Research Statement

Brett Geiger (geiger12@math.uh.edu)

1. Overview

My main research interests are in the fields of applied probability and mathematical biology with a focus on applications of large deviations to dynamical systems influenced by small random fluctuations (noise). I also have a secondary interest in nonequilibrium dynamics. Large deviations principles exist in a wide variety of probabilistic settings, including stochastic differential equations (SDEs) and discrete and continuous Markov processes. A fundamental question that arises is whether these principles can be applied to a probabilistic model in an *explicit and computationally efficient* way. In doing so, one obtains *concrete* information for the system; namely, the computation of rare events of interest.

Nonequilibrium dynamics refers to situations in which the dynamical model varies in time. The study of nonequilibrium dynamics is motivated by our desire to understand the dynamics of systems that evolve in time-varying environments. Examples of such systems include Sinai billiards with moving scatterers and open systems (systems with holes) in which mass is allowed to escape. The field of random dynamical systems treats the case in which the statistical properties of the variability of the model are assumed to be known. In contrast, no a priori knowledge of the statistical properties of the evolution of the dynamical model are assumed in the case of nonequilibrium dynamics. Therefore, the study of the statistical properties of a time-dependent system becomes a natural approach to take in order to understand the system over time. One property of particular interest is the behavior of the distances between pairs of initial distributions as distributions evolve.

During my graduate tenure, I have studied large deviations for Gaussian diffusions with delay, which can be applied locally to nonlinear SDEs. Funded through a research grant, I have studied applications of large deviations to genetic evolution of bacterial populations, namely Escherichia Coli (E. Coli), using results from numerous biological scientists and labs, including Professor Tim Cooper's lab at the University of Houston. I have also studied the statistical properties of a class of nonequilibrium open systems with high-dimensional phase spaces. My research has been conducted under the guidance of Professor Robert Azencott, Professor William Ott, and Professor Ilya Timofeyev.

2. Large Deviations

Small noise can dramatically affect underlying deterministic dynamics, such as transforming stable states into metastable states. This leads to positive probabilities of rare events of high interest, such as excursions away from nominally stable states or transitions between metastable states. These rare events play important functional roles in a wide range of settings, including genetics, biochemistry, and systems with multiple timescales [4, 5, 9] . Observing rare events from direct simulations or experimentation is often infeasible and computationally costly as one may need a large number of realizations on a long timescale in order to observe even one occurrence of said event. Large deviations theory and techniques allow one to quantify the probabilities of rare events as well as the most likely pathways by which the system achieves rare events without the use of direct simulation and experimentation. Large deviations is a rich field, mixing numerous theoretical aspects with computational techniques. On one hand, one must first either establish a large deviations principle for the model in question or rely on previously-proven principles which apply to the model. Establishing large deviations principles often requires the use of functional analysis, probability theory, topology, optimization, and analysis. On the other hand, applying these principles is a highly nontrivial computational variational problem and often involves simulation of Hamilton-Jacobi equations, which are computationally costly even in moderately high spatial dimensions.

2.1. Gaussian Diffusions with Delay.

Consider a family of random processes $X^\varepsilon(t)$ indexed by a small parameter $\varepsilon > 0$ and driven by the following generic small-noise SDE with drift b, diffusion $\varepsilon\sigma$, and no delays:

$$dX^\varepsilon(t) = b(X^\varepsilon(t))\,\mathrm{d}t + \varepsilon\sigma(X^\varepsilon(t))\,\mathrm{d}W(t).$$

Large deviations theory for SDEs of this form was developed by Freidlin and Wentzell [10]. Friedlin-Wentzell theory estimates the probability that the process $X^\varepsilon(t)$ lies within a small tube around any given continuous

path $\psi \in C([0,T],\mathbb{R}^d)$ in terms of the *action functional* $S_T(\psi)$ of ψ :

$$\mathbb{P}_x\left\{\sup_{0\le t\le T}|X^\varepsilon(t)-\psi(t)|\le\delta\right\}\approx\exp\left(-\varepsilon^{-1}S_T(\psi)\right).$$

Here, $\mathbb{P}_x$ denotes the probability conditioned on $X^\varepsilon(0)=x$, and we assume that $\psi(0)=x$.

