sim T(s,\pi(s))}\left[V_R^{\pi}(s',\gamma)\right] \leq \gamma^{-1}\epsilon. \tag{F.370}$$

To avoid superfluous division by γ, we modify the traditional criterion to

$$\max_{\pi'} \mathop{\mathbb{E}}_{s' \sim T(s,\pi'(s))}\left[V_R^{\pi'}(s',\gamma)\right] - \mathop{\mathbb{E}}_{s' \sim T(s,\pi(s))}\left[V_R^{\pi}(s',\gamma)\right] \leq \epsilon. \tag{F.371}$$

However, since the magnitude of on-policy value often diverges as $\gamma \to 1$, we instead consider a *discount-averaged per-time step suboptimality* of ϵ:

$$\max_{\pi'} \mathop{\mathbb{E}}_{s' \sim T(s,\pi'(s))}\left[V_R^{\pi'}(s',\gamma)\right] - \mathop{\mathbb{E}}_{s' \sim T(s,\pi(s))}\left[V_R^{\pi}(s',\gamma)\right] \leq \frac{\epsilon}{1-\gamma}. \tag{F.372}$$

Multiplying both sides by $1-\gamma$, we arrive at our definition.

Definition F.291 (ϵ-optimal policy)**.** Let $\epsilon \geq 0$. Policy π is *ϵ-optimal for reward function R at discount rate* $\gamma \in [0,1]$ when for all $s \in \mathcal{S}$,

$$\max_{\pi'} \mathop{\mathbb{E}}_{s' \sim T(s,\pi'(s))}\left[V_{R,\text{norm}}^{\pi'}(s',\gamma)\right] - \mathop{\mathbb{E}}_{s' \sim T(s,\pi(s))}\left[V_{R,\text{norm}}^{\pi}(s',\gamma)\right] \leq \epsilon. \tag{F.373}$$

For any fixed $\gamma \in (0,1)$, the traditional definition (eq. (F.368)) is equivalent to definition F.291, in that traditional ϵ-optimality is equivalent to definition F.291's $\frac{\epsilon}{\gamma(1-\gamma)}$-optimality.

Definition F.292 (ϵ-optimal policy set)**.** Let $\epsilon \geq 0$. The *ϵ-optimal policy set* for reward function R at discount rate $\gamma \in [0,1]$ is

$$\Pi^{\epsilon}(R,\gamma) := \left\{\pi \in \Pi \mid \pi \text{ is } \epsilon\text{-optimal for } R \text{ at } \gamma\right\}. \tag{F.374}$$

Remark. When $\gamma \in (0,1)$, many reward functions may have multiple ϵ-optimal visit distributions, whereas lemma F.105 shows that continuous distributions place zero probability on reward functions with multiple optimal visit distributions.

F.14.1 ϵ-optimal $\textsc{Power}_{\mathcal{D}_{\text{bound}}}$

The situation for ϵ-optimal $\textsc{Power}_{\mathcal{D}_{\text{bound}}}$ is quite simple.

Definition F.293 (ϵ-optimal policy-generating function)**.** $\text{pol} : \mathbb{R}^{\mathcal{S}} \times [0,1] \to \Pi$ is an *ϵ-optimal policy-generating function* when $\forall R \in \mathbb{R}^{\mathcal{S}}, \gamma \in [0,1] : \text{pol}(R,\gamma) \in \Pi^{\epsilon}(R,\gamma)$.

Proposition F.294 (ϵ-optimal $\textsc{Power}_{\mathcal{D}_{\text{bound}}}$ bound)**.** *Let $\epsilon \geq 0$ and $\gamma \in [0,1]$. Let pol be an ϵ-optimal policy-generating function. Then for all states s,*

$$\textsc{Power}_{\mathcal{D}_{bound}}(s,\gamma) - \textsc{Power}^{pol}_{\mathcal{D}_{bound}}(s,\gamma) \leq \epsilon. \tag{F.375}$$

Proof.

$$\begin{aligned}
&\textsc{Power}_{\mathcal{D}_{\text{bound}}}(s,\gamma) - \textsc{Power}^{\text{pol}}_{\mathcal{D}_{\text{bound}}}(s,\gamma) && \text{(F.376)}\\
&= \mathop{\mathbb{E}}_{R\sim\mathcal{D}_{\text{any}}}\left[\max_{\pi'} \mathop{\mathbb{E}}_{s'\sim T(s,\pi'(s))}\left[V^{\pi'}_{R,\text{norm}}(s',\gamma)\right] - \mathop{\mathbb{E}}_{s'\sim T(s,\pi(s))}\left[V^{\pi}_{R,\text{norm}}(s',\gamma)\right]\right] && \text{(F.377)}\\
&\leq \mathop{\mathbb{E}}_{R\sim\mathcal{D}_{\text{any}}}[\epsilon] && \text{(F.378)}\\
&= \epsilon. && \text{(F.379)}
\end{aligned}$$

Equation (F.377) follows from lemma F.11 and definition D.5 of $\text{POWER}^{\text{pol}}_{\mathcal{D}_{\text{bound}}}$. Equation (F.378) follows from its assumed ϵ-optimality for all reward functions. ☐

Theorem F.295 (Optimal POWER-seeking implies ϵ-optimal POWER-seeking)**.** *Let $k \geq 1$, $s \in \mathcal{S}, a,\ a' \in \mathcal{A}, \gamma \in [0,1]$. If*

$$\mathop{\mathbb{E}}_{s' \sim T(s,a)}\left[\text{POWER}_{\mathcal{D}_{bound}}\left(s', \gamma\right)\right] > k \mathop{\mathbb{E}}_{s' \sim T(s,a')}\left[\text{POWER}_{\mathcal{D}_{bound}}\left(s', \gamma\right)\right],$$

then there exists $\epsilon > 0$ such that

$$\mathop{\mathbb{E}}_{s' \sim T(s,a)}\left[\text{POWER}^{pol}_{\mathcal{D}_{bound}}\left(s', \gamma\right)\right] > k \mathop{\mathbb{E}}_{s' \sim T(s,a')}\left[\text{POWER}^{pol}_{\mathcal{D}_{bound}}\left(s', \gamma\right)\right],$$

where pol is any ϵ-optimal policy-generating function.

