AN ABSTRACT OF THE THESIS OF

Andrew Jensen for the degree of Master of Science in Mathematics presented on August 24, 2018.

Title: A Homotopy Strategy for Accelerated Sampling

Abstract approved:

Juan M. Restrepo

Markov Chain Monte Carlo methods may be used to determine normalizations and moments of distributions. However, these methods may perform poorly when starting from distributions that have little overlap with the target. We develop a homotopy based iterative process of incremental importance sampling to normalize distributions when observations can only be made from those distributions that differ from target. In addition to this topic, discussions are included of my experience in the NSF Risk and Uncertainty Traineeship Program (NRT) applying mathematical knowledge to a transdisciplinary study of the Dungeness crab.

A Homotopy Strategy for Accelerated Sampling

by
Andrew Jensen

A THESIS

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Master of Science

Presented August 24, 2018
Commencement June 2019

Master of Science thesis of Andrew Jensen presented on August 24, 2018

APPROVED:

Major Professor, representing Mathematics

Head of the Department of Mathematics

Dean of the Graduate School

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

Andrew Jensen, Author

Table of Contents

1 Background

1.1 History and Context

The universe in which we reside is governed by complicated and convoluted processes. From the probabilistic behavior of the quantum world to the chaotic interactions of global weather systems, there is uncertainty lurking in every physical process we try to understand. This necessitates the use of statistical and probabilistic methods that seek to quantify and reduce this uncertainty. In most systems, one issue is an inability to conceptualize the entire domain of the system mathematically. For example, weather systems can only be understood through discrete observations, such as weather stations and satellite data. This incomplete data set only gives us a sample of the entire system, from which we need to extrapolate. In this example, the variables of a weather phenomena, such as wind speed, pressure, and temperature, may be theoretically expressed as a function, but practically understood through limited data points. This data points represent a snapshot of this entire system however, and can be used to understand the larger picture.

The infeasibility of studying these systems by hand delayed much of their mathematical understanding until the development of computers. Early computers, engineered to solve mathematical problems and operated at research institutes such as Los Alamos National Labs, immediately began to see use in studying random processes. The term Monte Carlo, named after the famous casino, originated to describe the process of random processes, and soon expanded to include Markov Chain Monte Carlo (MCMC) methods in the 1950s [3]. The breakthrough occurred at Los Alamos due to the work of Metropolis, Von Neumann and others [7] studying statistical mechanics. The revelation that a system's exact dynamics were not needed, only a description of the random process, allowed these statistically guided methods to study complicated processes thus far largely untouched. The use of these

early computers such as the Electronic Numerical Integrator and Computer (ENIAC) enabled these large calculations to be performed, which led to the development of theories to model weather, trajectories, and the development of atomic weaponry [15]. The ensuing decades saw the development of new and more versatile and powerful algorithms, such as the 1970 extension of the Metropolis algorithm known as Metropolis-Hastings [3]. Following this was the invention of Gibbs sampling and further development of these techniques. Enabling progress in Bayesian inference and allowing these techniques to be applied to contribute to the growth of modern data and machine learning methods.

To discuss these methods in more detail let us define some basics of probability and statistics. In order to discuss sampling from a probability distribution, we need to defined what distinguishes a probability distribution from a general function. First let us define a *probability space*. A Probability space is defined as the triplet $(\Omega, \mathcal{F}, P)$, where Ω is a non-empty set, $\mathcal{F}$ is a $\sigma-$algebra on Ω, i.e. a collection of subsets of Ω such that the empty set is included, and the collection is closed under complement, countable unions, and countable intersections [Jones]. Finally, P, known as probability, is a measure on this space such that $P(\Omega) = 1$. Note that this is a particular case of a measure space under a finite measure. The set Ω may represent the set of all outcomes of a random system, and the elements of $\mathcal{F}$ are particular events [16]. In the context of weather modeling, a set Ω might represent all possible weather variables combinations at a particular point, where an event might comprise a particular range of variables, e.g. temperatures between 10°C and 15°C. Then a particular probability would be assigned to that event occurring.

For a particular probability space, we can define a random variable X as a measurable function from Ω. Assume for now that X is a real-valued function. For a given random variable we can define a cumulative density function (CDF) $F_X(x) = P[y \in \Omega : X(y) \leq x]$ [14]. Note that this function describes

the probability that our random variable falls below a particular number, and that $0 \leq F_X \leq 1$. We can now define a new Lebesgue-measurable function $f_X(x)$, denoted the probability density function (PDF). This function is formally defined as a Radon-Nikodym derivative, but for our purposes we define it as the derivative of the CDF. Since the CDF is a function over $\mathbb{R}$ rather than Ω, we will integrate with respect to the standard Lebesgue measure. We thus define f_X such that [14]

$$F_X(x) = \int_{-\infty}^{x} f_X(y)dy.$$

With the basics of probability established, let us proceed to define some basics of sampling. We define a sample as a randomly selected subset of some set Ω. As the number of samples increases, the sample will clearly converge to the entire set. However, we can sample in a more informed manner, such as in the case of inverse transform sampling. Suppose that we have a random variable with PDF f_X and CDF F_X. We wish to sample in such a manner to be representative of the distribution represented by f_X. Let us first define the inverse of F_X, denoted F_X^{-1}, such that $F_X\big(F_X^{-1}(x)\big) = x$. Note that F_X is increasing, but potentially not strictly. Therefore it may not be injective, meaning this may be a one-sided inverse. We claim that if $u \sim \mathcal{U}$ is the uniform random variable between $(0, 1)$, then the cumulative distribution function of $F_X^{-1}(u)$ is in fact F_X [10].

Proof: The CDF of $F_X^{-1}(u)$ is given by $P[u : F_X^{-1}(u) < x]$. Applying F to each side of the inequality results in $P[u : u < F_X(x)]$. This is equivalent to $F_X(x)$. Thus the CDF of $F_X^{-1}(u)$ is F_X and $F_X^{-1}(u)$ generates samples according to X.

Let us now proceed to describe several of the major MCMC and sampling methods described above. We begin by describing the process of standard Monte Carlo methods, for example to calculate an integral such as expectation. Suppose that we have a random variable X on a probability space such

that X has a probability density function f_X. Then the expectation of some function $g(X)$ can be given by

$$\mathrm{E}[g(X)] = \int_{\Omega} g(x)f(x)dx.$$

If we take a set of N samples $\{x_1, x_2, \ldots, x_N\}$ from the CDF of X in the previous manner, we can approximate this integral as

$$\mathrm{E}[g(X)] \approx \bar{g}_N = \frac{1}{N}\sum_{j=i}^{N} g(x_j).$$

By the law of large numbers, this will converge to our expected value if we let N tend to infinity. We can furthermore calculate the variance of $\bar{g}_N$ by another application of this sampling [10]:

$$\mathrm{Var}(\bar{g}_N) \approx \frac{1}{N^2}\sum_{j=1}^{N} \big(g(x_j) - \bar{g}_N\big)^2.$$

As mentioned, an extension of Monte-Carlo methods are Markov Chain Monte Carlo (MCMC) methods. In a standard process of Monte-Carlo sampling, each sample is taken from the same distribution. In an MCMC process, each new sample may be generated from a different distribution in a type of walk called a Markov Chain. The definition of a Markov Chain is a sequence of random variables X_i on a probability space such that they satisfy two conditions [18]:

a) Markov Property: For all i, $P[X_{i+1} = x_{i+1}|X_i = x_i, \ldots, X_1 = x_1] = P[X_{i+1} = x_{i+1}|X_i = x_i]$.

b) Time Homogeneity: $P[X_{i+1} = x|X_i = y] = P[X_{j+1} = x|X_j = y]$

This equivalently means that the value of the next step in a Markov Chain

only depends upon the value of the previous step, not upon the place in the chain or the initial values. For an MCMC process, this means that each step resamples based only upon the previous sample. The process should eventually converge to the target distribution. A classic example of an MCMC process is Metropolis Hastings, which progresses from a starting density $g(y|x)$ to a target distribution $f(x)$. We begin with a starting sample x_0 and generate a new sample y from $g(y, x_0)$. We calculate

$$A(y, x_0) = \min\left(1, \frac{f(y)}{f(x_0)}\frac{g(x_0|y)}{g(y|x_0)}\right).$$

Generating a random number $u \sim \mathcal{U}(0, 1)$, if $u \leq A(y, x_0)$ then we accept y and let $x_1 = y$. Otherwise we reject at let $x_1 = x_0$. We then proceed to increment and select a new value y to test and select x_2 [10].