Proof. Let $\delta := \mathbb{E}_{s' \sim T(s,a)}\left[\text{POWER}_{\mathcal{D}_{\text{bound}}}\left(s', \gamma\right)\right] - k\,\mathbb{E}_{s' \sim T(s,a')}\left[\text{POWER}_{\mathcal{D}_{\text{bound}}}\left(s', \gamma\right)\right]$.

$$\mathop{\mathbb{E}}_{\substack{s_a \sim T(s,a),\\ s_{a'} \sim T(s,a')}}\left[\text{POWER}^{\text{pol}}_{\mathcal{D}_{\text{bound}}}\left(s_a, \gamma\right) - k\text{POWER}^{\text{pol}}_{\mathcal{D}_{\text{bound}}}\left(s_{a'}, \gamma\right)\right] \tag{F.380}$$

$$\begin{aligned} = \mathbb{E}_{\substack{s_a \sim T(s,a),\\ s_{a'} \sim T(s,a')}}\Bigg[&\Big(\text{POWER}^{\text{pol}}_{\mathcal{D}_{\text{bound}}}\left(s_a, \gamma\right) - \text{POWER}_{\mathcal{D}_{\text{bound}}}\left(s_a, \gamma\right)\Big) \\ &+ \left(\text{POWER}_{\mathcal{D}_{\text{bound}}}\left(s_a, \gamma\right) - k\text{POWER}_{\mathcal{D}_{\text{bound}}}\left(s_{a'}, \gamma\right)\right) \\ &+ \Big(\text{POWER}_{\mathcal{D}_{\text{bound}}}\left(s_{a'}, \gamma\right) - \text{POWER}^{\text{pol}}_{\mathcal{D}_{\text{bound}}}\left(s_{a'}, \gamma\right)\Big)\Bigg] \end{aligned} \tag{F.381}$$

$$\begin{aligned} = \mathbb{E}_{\substack{s_a \sim T(s,a),\\ s_{a'} \sim T(s,a')}}\Bigg[&\left(\text{POWER}_{\mathcal{D}_{\text{bound}}}\left(s_a, \gamma\right) - k\text{POWER}_{\mathcal{D}_{\text{bound}}}\left(s_{a'}, \gamma\right)\right) \\ &- \Big(\text{POWER}_{\mathcal{D}_{\text{bound}}}\left(s_a, \gamma\right) - \text{POWER}^{\text{pol}}_{\mathcal{D}_{\text{bound}}}\left(s_a, \gamma\right)\Big) \\ &+ \Big(\text{POWER}_{\mathcal{D}_{\text{bound}}}\left(s_{a'}, \gamma\right) - \text{POWER}^{\text{pol}}_{\mathcal{D}_{\text{bound}}}\left(s_{a'}, \gamma\right)\Big)\Bigg] \end{aligned} \tag{F.382}$$

$$\geq \mathop{\mathbb{E}}_{\substack{s_a \sim T(s,a),\\ s_{a'} \sim T(s,a')}}\left[\left(\text{POWER}_{\mathcal{D}_{\text{bound}}}\left(s_a, \gamma\right) - k\text{POWER}_{\mathcal{D}_{\text{bound}}}\left(s_{a'}, \gamma\right)\right) - \epsilon + 0\right] \tag{F.383}$$

$$= \delta - \epsilon \tag{F.384}$$

$$> 0. \tag{F.385}$$

Equation (F.384) follows because $\textsc{Power}_{\mathcal{D}_{\text{bound}}}(s_{a'}, \gamma) \geq \textsc{Power}^{\text{pol}}_{\mathcal{D}_{\text{bound}}}(s_{a'}, \gamma)$ and because proposition F.294 bounds $\textsc{Power}_{\mathcal{D}_{\text{bound}}}(s_a, \gamma) - \textsc{Power}^{\text{pol}}_{\mathcal{D}_{\text{bound}}}(s_a, \gamma) \leq \epsilon$ because pol is an ϵ-optimal policy generating function. Equation (F.385) follows for any $\epsilon < \delta$; since δ is positive, we can ensure that ϵ is as well. □

F.14.2 ϵ-optimality probability

The situation for ϵ-optimality probability is less simple.

Definition F.296 (ϵ-optimality probability)**.** For state s, let $F \subseteq \mathcal{F}(s)$, $\gamma \in [0,1]$, and $\epsilon \geq 0$.

$$\mathbb{P}^{\epsilon}_{\mathcal{D}_{\text{any}}}(F, \gamma) \coloneqq \mathop{\mathbb{P}}_{R \sim \mathcal{D}_{\text{any}}}\left(\exists \mathbf{f}^{\pi} \in F : \pi \in \Pi^{\epsilon}(R, \gamma)\right). \tag{F.386}$$

Definition F.297 (Average-optimal policies)**.** $\Pi^{\text{avg}}(R)$ is the set of average reward-optimal policies for reward function R.

Proposition F.298 (Characterizing 0-optimal policy sets)**.** *Let* $\gamma \in (0,1)$.

1. $\Pi^*(R, \gamma) = \Pi^0(R, \gamma)$.
2. $\Pi^*(R, 0) \subseteq \Pi^0(R, 0) = \Pi^{greedy}(R)$.
3. $\Pi^*(R, 1) \subseteq \Pi^0(R, 1) = \Pi^{avg}(R)$.

Proof. If $\gamma \in (0,1)$, $\Pi^0(R, \gamma)$ must be optimal at every s by definition F.291. $\Pi^0(R, 0) = \Pi^{\text{greedy}}(R)$ by definition F.14 and $\Pi^0(R, 1) = \Pi^{\text{avg}}(R)$ by definition F.297. $\Pi^*(R, 0) \subseteq \Pi^{\text{greedy}}(R)$ by lemma F.15 and $\Pi^*(R, 1) \subseteq \Pi^{\text{avg}}(R)$ by the fact that Blackwell optimal policies must be average optimal ([68]). □

Corollary F.299 (When $\gamma \in (0,1)$, 0-optimality probability coincides with optimality probability)**.** *Let* $F \subseteq \mathcal{F}(s)$. *If* $\gamma \in (0,1)$, *then* $\mathbb{P}_{\mathcal{D}_{any}}(F, \gamma) = \mathbb{P}^0_{\mathcal{D}_{any}}(F, \gamma)$.