We will finally discuss the method of importance sampling, which is utilized in the homotopy process. Importance sampling is an extension of Monte-Carlo approximation, primarily used when we may not be able to sample from our intended distribution. Suppose that we are trying to calculate the expectation of some function $g(X)$ where X has PDF $f(x)$. In the case where we cannot or choose not to sample from $f(x)$, if we can find another distribution $q(x)$ that is close to proportional to $|f(x)|$ and that $q(x) > 0$ when $f(x) \neq 0$, then we can rewrite our integral and sample from q as so:

$$\mathrm{E}[g(X)] = \int_\Omega g(x)f(x)dx = \int_\Omega g(x)\frac{f(x)}{q(x)}q(x)dx = \mathrm{E}\left[g(x)\frac{f(x)}{q(x)}\right]$$

over the distribution q. This can then be approximated as

$$\mathrm{E}\left[g(x)\frac{f(x)}{q(x)}\right] = \frac{1}{N}\sum_{j=1}^{N} g(x_j)\frac{f(x_j)}{q(x_j)}$$

where x_j are sampled from q [10].

2 Homotopy With Importance Sampling

2.1 Motivation

As previously discussed, there are particular issues that may require sampling using a specific technique. If we cannot directly sample from our target distribution we must use a technique that samples from another distribution such as Metropolis-Hastings or Importance sampling. However, with both these cases, the starting distribution from which we sample must be close to our target distribution. We theorize that if we are forced to sample from a distribution that is further away from the target distribution, that we will see performance decrease shown by a longer burn-in period. We seek to avoid these problems by moving from our starting distribution to our target distribution in a more informed manner rather than through a random walk.

To do so, we develop a method of implementing repeated importance sampling over a homotopy path, in order to perform Monte-Carlo method integration and calculate normalization constants of nonnormalized distributions. Suppose we seek the normalization of a distribution $\beta(x)$ from which we cannot directly sample. We can however sample from a known distribution α, who resembles β but is distant in mean. We can begin with our observed distribution $\alpha(x)$ and define a continuous homotopy function $\phi_t(x)$, where $\phi_0(x) = \alpha(x)/Z_0$ and $\phi_1(x) = \beta(x)/Z_1$ where Z_0 and Z_1 are the distribution's respective normalization constants. Let us explicitly define this homotopy by $\phi_t(x) = (1/Z_t)\beta^t(x)\alpha^{(1-t)}(x)$, where we define Z_t as

$$Z_t = \int_{\mathbb{R}} \beta^t(x)\alpha^{(1-t)}(x)dx,$$

resulting in Z_0 being the normalization of α and Z_1 the normalization of β that we seek. Note that ϕ_t is a probability density function for every t. From our definition of Z_t, we see that by traveling a step size ϵ from t we obtain a

new normalization value of

$$\begin{aligned} Z_{t+\epsilon} &= \int_{\mathbb{R}} \beta^{t+\epsilon}(x)\alpha^{1-t-\epsilon}(x)dx \\ &= \int_{\mathbb{R}} \left(\frac{\beta(x)}{\alpha(x)}\right)^{\epsilon} \beta^{t}(x)\alpha^{1-t}(x)dx. \end{aligned}$$

Dividing by the previous term results in

$$\frac{Z_{t+\epsilon}}{Z_t} = \frac{\displaystyle\int_R \left(\frac{\beta(x)}{\alpha(x)}\right)^{\epsilon} \beta^{t}(x)\alpha^{1-t}(x)dx}{\displaystyle\int_R \beta^{t}(x)\alpha^{1-t}(x)dx} = \int_{\mathbb{R}} \left(\frac{\beta(x)}{\alpha(x)}\right)^{\epsilon} d\Phi_t(x)$$

where $\Phi_t(x)$ is the Cumulative Distribution Function of $\phi_t(x)$. We can consider this value $Z_{t+\epsilon}/Z_t$ as the integral of $(\beta/\alpha)^{\epsilon}$ taken with respect to the CDF of the previous step, introducing the notation:

$$\frac{Z_{t+\varepsilon}}{Z_t} = \left\langle \left(\frac{\beta(x)}{\alpha(x)}\right)^{\varepsilon} \right\rangle_t.$$

Assuming that we evaluate our homotopy over the course of M equally sized steps, we can let $\epsilon = 1/M$. We can write the quotient Z_1/Z_0 as

$$\begin{aligned} \frac{Z_1}{Z_0} &= \frac{Z_{1/M}}{Z_0} \cdot \frac{Z_{2/M}}{Z_{1/M}} \cdots \frac{Z_1}{Z_{(M-1)/M}} \\ &= \prod_{m=1}^{M} \frac{Z_{m/M}}{Z_{(m-1)/M}} \\ &= \prod_{m=1}^{M} \left\langle \left(\frac{\beta(x)}{\alpha(x)}\right)^{\frac{1}{M}} \right\rangle_{\frac{m}{M}} \end{aligned}$$

Taking a natural logarithm of each side we obtain

$$\ln Z_1 = \sum_{m=1}^{M} \ln \left\langle \left(\frac{\beta(x)}{\alpha(x)}\right)^{\frac{1}{M}} \right\rangle_{\frac{m}{M}} + \ln Z_0. \tag{1}$$

Supposing we know Z_0, (1) is a recursive schedule for the calculation of Z_1 in M homotopy steps. If we use a sample estimate for each of the homotopy calculations and we assume that each of these is calculated using N samples, (1) will be approximated as

$$\ln Z_1 \approx \sum_{m=1}^{M} \ln \left(\frac{1}{N} \sum_{j=1}^{N} \left(\frac{\beta\big(X_{m-1,j}(u)\big)}{\alpha\big(X_{m-1,j}(u)\big)} \right)^{\frac{1}{M}} \right) + \ln Z_0,$$

where $[X_m]_j$ is the j^{th} sample from the $(m-1)^{th}$ distribution

$$\frac{1}{Z_{m/M}} \beta^m(x) \alpha^{1-m}(x).$$

A balanced iterative variant would start with Z_0 known and then calculate, for $m = 1, ..., M$,

$$Z_m = Z_{m-1} \frac{1}{N} \sum_{j=1}^{N} \left(\frac{\beta\big(X_{m-1,j}(u)\big)}{\alpha\big(X_{m-1,j}(u)\big)} \right)^{\frac{1}{M}}$$

2.2 Example Calculation

As an example, we will build a homotopy between two Gaussian distributions centered at different means. Since we will be able to analytically solve for all intermediate steps of the procedure we can calculate deviations throughout the process. Let us start with the analytical computations for two arbitrary distributions and later directly apply the technique to two specific ones. Our homotopy γ_t will travel between distributions α, for which we assume the normalization is known, to a target distribution β, for which we seek to compute an estimation of $Z_1 = \int \beta(x)\, dx$. Define β as:

$$\beta(x) = e^{-\frac{(x-\mu_\beta)^2}{2\sigma_\beta^2}}.$$

We can however calculate the normalization as:

$$Z_1 = \frac{1}{\sqrt{4\pi\sigma_\beta^2}}.$$

This will allow us to determine the accuracy of our technique. We will examine how the homotopy process proceeds from starting Gaussian

$$\alpha = e^{-\frac{(x-\mu_\alpha)^2}{2\sigma_\alpha^2}}.$$

We will assume knowledge of

$$Z_0 = \int \alpha \, dx = \frac{1}{\sqrt{4\pi\sigma_\alpha^2}}.$$

Suppose that we have progressed to some intermediate normalization term Z_t (although note this could be the starting normalization Z_0). To progress to the next step $Z_{t+\epsilon}$ for some small change ϵ, we need to sample from the previous distribution. We can analytically solve for these values by computing the integral:

$$Z_t = \int_{-\infty}^{\infty} \beta^t(x)\alpha^{(1-t)}(x)dx = \frac{\sqrt{4\pi}\sigma_\beta\sigma_\alpha}{\sqrt{t\sigma_\alpha^2 - (t-1)\sigma_\beta^2}} \exp\left(-\frac{1}{4}\frac{(\mu_\beta - \mu_\alpha)^2(t-1)t}{\left(t\sigma_\alpha^2 - (t-1)\sigma_\beta^2\right)}\right).$$

We can find an analytical expression for our CDF at step t by:

$$\int_{-\infty}^{x} \frac{\beta^t\alpha^{(1-t)}}{Z_t}dy = \frac{1}{2}\left(1 + \operatorname{erf}\left[\frac{(1-t)(x-\mu_\alpha)\sigma_\beta^2 + t(x-\mu_\beta)\sigma_\alpha^2}{\sigma_\alpha\sigma_\beta\sqrt{2\left((1-t)\sigma_\beta^2 + t\sigma_\alpha^2\right)}}\right]\right)$$

where erf is the error function.

Assuming we have reached step m, to obtain our samples $[X_m]_j$ to proceed to step $m+1$, we can use this CDF to perform inverse transform sampling.