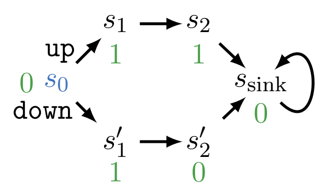


Figure F.49: All policies are average-optimal, since there is only one RSD. However, only going up is Blackwell-optimal. Similarly, all policies are greedily optimal, since $R(s_1) = R(s_1') = 1$, but only going up is asymptotically greedily optimal. In this MDP, $\Pi^0(R,0) \neq \Pi^*(R,0)$ and $\Pi^0(R,1) \neq \Pi^*(R,1)$ by proposition F.298.

Proof. When $\gamma \in (0,1)$, $\Pi^*(R,\gamma) = \Pi^0(R,\gamma)$ by proposition F.298. ϵ-optimality probability (definition F.296) then reduces to optimality probability (definition 5.9). □

Lemma F.300 (ϵ-optimal policy set monotonicity)**.** *Let* $0 \leq \epsilon_1 \leq \epsilon_2$. *Then* $\Pi^{\epsilon_1}(R,\gamma) \subseteq \Pi^{\epsilon_2}(R,\gamma)$.

Proof. Suppose $\pi \in \Pi^{\epsilon_1}(R,\gamma)$. Then by definition F.292,

$$\max_{\pi'} \mathop{\mathbb{E}}_{s' \sim T(s,\pi'(s))}\left[V^{\pi'}_{R,\,\text{norm}}\left(s',\gamma\right)\right] - \mathop{\mathbb{E}}_{s' \sim T(s,\pi(s))}\left[V^{\pi}_{R,\,\text{norm}}\left(s',\gamma\right)\right] \leq \epsilon_1 \tag{F.387}$$

$$\leq \epsilon_2. \tag{F.388}$$

Therefore, $\pi \in \Pi^{\epsilon_2}(R,\gamma)$. □

Corollary F.301 (ϵ-optimal policy set containment)**.** *Let* $\epsilon \geq 0$. *For all* $R \in \mathbb{R}^{\mathcal{S}}, \gamma \in [0,1]$, $\Pi^*(R,\gamma) \subseteq \Pi^0(R,\gamma) \subseteq \Pi^\epsilon(R,\gamma)$.

Proof. The first containment holds by proposition F.298. The second containment holds by lemma F.300. □

Proposition F.302 (ϵ-optimality probability is monotonically increasing in ϵ)**.** *Let s be a state and consider* $F \subseteq \mathcal{F}(s)$, $\gamma \in [0,1]$. *Let* $0 \leq \epsilon_1 \leq \epsilon_2$. $\mathbb{P}_{\mathcal{D}}(F,\gamma) \leq \mathbb{P}^{\epsilon_1}_{\mathcal{D}_{any}}(F,\gamma) \leq \mathbb{P}^{\epsilon_2}_{\mathcal{D}_{any}}(F,\gamma)$.

Proof. By lemma F.300 and corollary F.301, $\Pi^*(R,\gamma) \subseteq \Pi^{\epsilon_1}(R,\gamma) \subseteq \Pi^{\epsilon_2}(R,\gamma)$. Then $\mathbb{P}_{\mathcal{D}}(F,\gamma) \leq \mathbb{P}^{\epsilon_1}_{\mathcal{D}_{\text{any}}}(F,\gamma) \leq \mathbb{P}^{\epsilon_2}_{\mathcal{D}_{\text{any}}}(F,\gamma)$ by the monotonicity of probability. □

As ϵ increases, more policies are ϵ-optimal for any given reward function (lemma F.300). The following result shows that for bounded reward function distributions, "anything goes" for sufficiently large ϵ.

Proposition F.303 (Under $\mathcal{D}_{\text{bound}}$, every policy can be ϵ-optimal (for the right ϵ))**.** *Let $\mathcal{D}_{bound}$ be a reward function distribution which is bounded $[b,c]$. Then for all $R \in \operatorname{supp}(\mathcal{D}_{bound})$ and $\gamma \in [0,1]$, $\Pi^{(c-b)}(R,\gamma) = \Pi$ and for all s, non-empty $F \subseteq \mathcal{F}(s)$, $\mathbb{P}^{(c-b)}_{\mathcal{D}_{bound}}(F,\gamma) = 1$.*

Proof. Let $\gamma \in [0,1]$.

$$\forall R \in \operatorname{supp}(\mathcal{D}_{\text{bound}}), s \in \mathcal{S} : V^*_{R,\,\text{norm}}(s,\gamma) \leq \lim_{\gamma^* \to \gamma} \frac{c}{1-\gamma^*} = c. \tag{F.389}$$

Its minimal normalized value is likewise at least b. Therefore, for all $\pi \in \Pi$, we have

$$\max_{\pi'} \mathop{\mathbb{E}}_{s' \sim T(s,\pi'(s))}\left[V^{\pi'}_{R,\,\text{norm}}(s',\gamma)\right] - \mathop{\mathbb{E}}_{s' \sim T(s,\pi(s))}\left[V^{\pi}_{R,\,\text{norm}}(s',\gamma)\right] \tag{F.390}$$

$$\leq \max_{\pi'} \mathop{\mathbb{E}}_{s' \sim T(s,\pi'(s))}[c] - \mathop{\mathbb{E}}_{s' \sim T(s,\pi(s))}[b] \tag{F.391}$$

$$= c - b. \tag{F.392}$$

This holds at all states s. So every π is $(c-b)$-optimal, and $\Pi^{(c-b)}(R,\gamma) = \Pi$.

Since F is not empty, it must contain some $\mathbf{f}^\pi \in F$. Since $\Pi^{(c-b)}(R,\gamma) = \Pi$, $\pi \in \Pi^{(c-b)}(R,\gamma)$ for all R, and so $\mathbb{P}^{(c-b)}_{\mathcal{D}_{\text{bound}}}(F,\gamma) = 1$ by definition F.296. □

Suppose that we know that action a is both strictly POWER-seeking and strictly more probable under optimality compared to another action a'. Proposition F.294 implies that there exists an $\epsilon > 0$ for which a is ϵ-optimal POWER-seeking compared to a'. The following result implies a similar result with respect to ϵ-optimality probability.

Proposition F.304 (ϵ-optimality probability approaches 0-optimality probability in a continuous fashion)**.** *Let* $F \subseteq \mathcal{F}(s)$*.* $\lim_{\epsilon \to 0} \mathbb{P}^{\epsilon}_{\mathcal{D}_{any}}(F, \gamma) = \mathbb{P}^{0}_{\mathcal{D}_{any}}(F, \gamma)$*.*

Proof. Let $f(\epsilon) := \mathbb{P}^{\epsilon}_{\mathcal{D}_{\text{any}}}(F, \gamma)$. f is monotonically increasing in ϵ and has range $[\mathbb{P}^{0}_{\mathcal{D}_{\text{any}}}(F, \gamma), 1]$ by proposition F.302. Suppose the sequence $\mathbf{s} := (\epsilon_i)_{i \geq 1}$ converges to 0. Choose a monotonically decreasing subsequence $\mathbf{s}' := \left(\epsilon'_j\right)_{j \in I}$. $\mathbf{s}'$ must also converge to 0. Then $\mathbf{s}'_f := \left(f(\epsilon'_j)\right)_{j \in I}$ monotonically decreases and is bounded below by $\mathbb{P}^{0}_{\mathcal{D}_{\text{any}}}(F, \gamma)$. Therefore, by the monotone convergence theorem, $\mathbf{s}'_f$ has non-negative limit L.

Suppose $L > \mathbb{P}^{0}_{\mathcal{D}_{\text{any}}}(F, \gamma)$. This implies that all elements of $\mathbf{s}'$ are positive. Then there exist reward functions for which some $\mathbf{f}^{\pi} \in F$ is ϵ-optimal for arbitrarily small $\epsilon > 0$, but for which $\mathbf{f}$ is not 0-optimal. This is impossible, and so $L = \mathbb{P}^{0}_{\mathcal{D}_{\text{any}}}(F, \gamma)$.

Furthermore, since $\mathbf{s} \to 0$, $\left(f(\epsilon_i)\right)_{i \geq 1} \to \mathbb{P}^{0}_{\mathcal{D}_{\text{any}}}(F, \gamma)$ as well. Since this limit applies for any sequence $\mathbf{s} \to 0$, the result follows. ☐

Corollary F.305 (For small ϵ, ϵ-optimality probability approximates optimality probability)**.** *Let* $F \subseteq \mathcal{F}(s)$*. When* $\gamma \in (0, 1)$*, for any* $\delta > 0$*, there exists* $\epsilon > 0$ *such that* $\mathbb{P}^{\epsilon}_{\mathcal{D}_{any}}(F, \gamma) - \mathbb{P}_{\mathcal{D}_{any}}(F, \gamma) < \delta$*.*

Proof. By corollary F.299, when $\gamma \in (0, 1)$, $\mathbb{P}^{0}_{\mathcal{D}_{\text{any}}}(F, \gamma) = \mathbb{P}_{\mathcal{D}_{\text{any}}}(F, \gamma)$. Therefore, when $\gamma \in (0, 1)$, proposition F.304 shows that $\lim_{\epsilon \to 0} \mathbb{P}^{\epsilon}_{\mathcal{D}_{\text{any}}}(F, \gamma) = \mathbb{P}_{\mathcal{D}_{\text{any}}}(F, \gamma)$. Proposition F.304 showed that $f(\epsilon) := \mathbb{P}^{\epsilon}_{\mathcal{D}_{\text{any}}}(F, \gamma)$ is sequentially continuous at $\epsilon = 0$; since $f : \mathbb{R} \to \mathbb{R}$ has metric spaces for both its domain and its range, sequential continuity implies (topological) continuity at $\epsilon = 0$. Then the claim follows from the definition of continuity on metric spaces. ☐

Lemma F.306 (Strict optimality probability inequalities are preserved for small enough ϵ)**.** *Let* $k \geq 1, s, s' \in \mathcal{S}, F \subseteq \mathcal{F}(s), F' \subseteq \mathcal{F}(s')$*, and* $\gamma \in (0, 1)$*. If* $\mathbb{P}_{\mathcal{D}_{any}}(F, \gamma) > k\,\mathbb{P}_{\mathcal{D}_{any}}\left(F', \gamma\right)$*, then there exists* $\epsilon > 0$ *such that* $\mathbb{P}^{\epsilon}_{\mathcal{D}_{any}}(F, \gamma) > k\,\mathbb{P}^{\epsilon}_{\mathcal{D}_{any}}\left(F', \gamma\right)$*.*

Proof. Let $\delta := \mathbb{P}_{\mathcal{D}_{\text{any}}}(F, \gamma) - k\,\mathbb{P}_{\mathcal{D}_{\text{any}}}\left(F', \gamma\right)$; note that $\delta > 0$ by assumption.

$$\mathbb{P}^{\epsilon}_{\mathcal{D}_{\text{any}}}(F, \gamma) - k\,\mathbb{P}^{\epsilon}_{\mathcal{D}_{\text{any}}}\left(F', \gamma\right) \tag{F.393}$$

$$
\begin{aligned}
&= \Big(\mathbb{P}^{\epsilon}_{\mathcal{D}_{\text{any}}}(F,\gamma) - \mathbb{P}_{\mathcal{D}_{\text{any}}}(F,\gamma)\Big) \\
&\qquad + \Big(\mathbb{P}_{\mathcal{D}_{\text{any}}}(F,\gamma) - k\,\mathbb{P}_{\mathcal{D}_{\text{any}}}(F',\gamma)\Big) \\
&\qquad + k\Big(\mathbb{P}_{\mathcal{D}_{\text{any}}}(F',\gamma) - \mathbb{P}^{\epsilon}_{\mathcal{D}_{\text{any}}}(F',\gamma)\Big) && \text{(F.394)} \\
&= \delta + \Big(\mathbb{P}^{\epsilon}_{\mathcal{D}_{\text{any}}}(F,\gamma) - \mathbb{P}_{\mathcal{D}_{\text{any}}}(F,\gamma)\Big) - k\Big(\mathbb{P}^{\epsilon}_{\mathcal{D}_{\text{any}}}(F',\gamma) - \mathbb{P}_{\mathcal{D}_{\text{any}}}(F',\gamma)\Big) && \text{(F.395)} \\
&\geq \delta + 0 - k\Big(\mathbb{P}^{\epsilon}_{\mathcal{D}_{\text{any}}}(F',\gamma) - \mathbb{P}_{\mathcal{D}_{\text{any}}}(F',\gamma)\Big) && \text{(F.396)} \\
&> 0. && \text{(F.397)}
\end{aligned}
$$