Suppose that $u \sim U(0,1)$, a sample from the uniform distribution. If we assume that for a given sample $[X_m]_j$, our CDF equals u we have

$$u = \frac{1}{2}\left(1 + \text{erf}\left[\frac{(1-m/M)([X_m]_j - \mu_\alpha)\sigma_\beta^2 + (m/M)([X_m]_j - \mu_\beta)\sigma_\alpha^2}{\sigma_\alpha\sigma_\beta\sqrt{2\big((1-t)\sigma_\beta^2 + t\sigma_\alpha^2\big)}}\right]\right).$$

Solving for $[X_m]_j$ yields

$$[X_m]_j = \frac{(1-m/M)\mu_\alpha\sigma_\beta^2 + (m/M)\mu_\beta\sigma_\alpha^2}{(1-m/M)\sigma_\beta^2 + (m/M)\sigma_\alpha^2} + \frac{\sigma_\alpha\sigma_\beta\,\text{erfinv}[2u-1]}{\sqrt{(1-m/M)\sigma_\beta^2 + (m/M)\sigma_\alpha^2}},$$

where erfinv is the inverse error function. For the case of a shared mean, $\mu_\alpha = \mu_\beta = 0$, we have the simpler, transformed term:

$$[X_m]_j = \frac{\sigma_\alpha\sigma_\beta\,\text{erfinv}[2u-1]}{\sqrt{(1-m/M)\sigma_\beta^2 + (m/M)\sigma_\alpha^2}}.$$

We may evaluate the sum

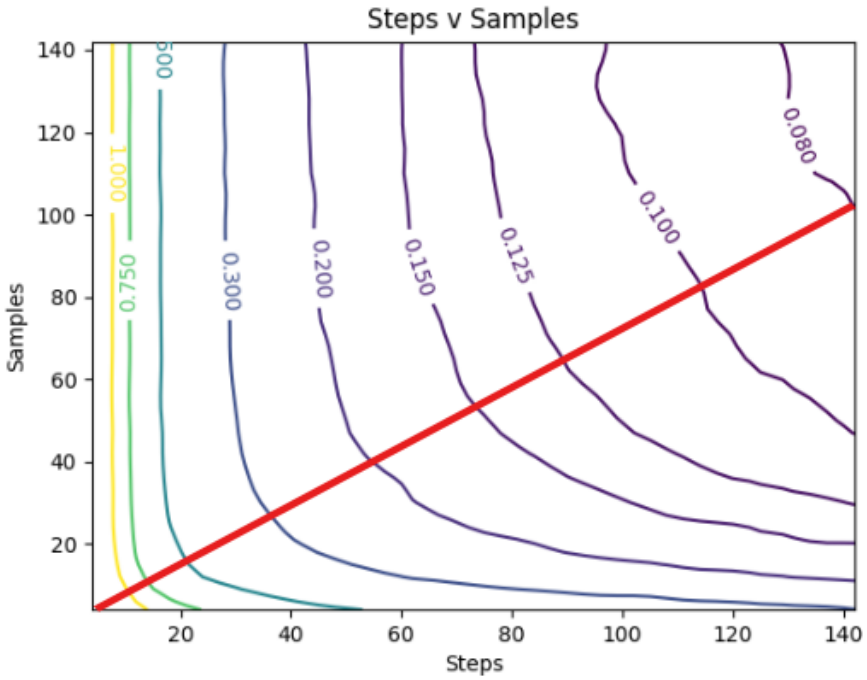
$$\frac{1}{N}\sum_{j=1}^{N}\left(\frac{\beta\big([X_m]_j\big)}{\alpha\big([X_m]_j\big)}\right)^\epsilon$$

to obtain the ratio Z_m/Z_{m-1}.

2.3 Optimizing the Algorithm

To test the procedure we implemented the algorithm in Python, utilizing the Numpy package for generation of random samples. We observed the case where we calculate the normalization of a target Gaussian based on sampling a different starting Gaussian at various means, and compared these performances to the standard Metropolis-Hastings algorithm. The first stage of analysis involved comparing how the relative number of steps versus sam-

Figure 1: Contour lines of constant error for steps vs samples



ples affected the error for a given effort. To establish these numbers, we ran the algorithm repeatedly for the same number of steps and samples, until it started to converge. We then compared the running mean of these values with an analytical value. By cycling through different numbers of samples and steps we were able to create contours for various errors from the analytical normalization constants.

By looking at approximate gradients along the plot, we are able to estimate a path of lowest error per effort. We see that as effort increases, the path of least error tends to favor lower sampling numbers and higher step sizes, indicating that we may tune our algorithm to emphasize a higher ratio of steps to samples as we proceed. The decision was made to optimize steps to minimize variance.

The next stage in optimization was considering how the variance in our approximate results changed as a function of stepsizes as the homotopy progressed. This required a calculation of said variance. We will consider variance at the m^{th} step $(t_{m-1}+\epsilon)$ assuming that we have already reached step $m-1$ at t_{m-1}. This will be a function of ϵ and the number of samples. Then we will be able to calculate the variance of the entire process as a function of all chosen step sizes and sample numbers.

Let us begin by restating our definitions. We are looking for the variance of

$$Z_m = Z_{m-1}\frac{1}{N_m}\sum_{j=1}^{N_m}\left(\frac{\beta\big(X_{m-1,j}(u)\big)}{\alpha\big(X_{m-1,j}(u)\big)}\right)^{\epsilon}$$

where N_m is the number of samples taken to reach step m and $X_{m-1,j}(u)$, the j^{th} sample from the $(m-1)^{th}$ distribution, are defined as

$$X_{m-1,j}(u) = \frac{\sigma_\alpha\sigma_\beta\,\mathrm{erfinv}[2u-1]}{\sqrt{(1-t_{m-1})\sigma_\beta^2 + t_{m-1}\sigma_\alpha^2}}$$

where u is uniform on $[0,1)$.

If we assume that Z_{m-1} is fixed, we know from properties of variance that

$$\mathrm{Var}\left(Z_{m-1}\frac{1}{N_m}\sum_{j=1}^{N_m}\left(\frac{\beta\big(X_{m-1,j}(u)\big)}{\alpha\big(X_{m-1,j}(u)\big)}\right)^{\epsilon}\right) = Z_{m-1}^2\,\mathrm{Var}\left(\frac{1}{N_m}\sum_{j=1}^{N_m}\left(\frac{\beta\big(X_{m-1,j}(u)\big)}{\alpha\big(X_{m-1,j}(u)\big)}\right)^{\epsilon}\right)$$

Which we can more concisely write as $Z_{m-1}^2\,\mathrm{Var}(Z_m/Z_{m-1})$. We can express this as $Z_{m-1}^2\,\mathrm{Var}(Z_m/Z_{m-1}) = Z_{m-1}^2\,\mathrm{E}[(Z_m/Z_{m-1})^2] - Z_{m-1}^2\,\mathrm{E}[Z_m/Z_{m-1}]^2$, where $\mathrm{E}[Z_m/Z_{m-1}]$ is the expected value of Z_m. Let $f(x)$ be the probability density function for our uniform distribution over $[0,1)$ giving us

$$f(x) = \begin{cases} 1 & \text{if } x \in [0,1) \\ 0 & \text{otherwise} \end{cases}.$$

We can then write our variance as:

$$Z_{m-1}^2 \operatorname{Var}(Z_m/Z_{m-1}) = Z_{m-1}^2 \operatorname{Var}\left(\frac{1}{N_m}\sum_{j=1}^{N_m}\left(\frac{\beta\big(X_{m-1,j}(f(x))\big)}{\alpha\big(X_{m-1,j}(f(x))\big)}\right)^{\epsilon}\right).$$

Since each term in the summand is an identical random variable, we know that the Covariance between them is simply the variance. Thus this can be expanded to the form:

$$\begin{aligned} Z_{m-1}^2 \operatorname{Var}\left(\frac{1}{N_m}\sum_{j=1}^{N_m}\left(\frac{\beta\big(X_{m-1,j}(f(x))\big)}{\alpha\big(X_{m-1,j}(f(x))\big)}\right)^{\epsilon}\right) &= Z_{m-1}^2 \frac{1}{N_m^2}\sum_{j=1}^{N_m^2} \operatorname{Var}\left(\left(\frac{\beta\big(X_{m-1,j}(f(x))\big)}{\alpha\big(X_{m-1,j}(f(x))\big)}\right)^{\epsilon}\right) \\ &= Z_{m-1}^2 \operatorname{Var}\left(\left(\frac{\beta\big(X_{m-1,j}(f(x))\big)}{\alpha\big(X_{m-1,j}(f(x))\big)}\right)^{\epsilon}\right). \end{aligned}$$

Since $\operatorname{Var}(X) = \mathrm{E}[X^2] - \mathrm{E}[X]^2$, we can write this as the difference of integrals:

$$\begin{aligned} Z_{m-1}^2 \operatorname{Var}(Z_m/Z_{m-1}) = Z_{m-1}^2 &\int_{-\infty}^{\infty}\left(\frac{\beta\big(X_{m-1,j}(f(x))\big)}{\alpha\big(X_{m-1,j}(f(x))\big)}\right)^{2\epsilon} dx \\ &- Z_{m-1}^2\left[\int_{-\infty}^{\infty}\left(\frac{\beta\big(X_{m-1,j}(f(x))\big)}{\alpha\big(X_{m-1,j}(f(x))\big)}\right)^{\epsilon} dx\right]^2, \end{aligned}$$

where

$$X_{m-1,j}(f(x)) = \frac{\sigma_\alpha \sigma_\beta \operatorname{erfinv}[2f(x)-1]}{\sqrt{(1-t_{m-1})\sigma_\beta^2 + t_{m-1}\sigma_\alpha^2}}$$

To compute these integrals we can make use of the Law of the Unconscious Statistician, namely that for a random variable X with PDF $f(x)$, the expectation of a function of X can be computed as $\mathrm{E}[g(X)] = \mathrm{E}[g(x)f(x)]$. In

our case this lets us rewrite our integrals as

$$Z_{m-1}^2 \int_{-\infty}^{\infty} \left(\frac{\beta\big(X_{m-1,j}(x)\big)}{\alpha\big(X_{m-1,j}(x)\big)} \right)^{2\epsilon} f(x)dx$$

and

$$Z_{m-1}^2 \left[\int_{-\infty}^{\infty} \left(\frac{\beta\big(X_{m-1,j}(x)\big)}{\alpha\big(X_{m-1,j}(x)\big)} \right)^{\epsilon} f(x)dx \right]^2$$

respectively.

Since $f(x)$ is simply the indicator function over the interval $[0, 1)$, these integrals reduce to

$$Z_{m-1}^2 \int_{0}^{1} \left(\frac{\beta\big(X_{m-1,j}(x)\big)}{\alpha\big(X_{m-1,j}(x)\big)} \right)^{2\epsilon} dx$$

and

$$Z_{m-1}^2 \left[\int_{0}^{1} \left(\frac{\beta\big(X_{m-1,j}(x)\big)}{\alpha\big(X_{m-1,j}(x)\big)} \right)^{\epsilon} dx \right]^2$$

respectively. It now becomes a matter of solving such integrals. Since β and α are both Gaussian shaped distributions with different parameters, the

fraction β/α can be simplified. We see that

$$\begin{aligned}
\frac{\beta(X)}{\alpha(X)} &= \frac{e^{-\frac{(X-\mu_\beta)^2}{2\sigma_\beta^2}}}{e^{-\frac{(X-\mu_\alpha)^2}{2\sigma_\alpha^2}}} \\
&= e^{\frac{(X-\mu_\alpha)^2}{2\sigma_\alpha^2}-\frac{(X-\mu_\beta)^2}{2\sigma_\beta^2}} \\
&= e^{\frac{\sigma_\beta^2(X-\mu_\alpha)^2-\sigma_\alpha^2(X-\mu_\beta)^2}{2\sigma_\alpha^2\sigma_\beta^2}} \\
&= e^{\frac{(\sigma_\beta^2-\sigma_\alpha^2)X^2+(2\mu_\beta\sigma_\alpha^2-2\mu_\alpha\sigma_\beta^2)X+(\mu_\alpha^2\sigma_\beta^2-\mu_\beta^2\sigma_\alpha^2)}{2\sigma_\alpha^2\sigma_\beta^2}}
\end{aligned}$$

Therefore our first integral can be rewritten as:

$$Z_{m-1}^2\int_0^1\left(e^{\frac{(\sigma_\beta^2-\sigma_\alpha^2)X_{m-1,j}^2(x)+(2\mu_\beta\sigma_\alpha^2-2\mu_\alpha\sigma_\beta^2)X_{m-1,j}(x)+(\mu_\alpha^2\sigma_\beta^2-\mu_\beta^2\sigma_\alpha^2)}{2\sigma_\alpha^2\sigma_\beta^2}}\right)^{2\epsilon}dx.$$

Before making a substitution for $X_{m-1,j}(x)$ we can write our coefficients as

$$a=\frac{-\epsilon(\mu_\beta-\mu_\alpha)^2\left[(1-t_{m-1})^2\sigma_\beta^2+t_{m-1}^2\sigma_\alpha^2\right]}{2\left[(1-t_{m-1})\sigma_\beta^2+t_{m-1}\sigma_\alpha^2\right]^2}$$

$$b=\frac{\epsilon(\mu_\alpha-\mu_\beta)\sigma_\beta\sigma_\alpha}{\left[(1-t_{m-1})\sigma_\beta^2+t_{m-1}\sigma_\alpha^2\right]^{3/2}}$$

$$c=\frac{\epsilon(\sigma_\alpha^2-\sigma_\beta^2)}{2\left[(1-t_{m-1})\sigma_\beta^2+t_{m-1}\sigma_\alpha^2\right]}.$$

Making the substitution for $X_{m-1,j}(x)$ yields an integral

$$Z_{m-1}^2 \int_0^1 e^{2a+2b\,\mathrm{erfinv}(2x-1)+2c\,\mathrm{erfinv}^2(2x-1)} dx$$

with the above defined coefficients. We can integrate this via the substitution $y = \mathrm{erfinv}(2x-1)$ noting that

$$dy = \sqrt{\pi} e^{\mathrm{erfinv}(2x-1)^2} dx = \sqrt{\pi} e^{y^2} dx.$$

Therefore our integral becomes

$$\frac{Z_{m-1}^2}{\sqrt{\pi}} \int_{-\infty}^{\infty} \frac{e^{2a+2by+2cy^2}}{e^{y^2}} dy = \frac{Z_{m-1}^2}{\sqrt{\pi}} \int_{-\infty}^{\infty} e^{a+by+(c-1)y^2} dy$$

which when evaluated yields

$$Z_{m-1}^2 \,\mathrm{E}[(Z_m/Z_{m-1})^2] = \frac{Z_{m-1}^2}{\sqrt{1-2c}} e^{2a+\frac{b^2}{1-2c}}$$

The second integral can easily be verified to be

$$\begin{aligned} Z_{m-1}^2 \,\mathrm{E}[Z_m/Z_{m-1}]^2 &= \left(\frac{Z_{m-1}^2}{\sqrt{1-c}} e^{a+\frac{b^2}{4(1-c)}} \right)^2 \\ &= \frac{Z_{m-1}^2 e^{2a+\frac{b^2}{2(1-c)}}}{1-c}. \end{aligned}$$

Finally we can write the variance of Z_m given Z_{m-1} as

$$\begin{aligned} \mathrm{Var}(Z_m) &= \frac{Z_{m-1}^2 e^{\left(2a+\frac{b^2}{1-2c}\right)}}{\sqrt{1-2c}} - \frac{Z_{m-1}^2 e^{\left(2a+\frac{b^2}{2(1-c)}\right)}}{1-c} \\ &= \mathrm{Var}(Z_m) = Z_{m-1}^2 e^{2a} \left(\frac{e^{\left(\frac{b^2}{1-2c}\right)}}{\sqrt{1-2c}} - \frac{e^{\left(\frac{b^2}{2-2c}\right)}}{1-c} \right) \end{aligned}$$

again where:

$$a = \frac{-\epsilon(\mu_\beta - \mu_\alpha)^2\left[(1-t_{m-1})^2\sigma_\beta^2 + t_{m-1}^2\sigma_\alpha^2\right]}{2\left[(1-t_{m-1})\sigma_\beta^2 + t_{m-1}\sigma_\alpha^2\right]^2}$$

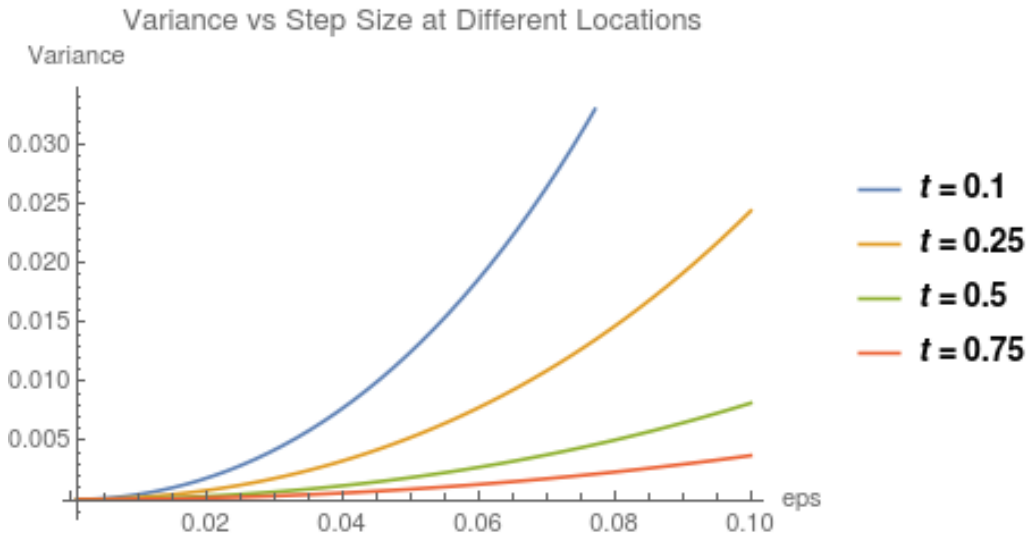
$$b = \frac{\epsilon(\mu_\alpha - \mu_\beta)\sigma_\beta\sigma_\alpha}{\left[(1-t_{m-1})\sigma_\beta^2 + t_{m-1}\sigma_\alpha^2\right]^{3/2}}$$

$$c = \frac{\epsilon(\sigma_\alpha^2 - \sigma_\beta^2)}{2\left[(1-t_{m-1})\sigma_\beta^2 + t_{m-1}\sigma_\alpha^2\right]}.$$