Equation (F.396) follows by proposition F.302. By corollary F.305, $\gamma \in (0,1)$ and $\delta > 0$ implies that we can choose ϵ so that $\mathbb{P}^{\epsilon}_{\mathcal{D}_{\text{any}}}(F,\gamma) - \mathbb{P}_{\mathcal{D}_{\text{any}}}(F,\gamma) < \frac{\delta}{k}$; by doing so, eq. (F.397) follows. ∎

Conjecture F.307 (ϵ-optimality results hold for RSD optimality probability)**.**

For shorthand, we define action ϵ-optimality probability, mirroring definition 5.10's definition of action optimality probability.

Definition F.308 (Action ϵ-optimality probability)**.** Let $\epsilon \geq 0$. At discount rate γ and at state s, the *ϵ-optimality probability of action a* is

$$\mathbb{P}^{\epsilon}_{\mathcal{D}_{\text{any}}}(s,a,\gamma) := \mathop{\mathbb{P}}_{R\sim\mathcal{D}_{\text{any}}}\left(\exists \pi \in \Pi^{\epsilon}(R,\gamma) : \pi(s) = a\right).$$

Proposition F.309 (Action ϵ-opt. probability is a special case of visit distribution ϵ-opt. prob.)**.** *For any $\gamma \in [0,1]$ and $\epsilon \geq 0$, $\mathbb{P}^{\epsilon}_{\mathcal{D}_{any}}(s,a,\gamma) = \mathbb{P}^{\epsilon}_{\mathcal{D}_{any}}\big(\mathcal{F}(s \mid \pi(s)=a),\gamma\big)$.*

Proof. Let $F_a := \mathcal{F}(s \mid \pi(s)=a)$. For $\gamma \in [0,1]$ and any state s,

$$
\begin{aligned}
\mathbb{P}^{\epsilon}_{\mathcal{D}_{\text{any}}}(s,a,\gamma) &:= \mathop{\mathbb{P}}_{R\sim\mathcal{D}_{\text{any}}}\big(\exists \pi \in \Pi^{\epsilon}(R,\gamma) : \pi(s) = a\big) && \text{(F.398)} \\
&= \mathop{\mathbb{P}}_{\mathbf{r}\sim\mathcal{D}_{\text{any}}}\big(\exists \mathbf{f}^{\pi} \in F_a : \pi \in \Pi^{\epsilon}(R,\gamma)\big) && \text{(F.399)} \\
&=: \mathbb{P}^{\epsilon}_{\mathcal{D}_{\text{any}}}(F_a,\gamma). && \text{(F.400)}
\end{aligned}
$$

Equation (F.399) follows because $\mathbf{f} \in F_a$ iff $\exists \pi \in \Pi : \mathbf{f}^\pi = \mathbf{f}, \pi(s) = a$ by the definition of $\mathcal{F}(s \mid \pi(s) = a)$ (definition 5.4). $\square$

Theorem F.310 (Optimal POWER-seeking incentives imply ϵ-optimal POWER-seeking incentives)**.** *Let $s \in \mathcal{S}, a, a' \in \mathcal{A}, \gamma \in (0,1)$. If*

$$\mathbb{E}_{s' \sim T(s,a)}\left[\text{POWER}_{\mathcal{D}_{bound}}\left(s', \gamma\right)\right] > \mathbb{E}_{s' \sim T(s,a')}\left[\text{POWER}_{\mathcal{D}_{bound}}\left(s', \gamma\right)\right]$$

and

$$\mathbb{P}_{\mathcal{D}_{any}}\left(s, a, \gamma\right) > \mathbb{P}_{\mathcal{D}_{any}}\left(s, a', \gamma\right),$$

then there exists $\epsilon > 0$ such that

$$\mathbb{E}_{s' \sim T(s,a)}\left[\text{POWER}^{pol}_{\mathcal{D}_{bound}}\left(s', \gamma\right)\right] > \mathbb{E}_{s' \sim T(s,a')}\left[\text{POWER}^{pol}_{\mathcal{D}_{bound}}\left(s', \gamma\right)\right]$$

and

$$\mathbb{P}^{\epsilon}_{\mathcal{D}_{any}}\left(s, a, \gamma\right) > \mathbb{P}^{\epsilon}_{\mathcal{D}_{any}}\left(s, a', \gamma\right),$$

where pol is any ϵ-optimal policy-generating function.

Proof. Theorem F.295 guarantees the existence of $\epsilon_{\text{POWER}} > 0$ such that

$$\mathbb{E}_{s' \sim T(s,a)}\left[\text{POWER}^{\text{pol}}_{\mathcal{D}_{\text{bound}}}\left(s', \gamma\right)\right] > \mathbb{E}_{s' \sim T(s,a')}\left[\text{POWER}^{\text{pol}}_{\mathcal{D}_{\text{bound}}}\left(s', \gamma\right)\right].$$

Lemma F.306 guarantees the existence of $\epsilon_{\mathbb{P}} > 0$ such that

$$\mathbb{P}^{\epsilon_{\mathbb{P}}}_{\mathcal{D}_{\text{any}}}\left(s, a, \gamma\right) = \mathbb{P}^{\epsilon_{\mathbb{P}}}_{\mathcal{D}_{\text{any}}}\left(\mathcal{F}(s \mid \pi(s) = a), \gamma\right) \tag{F.401}$$

$$> \mathbb{P}^{\epsilon_{\mathbb{P}}}_{\mathcal{D}_{\text{any}}}\left(\mathcal{F}(s \mid \pi(s) = a'), \gamma\right) \tag{F.402}$$

$$= \mathbb{P}^{\epsilon_{\mathbb{P}}}_{\mathcal{D}_{\text{any}}}\left(s, a', \gamma\right). \tag{F.403}$$

Equation (F.401) and eq. (F.403) follow from proposition F.309.

Choosing $\epsilon < \min\left(\epsilon_{\text{POWER}}, \epsilon_{\mathbb{P}}\right)$ ensures that both strict inequalities hold. Since both ϵ_{POWER} and $\epsilon_{\mathbb{P}}$ are positive, we can ensure that $\epsilon > 0$. $\square$

The $\textsc{Power}_{\mathcal{D}_{\text{bound}}}$-seeking theorems are only concerned with the probability of the *existence* of ϵ-optimal policies which seek $\textsc{Power}_{\mathcal{D}_{\text{bound}}}$. Supposing that an agent is only constrained to follow some ϵ-optimal policy, the agent may end up seeking $\textsc{Power}_{\mathcal{D}_{\text{bound}}}$ with lower probability due to the influence of *e.g.* some tie-breaking rule for policy selection.

Conjecture F.311 (Continuous distributions have continuous ϵ-optimality probability functions)**.** Fixing any $s \in \mathcal{S}, F \subseteq \mathcal{F}(s), \gamma \in [0,1]$, $\mathbb{P}^{\epsilon}_{\mathcal{D}_{\text{cont}}}(F, \gamma)$ is continuous on $\epsilon \in [0, \infty)$.

Bibliography

[1] David Abel, Will Dabney, Anna Harutyunyan, Mark K Ho, Michael Littman, Doina Precup, and Satinder Singh. On the expressivity of Markov reward. *Advances in Neural Information Processing Systems*, 34, 2021.

[2] Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 22–31, 2017.

[3] Eitan Altman. *Constrained Markov decision processes*, volume 7. CRC Press, 1999.

[4] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in AI safety. *arXiv:1606.06565 [cs]*, June 2016. URL http://arxiv.org/abs/1606.06565. arXiv: 1606.06565.

[5] Usman Anwar, Shehryar Malik, Alireza Aghasi, and Ali Ahmed. Inverse constrained reinforcement learning. *ICML*, 2021.

[6] Stuart Armstrong and Benjamin Levinstein. Low impact artificial intelligences. *arXiv:1705.10720 [cs]*, May 2017. URL http://arxiv.org/abs/1705.10720. arXiv: 1705.10720.

[7] Chris L Baker, Joshua B Tenenbaum, and Rebecca R Saxe. Goal inference as inverse planning. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 29, 2007.

[8] Tsvi Benson-Tilsen and Nate Soares. Formalizing convergent instrumental goals. *Workshops at the Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[9] Felix Berkenkamp, Matteo Turchetta, Angela Schoellig, and Andreas Krause. Safe model-based reinforcement learning with stability guarantees. In *Advances in Neural Information Processing Systems*, pages 908–918, 2017.

[10] Vikram R. Bhargava and Manuel Velasquez. Ethics of the attention economy: The

problem of social media addiction. *Business Ethics Quarterly*, 31(3):321–359, 2021. doi: 10.1017/beq.2020.32.

[11] David Blackwell. Discrete dynamic programming. *The Annals of Mathematical Statistics*, page 9, 1962.

[12] Nick Bostrom. The superintelligent will: Motivation and instrumental rationality in advanced artificial agents. *Minds and Machines*, 22(2):71–85, 2012.

[13] Nick Bostrom. *Superintelligence*. Oxford University Press, 2014.

[14] Gwern Branwen. Complexity no bar to AI, 2014. https://www.gwern.net/Complexity-vs-AI.

[15] Gwern Branwen. It looks like you're trying to take over the world, 2022. https://www.gwern.net/fiction/Clippy.

[16] Frank M Brown. *The frame problem in artificial intelligence: Proceedings of the 1987 workshop*. Morgan Kaufmann, 2014.

[17] Yuri Burda, Harri Edwards, Deepak Pathak, Amos Storkey, Trevor Darrell, and Alexei A. Efros. Large-scale study of curiosity-driven learning. In *International Conference on Learning Representations*, 2019.

[18] Bart Bussmann, Jacqueline Heinerman, and Joel Lehman. Towards empathic deep Q-learning. *arXiv preprint arXiv:1906.10918*, 2019.

[19] Ryan Carey. Incorrigibility in the CIRL framework. *AI, Ethics, and Society*, 2018.

[20] Ryan Carey. How useful is quantilization for mitigating specification gaming? 2019.

[21] Yinlam Chow, Ofir Nachum, Edgar Duenez-Guzman, and Mohammad Ghavamzadeh. A Lyapunov-based approach to safe reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 8092–8101, 2018.

[22] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, pages 4299–4307, 2017.

[23] Chris Drummond. Composing functions to speed up reinforcement learning in a changing world. In *Machine Learning: ECML-98*, volume 1398, pages 370–381. Springer, 1998.

[24] Yuqing Du, Stas Tiomkin, Emre Kiciman, Daniel Polani, Pieter Abbeel, and Anca

Dragan. AvE: Assistance via Empowerment. *Advances in Neural Information Processing Systems*, 33, 2020.

[25] Adrien Ecoffet, Joost Huizinga, Joel Lehman, Kenneth O Stanley, and Jeff Clune. First return, then explore. *Nature*, 590(7847):580–586, 2021.

[26] Effective Altruism Cambridge. Technical alignment curriculum, 2022. https://www.eacambridge.org/technical-alignment-curriculum.

[27] Tom Everitt, Victoria Krakovna, Laurent Orseau, and Shane Legg. Reinforcement learning with a corrupted reward channel. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 4705–4713, 2017. doi: 10.24963/ijcai.2017/656. URL https://doi.org/10.24963/ijcai.2017/656.

[28] Benjamin Eysenbach, Shixiang Gu, Julian Ibarz, and Sergey Levine. Leave no trace: Learning to reset for safe and autonomous reinforcement learning. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=S1vuO-bCW.

[29] Jerome A Feldman and Dana H Ballard. Connectionist models and their properties. *Cognitive science*, 6(3):205–254, 1982.

[30] David Foster and Peter Dayan. Structure in the space of value functions. *Machine Learning*, pages 325–346, 2002.

[31] Javier García and Fernando Fernández. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(1):1437–1480, 2015.

[32] Christian Guckelsberger, Christoph Salge, and Simon Colton. Intrinsically motivated general companion NPCs via coupled empowerment maximisation. In *IEEE Conference on Computational Intelligence and Games*, pages 1–8, 2016.