To understand what this calculation tells us we must look at the relationship between the variance and the size of ϵ. Doing so in general becomes difficult due to how convoluted the term for variance is. Therefore we can look at a specific case. We will continue to use this case for numerical testing of the algorithm. Let us define as our target the standard normal $\beta \sim \mathcal{N}(0,1)$, and our starting distribution as $\alpha \sim \mathcal{N}(2,2)$, a wider Gaussian centered two deviations from the target. Our variance then reduces to the form:

$$\begin{aligned}\mathrm{Var}[Z_m] = Z_{m-1}^2 e^{\frac{-4\epsilon\left((1-t)^2+4t^2\right)}{(1+3t)^2}} \cdot \\ \cdot \left(\frac{\sqrt{1+3t}\, e^{\left(\frac{16\epsilon^2}{(1+3t)^3-3(1+3t)^2\epsilon}\right)}}{\sqrt{1+3t-3\epsilon}} - \frac{2(1+3t)e^{\left(\frac{16\epsilon^2}{2(1+3t)^3-3(1+3t)^2\epsilon}\right)}}{2(1+3t)-3\epsilon} \right)\end{aligned}$$

We can plot this as function of ϵ for various t values and observe the following results:



This figure depicts Variance on the vertical axis and step size (ϵ) on the horizontal axis. Each curve is a constant t value, showing the relation between variance and ϵ as the homotopy moves from $t = 0$ to $t = 1$. We see that for lower t values, closer to the beginning of the homotopy process, variance increases faster than at t values closer to the end of the homotopy process. This change in variance as a function of t for a fixed ϵ is similar, exponentially decreasing. This informs us that if we fix the number of step sizes through the homotopy process, we can tune the process such that smaller step sizes are taken towards the beginning, and smaller step sizes are taken towards the end. Due to the inability to solve our equation for variance, experimentation was taken to find the step size values. For the m^{th} step we chose a step size

$$\epsilon_m = \frac{2m}{M(M+1)}$$

where M is the number of step sizes. If we sum these steps we obtain

$$\sum_{m=1}^{M} \frac{2m}{M(M+1)} = \frac{2}{M(M+1)} \frac{M(M+1)}{2} = 1$$

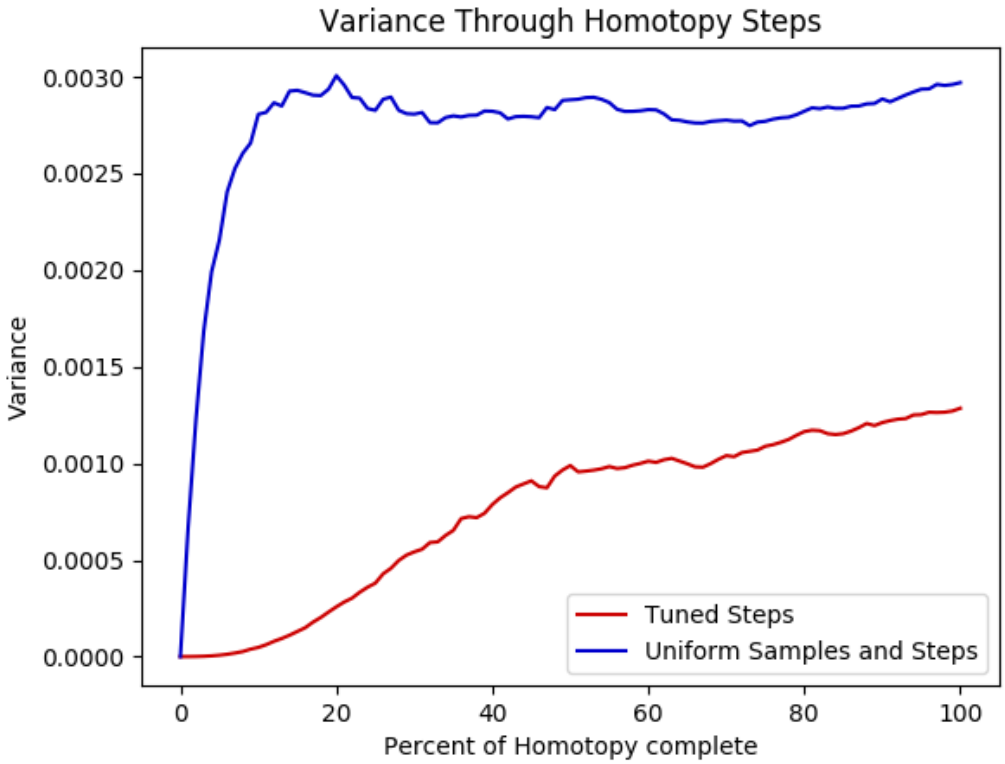


Figure 2: Variance plotted against step sizes

Implementing this optimization leads to the following performance:

We see in this figure variance plotted on the vertical axis verses percent completion of a homotopy procedure. We see in blue the performance of the standard homotopy with uniform samples and step sizes. In green we see a simple bisected sample number optimization, we we increase samples towards the beginning. In red is shown the tuned step sizes according to the above method. Notice that the beginning variance is reduced significantly. The final variance of the untuned process is .002.9693 and of the tuned is .0012848.

We can compare this performance versus the standard Metropolis Hast-

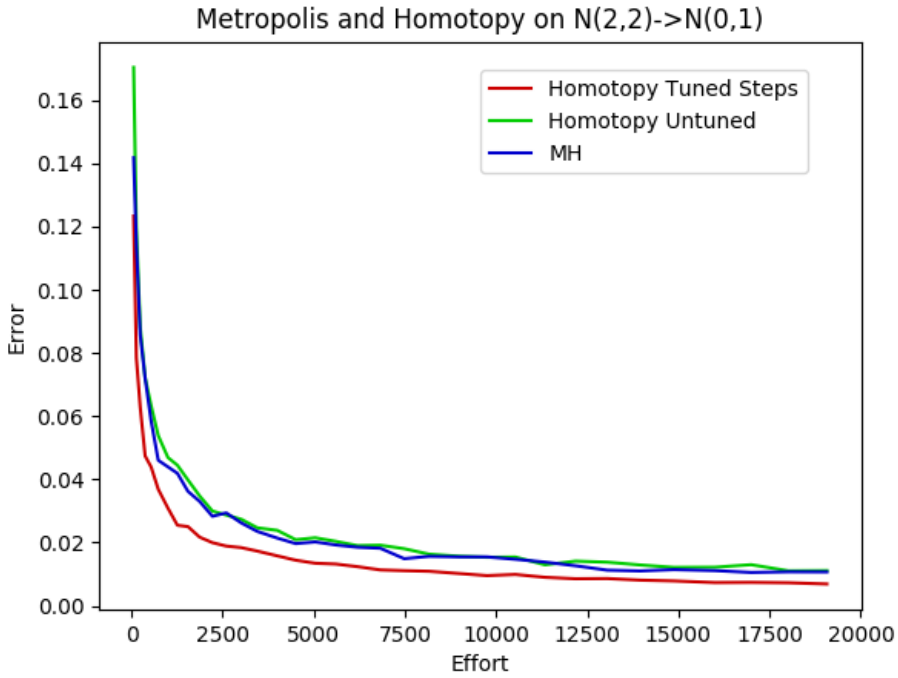


Figure 3: Comparison of tuned homotopy versus Metropolis Hastings

ings algorithm as a benchmark. We ran each algorithm 350 times from a starting distribution of $\alpha \sim \mathcal{N}(2,2)$ to a target distribution of $\beta \sim \mathcal{N}(0,1)$, and took the average error between them and the analytical normalization of β. The results are shown below:

On the horizontal axis is given effort, defined for MH as the number of trials, and for the homotopy procedure as the product of samples per step and number of steps. On the vertical axis is the relative error. We see in blue the benchmark algorithm Metropolis Hastings, performing equivalently with the untuned homotopy procedure in green. In red we see the tuned homotopy procedure outperforming the other two techniques. This adds credence to our hypothesis that the homotopy procedure will outperform a standard method for distributions with distant means.

3 Transdisciplinary Work on the Dungeness Crab

3.1 Motivation for Transdisciplinary Research

In addition to research within the math department at OSU, I participated under NSF grant #1545188: "Risk and uncertainty quantification and communication in marine science and policy" as part of the NSF Research Traineeship Program (NRT) at Oregon State. This chapter is fulfilling the Interdisciplinary Chapter requirements of the Fellowship. As such, I was one of five members of a transdisciplinary team of graduate students from the fields of biology, marine resource management, and fisheries genetics. We collaborated to investigate the relationship between changing ocean conditions and the species of *Cancer magister* (Dungeness crab), and the resulting impacts on the Oregon Coastal fishery and communities.

The project was aimed at investigating coupled human-natural systems that transcended geographic boundaries; merged the public, private and academic worlds; and incorporated information from numerous fields of study. As such, it was a candidate for developing research beyond a singular discipline or even the simple collaboration between them. To effectively study these complicated relationships through social systems, crab ecology, ocean conditions, and data statistics, we needed to employ transdisciplinary methods. We define transdisciplinarity as "transcending disciplinary world views" [8], using frameworks for research to bridge academic disciplines and the public and private sectors. As stated by Bonebreak et al., "ecological, conservation and social research on species redistribution can best be achieved by working across disciplinary boundaries to develop and implement solutions to climate change challenges" [2]. There have been successful applications of these transdisciplinary methods in similar marine contexts such as Benham's study of the Great Barrier Reef [1]. By employing these transdisciplinary

methods to our own project we can combine our disciplinary expertise in such a manner to tackle complex problems involving the environment, crustaceans, and people.

The first step in our transdisciplinary process was in our collaboration. Every member of the group participated in the process of creating questions and hypotheses, in collecting data sources and in their organization and analyses, and in the interpretations and conclusions we ultimately drew. The variety of knowledge and experience each of us brought to the table mollified many of difficulties we had, and allowed all of us to contribute in a balanced manner. In the end, each of us expanded our knowledge of the Dungeness crab fishery, ecology, and ecosystems. The general work we produced, which is outlined in detail in our Transdisciplinary Report, consisted of studying the relation between ocean conditions and the geographic distribution of Dungeness crab catch along the West Coast, looking at reliance of coastal communities on the Dungeness crab in relation to their socio-economic vulnerability, and looking at how environmental conditions could impact the larval stages of Dungeness crab, and the implications for the future catch. This last objective is presented below as a case study in how mathematics can be applied to studying a transdisciplinary project, the challenges that occurred, and the differences between seeing a problem from a mathematical lens verses a transdisciplinary perspective.

3.2 Biological Background

Dungeness crab is among the most economically important West Coast fisheries [9]. However, it has presented a constantly changing distribution along the coast, leading to varied landings both year-to-year and spatially between ports. We were seeking to therefore analyze and reduce some of this uncertainty as we looked for the relationship between Dungeness crab distribution and climate. In particular, understanding this complicated relationship may be important in the face of changing oceanic conditions and climate change

[6]. Although Dungeness crab can be legally caught only in the adult stages, it is in the younger life stages where the crabs are at most influence from environmental conditions [13][11].

The species of Dungeness crab is distributed as far south as Santa Barbara California to the Pribolof Islands of Alaska. They are organized in three ecosystems, the California Current off the West Coast of the contiguous United States, the Salish Sea between Washington state and British Colombia, and in the Gulf of Alaska [17]. For the purpose of this investigation we group the Dungeness crab into a southern population in the California Current ecosystem, and a northern population encompassing the Gulf of Alaska and Salish Sea ecosystems. Dungeness crab may begin their life cycles in any of these ecosystems and experience different life-cycle timings depending on the latitude of their origin. According to Rasmuson in [9], the southern population of Dungeness crab generally lay their eggs within a given Fall or Winter leading to hatching within the next December to January. Following this hatching, the crabs spend 3-4 months in a pelagic larval stage being transported by ocean currents, and finally settle before July of that year. The movement of larval stages ultimately determines where where the Dungeness crab mature and reach adulthood. However, the larvae are moved entirely based upon oceanic conditions. The dispersal of the larvae is jointly determined by current strength [11], the timing of the spring transition [13], and upwelling [12]. The northern population of Dungeness crab hatch at a later time however, and may not settle in the coast until after the beginning of August, taking as long as October. Since we observe larvae in Coos Bay through October, it is theorized that the late season larvae, those found after August 1st, have been transported from a north population. To confirm this hypothesis, genetic testing was performed by group member Elizabeth Lee, and it was found that there was a genetic difference between early season and late season megalopae, indicating that these late season megalopae were not local.

As stated above, the environmental forces that drive this megalopae transport are Pacific Decadal Oscillation (PDO) [12], Upwelling[11], and the timing of the Spring Transition [13]. The way that these environmental conditions impact megalopae transport is primarily through their relation with the California and Davidson currents. Larvae are transported offshore and then northward during their winter stage by the Davidson current, but then after the Spring Transition date they are returned southwards along the California current. It is during this time that the northern crab populations may be transported southward [11]. The recruitment patterns that we see are heavily influenced by the strength of this California current [5]. The megalopae are brought inshore by upwelling events that occur between April and as late as September [4][9], and a larger amount of upwelling should be linked to more megalopae presence in shore. Finally, PDO is a calculated index that describes changes in temperature anomalies in the mid-latitude Pacific Ocean. It describes a pattern of change in ocean temperature gradients where a positive index represents a warmer eastern ocean and colder west ocean and a negative index represents the opposite pattern. A year with a negative average PDO index is associated with a stronger California current, which is expected to increase southward transport of late-season megalopae. A year with positive average PDO index is associated with a weaker California current and therefore expected to transport fewer megalopae to Coos Bay during the late-season.

To explore the connection between Dungeness crab late season megalopae abundance and oceanic conditions, the choice was made to investigate several predictive models: a multinomial logistic regression model (MNLR), an Auto-regressive Integrated Moving Average (ARIMA) model with seasonality and explanatory environmental variables, and a Maximum Entropy model (MaxEnt). The choice of these three models was to cover a broad predictive range for varying scalings. The MNLR model was used to provide categorization of *low (L)*, *medium (M)*, and *high (H)* presence of late-season megalopae

as informed by previous years and environmental conditions for general long term predicting. The ARIMA and MaxEnt models were designed to produce near future seasonal predictions of megalopae presence distributions based on pertinent environmental data and the trendline through previous years.

Unfortunately, success was not obtained with the ARIMA and MaxEnt models. There was a lack of sufficient geographic data to employ the MaxEnt modeling. The MaxEnt framework we attempted to utilize was designed to incorporate geographically distributed data, and trying to substitute temporally distributed data did not properly function. The ARIMA model looked for reoccurring trend lines and seasonality, but the high variability due to one daily sample at one geographic location led to indistinct predictions for late-season presence. Either more samples were required for each data point to reinforce relations with climate data, or more sample years were necessary to extend the trend-line. In particular, with the years 2002-2005 missing from the dataset, the jump in years lead to difficulty establishing a year-to-year trend-line that would mirror PDO.

We ultimately chose to focus on the MNLR model, as its simplicity catered to the limited data. We deemed it better to model the situation more accurately in a simple manner than provide a detailed but poorly performing forecast. This led to the finalization of a MNLR model to answer the simple question of whether or not there will be a late season pulse based on environmental conditions.

The multinomial logistic regression technique is an extension of a binary logistic regression, which seeks to use maximum likelihood estimation to place dependent variables into two categories based on independent variables. We explored both binary and multinomial cases, so we will develop the theory behind both. We begin with a logistic function of a variable t given by

$$f(t) = \frac{1}{1 + e^{-t}}$$

Assuming that t linearly depends upon n explanatory variables x_i, we can

write t as $t = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$. Writing our coefficients and variables as the n dimensional vectors $\boldsymbol{\beta} = [b_1 \cdots b_n]$ and $\mathbf{x} = [x_1 \cdots x_n]$, we then have logistic function

$$p(\mathbf{x}) = \frac{1}{1 + e^{(-\beta_0 - \boldsymbol{\beta} \cdot \mathbf{x})}} \tag{2}$$

We can interpret $p(\mathbf{x})$ as the probability that the set of explanatory variables $\mathbf{x}$ informs a placement of a dependent variable in the first category. In our binary case, let this first category be denoted L for low and the second category H for high. Accordingly, $1 - p(\mathbf{x})$ is the probability of a dependent variable falling in the category H. Solving for the inverse of (2) yields the logit equation:

$$\ln\left(\frac{p(\mathbf{x})}{1 - p(\mathbf{x})}\right) = \beta_0 + \boldsymbol{\beta} \cdot \mathbf{x}, \tag{3}$$

which tells us the log-odds of a dependent variable being assigned to the first category over the second.