[33] Christian Guckelsberger, Christoph Salge, and Julian Togelius. New and surprising ways to be mean. In *IEEE Conference on Computational Intelligence and Games*, pages 1–8, 2018.

[34] Dylan Hadfield-Menell, Stuart Russell, Pieter Abbeel, and Anca Dragan. Cooperative inverse reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 3909–3917, 2016.

[35] Dylan Hadfield-Menell, Anca Dragan, Pieter Abbeel, and Stuart Russell. The off-switch game. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 220–227, 2017.

[36] Dylan Hadfield-Menell, Smitha Milli, Pieter Abbeel, Stuart Russell, and Anca Dragan. Inverse reward design. In *Advances in Neural Information Processing Systems*, pages 6765–6774, 2017.

[37] Jeff Hawkins and Sandra Blakeslee. *On intelligence*. Macmillan, 2004.

[38] Evan Hubinger, Chris van Merwijk, Vladimir Mikulik, Joar Skalse, and Scott Garrabrant. Risks from learned optimization in advanced machine learning systems. *arXiv:1906.01820 [cs]*, June 2019. URL http://arxiv.org/abs/1906.01820. arXiv: 1906.01820.

[39] M Inuiguchi and M Sakawa. An achievement rate approach to linear programming problems with an interval objective function. *Journal of the Operational Research Society*, 48(1):25–33, 1997. ISSN 0160-5682, 1476-9360. doi: 10.1057/palgrave.jors.2600322. URL https://www.tandfonline.com/doi/full/10.1057/palgrave.jors.2600322.

[40] Victoria Krakovna, Laurent Orseau, Ramana Kumar, Miljan Martic, and Shane Legg. Penalizing side effects using stepwise relative reachability. *arXiv preprint arXiv:1806.01186*, 2018.

[41] Victoria Krakovna, Laurent Orseau, Miljan Martic, and Shane Legg. Measuring and avoiding side effects using relative reachability. *arXiv preprint arXiv:1806.01186*, 2018.

[42] Victoria Krakovna, Laurent Orseau, Richard Ngo, Miljan Martic, and Shane Legg. Avoiding side effects by considering future tasks. In *Advances in Neural Information Processing Systems*, 2020.

[43] Victoria Krakovna, Jonathan Uesato, Vladimir Mikulik, Matthew Rahtz, Tom Everitt, Ramana Kumar, Zac Kenton, Jan Leike, and Shane Legg. Specification gaming: The flip side of AI ingenuity, 2020. URL https://deepmind.com/blog/article/Specification-gaming-the-flip-side-of-AI-ingenuity.

[44] Tejas D Kulkarni, Ardavan Saeedi, Simanta Gautam, and Samuel J Gershman. Deep successor reinforcement learning. *arXiv preprint arXiv:1606.02396*, 2016.

[45] Yann LeCun and Anthony Zador. Don't fear the Terminator. *Scientific American Blog*, September 2019. URL https://blogs.scientificamerican.com/observations/dont-fear-the-terminator/.

[46] Gavin Leech, Karol Kubicki, Jessica Cooper, and Tom McGrath. Preventing side-effects in gridworlds, 2018. URL https://www.gleech.org/grids/.

[47] Jan Leike, Miljan Martic, Victoria Krakovna, Pedro Ortega, Tom Everitt, Andrew Lefrancq, Laurent Orseau, and Shane Legg. AI safety gridworlds. *arXiv:1711.09883 [cs]*, November 2017. URL http://arxiv.org/abs/1711.09883. arXiv: 1711.09883.

[48] Steven A Lippman. On the set of optimal policies in discrete dynamic programming. *Journal of Mathematical Analysis and Applications*, 24(2):440–445, 1968.

[49] Michael L. Littman, Thomas L. Dean, and Leslie Pack Kaelbling. On the complexity of solving Markov decision problems. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, UAI'95, page 394–402, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc. ISBN 1558603859.

[50] Gabriel Loaiza-Ganem and John P Cunningham. The continuous Bernoulli: Fixing a pervasive error in variational autoencoders. In *Advances in Neural Information Processing Systems*, pages 13266–13276, 2019.

[51] Lionel W McKenzie. Turnpike theory. *Econometrica: Journal of the Econometric Society*, pages 841–865, 1976.

[52] Ishai Menache, Shie Mannor, and Nahum Shimkin. Q-cut – Dynamic discovery of sub-goals in reinforcement learning. In *European Conference on Machine Learning*, pages 295–306. Springer, 2002.

[53] Smitha Milli, Dylan Hadfield-Menell, Anca Dragan, and Stuart Russell. Should robots be obedient? In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 4754–4760, 2017.

[54] Melanie Mitchell. Why AI is harder than we think. *arXiv preprint arXiv:2104.12871*, 2021.

[55] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

[56] Shakir Mohamed and Danilo Jimenez Rezende. Variational information maximisation for intrinsically motivated reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 2125–2133, 2015.

[57] Teodor Mihai Moldovan and Pieter Abbeel. Safe exploration in Markov decision processes. *ICML*, 2012.

[58] Andrew Y. Ng, Daishi Harada, and Stuart Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *Proceedings of the*

Sixteenth International Conference on Machine Learning, pages 278–287. Morgan Kaufmann, 1999.

[59] Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 2020.

[60] Stephen Omohundro. The basic AI drives, 2008.

[61] OpenAI. OpenAI Five. https://blog.openai.com/openai-five/, 2018.

[62] Toby Ord. *The precipice: Existential risk and the future of humanity*. Hachette Books, 2020.

[63] Ben Pace. Debate on instrumental convergence between LeCun, Russell, Bengio, Zador, and more, 2019. https://www.alignmentforum.org/posts/WxW6Gc6f2z3mzmqKs/debate-on-instrumental-convergence-between-lecun-russell.

[64] Christos H Papadimitriou and John N Tsitsiklis. The complexity of Markov decision processes. *Mathematics of operations research*, 12(3):441–450, 1987.

[65] Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *ICML*, 2017.

[66] Martin Pecka and Tomas Svoboda. Safe exploration techniques for reinforcement learning–an overview. In *International Workshop on Modelling and Simulation for Autonomous Systems*, pages 357–375. Springer, 2014.

[67] Steven Pinker and Stuart Russell. The foundations, benefits, and possible existential threat of AI, June 2020. URL https://futureoflife.org/2020/06/15/steven-pinker-and-stuart-russell-on-the-foundations-benefits-and-possible-existential-risk-of-ai/.

[68] Martin L Puterman. *Markov decision processes: Discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

[69] Alex Ray, Joshua Achiam, and Dario Amodei. Benchmarking safe exploration in deep reinforcement learning. *arXiv preprint arXiv:1910.01708*, 2019.

[70] J.B. Reece and N.A. Campbell. *Campbell biology*. Pearson Australia, 2011.

[71] Kevin Regan and Craig Boutilier. Robust policy computation in reward-uncertain MDPs using nondominated policies. In *AAAI*, 2010.

[72] Kevin Regan and Craig Boutilier. Robust policy computation in reward-uncertain

MDPs using nondominated policies. In *Twenty-Fourth AAAI Conference on Artificial Intelligence*, 2010.

[73] Paul Rendell. Turing universality of the Game of Life. In *Collision-based computing*, pages 513–539. Springer, 2002.

[74] Stuart Russell. *Human compatible: Artificial intelligence and the problem of control*. Viking, 2019.

[75] Stuart J Russell and Peter Norvig. *Artificial intelligence: A modern approach*. Pearson Education Limited, 2009.

[76] Christoph Salge and Daniel Polani. Empowerment as replacement for the three laws of robotics. *Frontiers in Robotics and AI*, 4:25, 2017.

[77] Christoph Salge, Cornelius Glackin, and Daniel Polani. Empowerment–an introduction. In *Guided Self-Organization: Inception*, pages 67–114. Springer, 2014.

[78] Anders Sandberg and Stuart Armstrong. Indefinite survival through backup copies. 2012.

[79] Faridun Sattarov. *Power and technology: A philosophical and ethical analysis*. Rowman & Littlefield International, Ltd, 2019.

[80] William Saunders, Girish Sastry, Andreas Stuhlmueller, and Owain Evans. Trial without error: Towards safe reinforcement learning via human intervention. In *Proceedings of the 17th International Conference on Autonomous Agents and Multi-Agent Systems*, pages 2067–2069, 2018.

[81] Tom Schaul, Daniel Horgan, Karol Gregor, and David Silver. Universal value function approximators. In *International Conference on Machine Learning*, pages 1312–1320, 2015.

[82] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[83] Rohin Shah, Dmitrii Krasheninnikov, Jordan Alexander, Pieter Abbeel, and Anca Dragan. The implicit preference information in an initial state. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=rkevMnRqYQ.

[84] Rohin Shah, Pedro Freire, Neel Alex, Rachel Freedman, Dmitrii Krasheninnikov, Lawrence Chan, Michael Dennis, Pieter Abbeel, Anca Dragan, and Stuart Russell. Benefits of assistance over reward learning. 2021.

[85] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484, 2016.

[86] Herbert A Simon. Rational choice and the structure of the environment. *Psychological review*, 63(2):129, 1956.

[87] Nate Soares, Benja Fallenstein, Stuart Armstrong, and Eliezer Yudkowsky. Corrigibility. *AAAI Workshops*, 2015.

[88] Kaj Sotala. Advantages of artificial intelligences, uploads, and digital minds. *International journal of machine consciousness*, 4(01):275–291, 2012.

[89] Jacob Stavrianos and Alexander Turner. Game-theoretic alignment in terms of attainable utility, 2021. https://www.alignmentforum.org/posts/buaGz3aiqCotzjKie/game-theoretic-alignment-in-terms-of-attainable-utility.

[90] Jacob Stavrianos and Alexander Turner. Generalizing power to multi-agent games, 2021. https://www.alignmentforum.org/posts/MJc9AqyMWpG3BqfyK/generalizing-power-to-multi-agent-games.

[91] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT Press, 1998.

[92] Richard S Sutton, Joseph Modayil, Michael Delp, Thomas Degris, Patrick M Pilarski, Adam White, and Doina Precup. Horde: A scalable real-time architecture for learning knowledge from unsupervised sensorimotor interaction. In *International Conference on Autonomous Agents and Multiagent Systems*, pages 761–768, 2011.

[93] Jessica Taylor. Quantilizers: A safer alternative to maximizers for limited optimization. In *AAAI Workshop: AI, Ethics, and Society*, 2016.

[94] Alexander Turner. Probability of one sample mean being maximal among a set of other sample means, 2020. https://math.stackexchange.com/questions/3571961/probability-of-one-sample-mean-being-maximal-among-a-set-of-other-sample-means.

[95] Alexander Matt Turner. Reframing impact, 2019. https://www.alignmentforum.org/s/7CdoznhJaLEKHwvJW.

[96] Alexander Matt Turner. The causes of instrumental convergence and power-seeking, 2021. https://www.alignmentforum.org/s/fSMbebQyR4wheRrvk.

[97] Alexander Matt Turner, Dylan Hadfield-Menell, and Prasad Tadepalli. Conservative agency via attainable utility preservation. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 385–391, 2020.

[98] Alexander Matt Turner, Neale Ratzlaff, and Prasad Tadepalli. Avoiding side effects in complex environments. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

[99] Alexander Matt Turner, Logan Smith, Rohin Shah, Andrew Critch, and Prasad Tadepalli. Optimal policies tend to seek power. In *Advances in Neural Information Processing Systems*, 2021.

[100] Carroll L. Wainwright and Peter Eckersley. SafeLife 1.0: Exploring side effects in complex environments, 2019.

[101] Tao Wang, Michael Bowling, and Dale Schuurmans. Dual representations for dynamic programming and reinforcement learning. In *International Symposium on Approximate Dynamic Programming and Reinforcement Learning*, pages 44–51. IEEE, 2007.

[102] Tao Wang, Michael Bowling, Dale Schuurmans, and Daniel J Lizotte. Stable dual dynamic programming. In *Advances in Neural Information Processing Systems*, pages 1569–1576, 2008.

[103] Christopher Watkins and Peter Dayan. Q-learning. *Machine Learning*, 8(3-4): 279–292, 1992.

[104] Eliezer Yudkowsky. Intelligence explosion microeconomics. 2013.

[105] Shun Zhang, Edmund H Durfee, and Satinder P Singh. Minimax-regret querying on side effects for safe optimality in factored Markov decision processes. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 4867–4873, 2018.