For the standard multinomial case, where we will have M categories $C_1, \ldots, C_M$, we need to construct $M - 1$ logit functions to express the log-odds between every possible category. We therefore require $M - 1$ vectors $\boldsymbol{\beta}^1, \ldots \boldsymbol{\beta}^{M-1}$ each composed of n terms for each explanatory variable i.e. $\boldsymbol{\beta}^i = [\beta_1^i \cdots \beta_n^i]$. To produce the logit functions, we can consider one category a *pivot* category and produce $M - 1$ binary logit functions for the other categories. Allowing C_M to be this pivot, we obtain $M - 1$ equation for i in $\{1, 2, \ldots, M - 1\}$ of the form

$$\ln\left(\frac{p_i(\mathbf{x})}{p_M(\mathbf{x})}\right) = \beta_0^i + \boldsymbol{\beta}^i \cdot \mathbf{x}, \tag{4}$$

where $p_i(\mathbf{x})$ is the probability that a dependent variable falls within the ith category C_i, given independent variable $\mathbf{x}$.

In our case, our categories obey an ordinal relationship, where each increasing category represents a subsequent interval of values corresponding to the ratio of late season megalopae. In other words for values $a < b < c < d$,

we have $L = [a, b]$, $M = (b, c]$, and $H = (c, d]$. This allows us to produce new functions with probabilities P_b, P_c, and P_d where $P_i(\mathbf{x})$ is the probability that our dependent variable y is less than i, meaning it is contained in one of the categories lower than the border point i. Due to this, we obtain new logit equations for our ordinal categorization:

$$\ln\left(\frac{P_i(\mathbf{x})}{1 - P_i(\mathbf{x})}\right) = \beta_0^i + \boldsymbol{\beta} \cdot \mathbf{x}, \tag{5}$$

which is equivalent to giving the logodds of a dependent variable being less than i versus greater than i. For this unique ordinal case, since each probabilistic equation overlaps with those less than it, we can use the same β parameter for each explanatory variable, although each equation with have different intercepts.

The ultimate aim of the model is to use training data to produce β parameters. These can be used to predict the probabilities that a dependent variable y falls within a category given known explanatory variables $\mathbf{x}$. Our training data is in the form: X_0 and $\mathbf{y}_0$, where X_0 is a k by n array of n known explanatory variables for k years, and $\mathbf{y}_0$ is a k-vector of dependent variables for the respective years. In our model, the explanatory variables were ocean conditions, and our dependent variable was late-season presence. We can categorize each year based on the presence value of entries in $\mathbf{y}_0$.

There are multiple ways to generate our predictive coefficients $\boldsymbol{\beta}^i$, but most rely on maximizing log-likelihood in a process called Maximum Likelihood Estimation. For the two category case, we have defined the probability that y belongs in category L given $\mathbf{x}$ as $p(\mathbf{x})$ in (2). If we assume that $\boldsymbol{\beta}$ is also a variable in this case, we can write this probability as $p(\mathbf{x}, \boldsymbol{\beta})$. If we consider a particular trial year, we obtain a Bernoulli distribution with probability $p(\mathbf{x}, \boldsymbol{\beta})$ that y is in category L. Defining the indicator function

$$\chi_L(y) = \begin{cases} 1 & y \in L \\ 0 & y \in H \end{cases} \tag{6}$$

we can write the likelihood of a particular trial j as

$$\mathcal{L}_j(\boldsymbol{\beta}) = p(\mathbf{x}_j, \boldsymbol{\beta})^{\chi_L(y_j)}\big(1 - p(\mathbf{x}_j, \boldsymbol{\beta})\big)^{(1-\chi_L(y_j))} \tag{7}$$

where $\mathbf{x}_j$ are the explanatory variables from the jth year and y_j is the dependent variable from the jth year. To compute the likelihood using all the training years we take the product of each equation (7) for each j:

$$\mathcal{L}(\boldsymbol{\beta}) = \prod_{j=1}^{k} \mathcal{L}_j(\boldsymbol{\beta}) = \prod_{j=1}^{k} p(\mathbf{x}_j, \boldsymbol{\beta})^{\chi_L(y_j)}\big(1 - p(\mathbf{x}_j, \beta)\big)^{(1-\chi_L(y_j))}. \tag{8}$$

It then remains to maximize $\mathcal{L}$ over the vector $\boldsymbol{\beta}$.

3.3 Data Selection and Treatment

Daily counts of Dungeness crab larvae (megalopae), were obtained via light traps in Coos Bay within the years of 1997 to 2001 and from 2006 to 2017 (provided by Dr. Alan Shanks of the University of Oregon). Additional daily counts were obtained for the years of 2016 and 2017 in Yaquina Bay (provided by Elizabeth Lee of Oregon State University). The data is in the

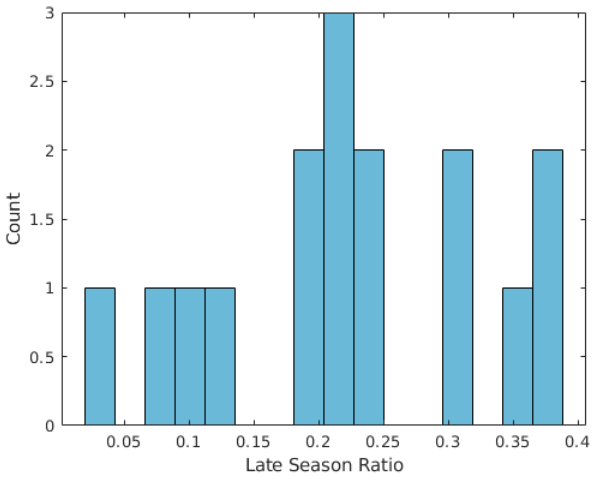
year	1998	1999	2000	2001	2006	2007	2008	2009
LSP	0.0348	0.2382	0.2263	0.1997	0.3844	0.3571	0.2049	0.3040
year	2010	2011	2012	2013	2014	2015	2016	2017
LSP	0.0661	0.3758	0.3155	0.2292	0.2092	0.0978	0.1330	0.1984

Table 1: Late-season presence ratios per year

form of daily count numbers for each year ranging from approximately the first of April through October or November. I chose to use data from the first of April through the second of October for each year, as this captured the majority of all counts. The data was natural log-transformed to reduce

skewing. The division between late and early season megalopae was defined as August 1st, which is 120 days into the gathered data. To produce its relative abundance, we simply took the sum of all counts after this date for a given year and divided over the total count for the entire year, calculating the relative abundance of late-season megalopae each year. Let us denote this as LSP for Late-Season Proportion.

Figure 4: Late Season Megalopae Proportions

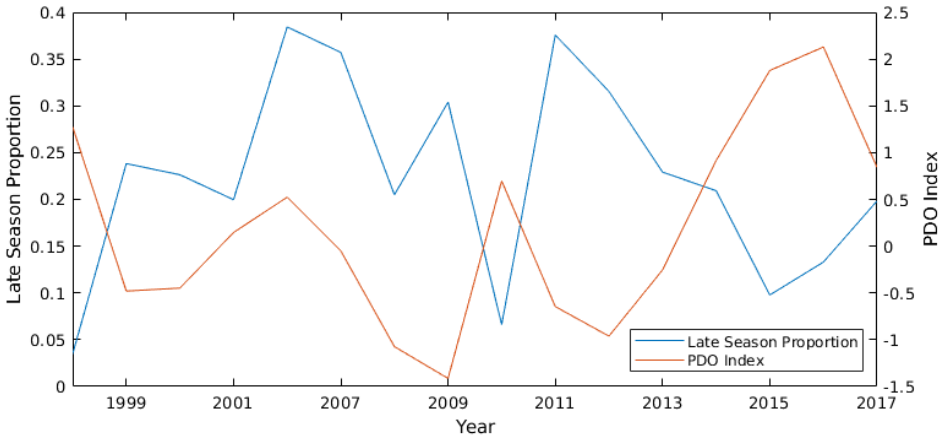


Using these ratios we produced a simple histogram (seen right) to locate possible breakages with which to divide the data into categories. Based on simple analysis of the figure, we postulated that the division between low and medium levels of megalopae proportion would fall within the interval $(.13, .17)$, and the division between medium and high late season proportion would fall in the interval $(.25, .29)$. The exact values of these devisions were to be determined by optimization over these intervals. We began with by categorizing our low presence interval $L = [0, a]$, our medium presence interval $M = (a, b]$ and our high presence interval $H = (b, .4]$, where a is the division between .13 and .17, and b the devision between .25 and .29.

The next task was to select the environmental data to use as explanatory variables. We know that there is correlation between PDO and megalopae count from [12]. Our hypothesis for late season megalopae states that in low PDO (cool) years, we should see a stronger California current, and therefore a greater transport of the current driven northern megalopae. In high PDO (warm) years, we should see a weaker California current and a drop in late

season megalopae presence. We computed a correlation coefficient between yearly average PDO and late season proportion of -0.6194, indicating a moderate negative correlation, reinforcing the hypothesis that a year of negative average PDO would lead to larger southern transport of northern megalopae due to a stronger California Current.

Figure 5: Late Season Proportion Plotted with PDO



Other explanatory variables considered for inclusion in the model were yearly average El Ñino Southern Oscillation (ENSO) index values, both Southern Oscillation Index (SOI) and Multivariate ENSO Index (MEI). We found that late-season megalopae presence correlated with SOI with coefficient 0.6803 and negatively with MEI with coefficient -0.7084. This may be explained by how a positive PDO phase can produce climate patterns very similar to El Niño, accounting for the similarity between these climate variables' correlations. We also considered offshore temperatures taken outside of Coos Bay negatively correlated with coefficient $r = -0.6542$ when considered against late-season megalopae presence. Since PDO is a function of temperature variation in the Pacific northern hemisphere, this is unsurprising to see similar correlations. In fact, our values for PDO and temperature themselves had a correlation coefficient of .8755. Therefore we elected to use only one of these environmental variables as an explanatory variable. Other environmen-

tal conditions that we tested included upwelling (UW) and Spring Transition (STI) dates. Both proved to be uncorrelated, with correlation coefficients of -0.1212 and 0.0999 respectively.

Variable	STI	UW	PDO	NPGO	SOI	MEI	SST
Correlation	.0999	-.1212	-.06194	-.0899	.6803	-.7084	-.6452

Table 2: Correlations of environmental variables

The data finally chosen to be run against the model were PDO indices and ENSO MEI.

3.4 Constructing the Model

The model was constructed in MATLAB R2017b to be run as a script. In setting up the model, optimization was required for several parameters. Foremost, determining where to divide categories within our postulated intervals. We iterated the model along the divisions until we produced the lowest p-values between our variables and categorizing probabilities. We discovered that the best placing locations for edges were at $a = 1.4$ and $b = 2.6$ yielding categories $L = [0, 1.4]$, $M = (1.4, 1.6]$ and $H = (1.6, .4]$. These results were obtained by running the model on various training sets and comparing the resulting p-values and confusion matrices. The MNLR framework was run using the two explanatory variables PDO and ENSO MEI in two different models and produced the following p-value results:

Variable	PDO	ENSO MEI
p value med/high	.1138	.2010
p value low/med	0.0099	.0337

Table 3: p-values from explanatory variables

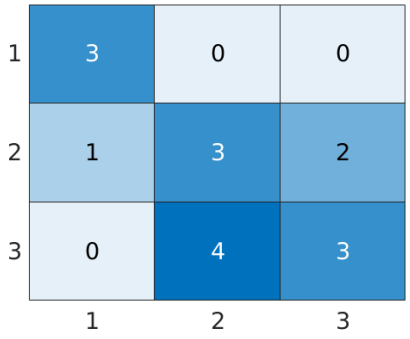


Figure 6: Heat map for one year predictions on three categories

Testing with single year forecasting over all years yielded the confusion matrix on the right. These results show accurate placement in the low category, but clear issues with categorizations to medium and high. Combined with the non-significant p-values for these category distinctions, this shows that the model could not clearly distinguish between high and medium presence of late-season megalopae. This might be due to lack of data, or due to less clear influence from environmental data at that level. Therefore the model was rerun with only two categories by merging medium and high, resulting in the following improved heatmap shown left. In this case, we the expected merging of the lower right hand boxes, leading to almost perfect categorization. However, the informational output of our model has been reduced to a binary result, the least amount of non-trivial information. This model only answers the question of whether or not there will be a late-season presence, giving no greater detail to the projected magnitude. However, with the current limited data available, this may be the most that can be done and reliably dependent upon.

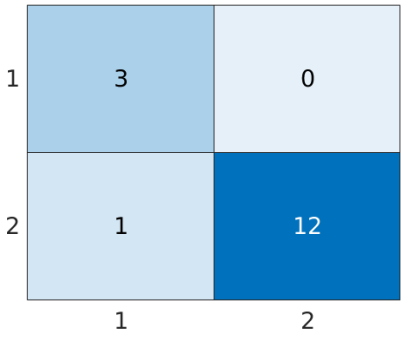


Figure 7: Heat map for one year predictions on two categories

3.5 Reflection

The NRT program provided me with a novel insight into the world of transdisciplinary research. An expansive undertaking such as studying the complex interactions between climate, Dungeness crab and society requires not only research from various fields, but the integration of those fields. As alluded to above, an example of this is in data gathering. We learned that in order to be fully prepared to mathematically model a particular biological phenomenon, the mathematicians should be involved in the data gathering process. We learned though this NRT experiment that transdisciplinary techniques can be very powerful tools for studying and learning about a subject, but may not cater as well towards individual disciplinary topics. For example, in our project we spent considerable time studying the statistical correlations between Dungeness crab catch data and environmental variables, we looked at how coastal community's vulnerabilities would make them susceptible to fluctuations in this catch, and of course we looked at modeling megalopae late-season presence. This work however, did not necessarily lend itself to as much mathematical development as it did towards data and statistical analysis.

In contrast to the tempered mathematical development I experienced during the NRT program, I learned an incredible amount about other disciplines, fully embracing the transdisciplinary philosophy. In particular, I studied oceanography to better understand the systems that drive crab larval transport. I researched the life stages of the Dungeness crab and how they interact with their environment. I was introduced to the west-coast fishery, an industry entirely foreign to me, and gained more of an appreciation of the processes, economics, and social frameworks within these coastal communities. I learned about the quantitative and qualitative methods of studying risk, and applied notions of vulnerability and resource reliance to the coastal communities. Finally, I got to see first hand how data can be used or may

prove less fruitful, giving me a glimpse beyond the theoretical mathematics I had been previously experienced with. I believe that this project has led to me growing as a graduate student and researcher, and helped me understand what role mathematics can play in relation to other sciences. I realized that my role in the group transcended my role as a mathematician, and included a much larger spectrum of work and research.

List of Figures

List of Tables

References

[1] C. F. Benham and K. A. Daniell, *Putting transdisciplinary research into practice: A participatory approach to understanding change in coastal social-ecological systems*, Ocean & Coastal Management, 128 (2016), pp. 29–39.

[2] T. C. Bonebrake, C. J. Brown, J. D. Bell, J. L. Blanchard, A. Chauvenet, C. Champion, I.-C. Chen, T. D. Clark, R. K. Colwell, F. Danielsen, et al., *Managing consequences of climate-driven species redistribution requires integration of ecology, conservation and social science*, Biological Reviews, 93 (2018), pp. 284–305.

[3] C. e. a. Geyer, *Handbook of Markov Chain Monte Carlo*, Chapman and Hall/CRC, 2013.

[4] G. S. Jamieson and A. C. Phillips, *Occurrence of cancer crab(c. magister and c. oregonensis) megalopae off the west coast of vancouver island, british columbia.*, Fishery Bulletin, 86 (1988), pp. 525–542.

[5] R. G. Lough, *Dynamics of crab larvae (anomura, brachyura) off the central oregon coast, 1969-1971*, (1975).

[6] K. N. Marshall, I. C. Kaplan, E. E. Hodgson, A. Hermann, D. S. Busch, P. McElhany, T. E. Essington, C. J. Harvey, and E. A. Fulton, *Risks of ocean acidification in the california current food web and fisheries: ecosystem model projections*, Global Change Biology, 23, pp. 1525–1539.

[7] N. e. a. Metropolis, *Equation of state calculations by fast computing machines*, J. Chem. Phys., 21 (1953), p. 1087.

[8] M. O'Rourke, S. Crowley, S. D. Eigenbrode, and J. Wulfhorst, *Enhancing communication & collaboration in interdisciplinary research*, Sage Publications, 2013.

[9] L. Rasmuson, *The biology, ecology, and fishery of the dungeness crab, cancer magister*, Advances in Marine Biology, 65 (2013), pp. 95–148.

[10] C. Robert and G. Casella, *Monte Carlo Statistical Methods*, Springer-Verlag, 2004.

[11] A. Shanks, G. Roegner, and J. Miller, *Predicting the future commercial catch of dungeness crabs.*, (2010).

[12] A. L. Shanks, *Atmospheric forcing drives recruitment variation in the dungeness crab (cancer magister), revisited*, Fisheries Oceanography, 22 (2013), pp. 263–272.

[13] A. L. Shanks and G. C. Roegner, *Recruitment limitation in dungeness crab populations is driven by variation in atmospheric forcing*, Ecology, 88 (2007), pp. 1726–1737.

[14] A. N. Shiryae, *Probability*, Springer-Verlag, 1996.

[15] J. Smagorjnsky, *The beginnings of numerical weather prediction and general circulation modeling: Early recollections*, in Theory of Climate, B. Saltzman, ed., vol. 25 of Advances in Geophysics, Elsevier, 1983, pp. 3 – 37.

[16] E. Waymire and B. R., *A Basic Course in Probability*, Springer Science+Business Media, 2007.

[17] P. W. Wild and R. N. Tasto, *Life history, environment, and mariculture studies of the Dungeness crab, Cancer magister, with emphasis on the central California fishery resource*, State of California. The Resources Agency. Department of Fish and Game, 1983.

[18] W. Woess, *Denumerable Markov Chains: Generating Functions, Boundary Theory, Random Walks on Trees*, EMS textbooks in mathematics, European Mathematical Society, 2009.