For fixed time T and points p,q in the state space, the path $\hat{\psi}$ that minimizes $S_T(\psi)$ under the constraints $\psi(0)=p$ and $\psi(T)=q$ is the most likely transition path starting at p and reaching q at time T. A second minimization over T provides the most likely transition path from p to q and the energy $V(p,q)$ associated with this optimal path. Often called the quasi-potential, V is central to the quantification of large deviations on long timescales [10].

Delay plays a crucial role in many biochemical processes. For instance, delay is an inherent component in the dynamics of genetic regulatory networks due to protein production. The chemical Langevin equations that model these networks necessarily have delay in both the drift and diffusion [11], yielding nonlinear stochastic differential equations with delay (delay SDEs). The magnitude of the time delay may be significant so that one cannot view a system with delay as a small perturbation of a system with no delay. Therefore, the necessary presence of delay creates an infinite-dimensional nonlinear system where quantifying events and characteristics of interest (e.g. most likely escape routes from metastable states) becomes particularly difficult.

For nonlinear delay SDEs, it is possible to compute a linear noise approximation [6] that is valid in a neighborhood of a given metastable state, which yields a Gaussian diffusion with delay. Consequently, we focused on Gaussian diffusions with delay for which we have rigorously developed and implemented a *fully explicit* large deviations framework enabling fast numerical computation of optimal transition paths and quasi-potentials. Our methodology *does not require* the numerical solution of Hamilton-Jacobi equations, a significant positive that enables efficient computations in high spatial dimensions. Furthermore, our framework allows analysis of nonlinear delay SDEs locally, which is advantageous since direct analysis of nonlinear models using large deviations principles is quite challenging. We demonstrated this by computing optimal escape trajectories from a small neighborhood of a metastable state of the co-repressive toggle switch [11], a bistable genetic circuit driven by a nonlinear Langevin equation. The co-repressive toggle switch also highlights the necessity of the delay term in the general framework that follows as delay in protein production can significantly affect the dynamics of gene regulatory networks [11].

We thus centered our study on the Itô delay SDE

$$\begin{cases}\mathrm{d}X_t^\varepsilon=(a+BX_t^\varepsilon+CX_{t-\tau}^\varepsilon)\,\mathrm{d}t+\varepsilon\Sigma\,\mathrm{d}W_t,\\ X_t^\varepsilon=\gamma(t)\text{ for }t\in[-\tau,0].\end{cases}\tag{1}$$

Here $X_t^\varepsilon\in\mathbb{R}^d$, t denotes time, $\tau\geqslant 0$ is the delay, $a\in\mathbb{R}^d$, B and C are real $d\times d$ matrices, W_t denotes standard n-dimensional Brownian motion, $\Sigma\in\mathbb{R}^{d\times n}$ denotes the diffusion matrix, and $\varepsilon>0$ is a small noise parameter. The initial history of the process is given by the Lipschitz continuous curve $\gamma:[-\tau,0]\to\mathbb{R}^d$. We worked with fixed delay to simplify the presentation – all of our results apply just as well to multiple delays and to random delay distributed over a finite time interval.

For the stochastic process X_t verifying (1), what is the most likely transition path from p to q given that $X_0=p$ and $X_T=q$, and what is the energy associated to this path? To address these questions, we first proved that the stochastic process X_t verifying (1) is a Gaussian process. Large deviations results for Gaussian processes are well-developed [2]; however, these results exist for centered (mean zero) processes. Therefore, we defined a process Z_t by $X_t=m(t)+\varepsilon Z_t$ where $m(t)=\mathbb{E}[X_t]$. We then have that Z_t is a centered Gaussian diffusion with delay so that large deviations results can be applied to Z_t. Since X_t is a deterministic shift of Z_t, probabilistic estimates for Z_t immediately translate to probabilistic estimates for X_t. The action functional in this setting is given by the Cramer "energy" $\lambda(f)$ of a continuous path $f\in C([0,T])$. In general, the Cramer energy λ associated with a probability μ is the Legendre dual of the log-Laplace transform of μ. This energy functional takes a more explicit form in the case of (centered) Gaussian processes [2]. Thus, finding the most likely transition path and associated energy involved a minimization of this energy functional over all paths realizing the desired transition. Using Lagrange multipliers for the minimization, we calculated explicit formulas for the most likely transition path and associated energy *independent of the small noise*

parameter ε, which are given by

$$h^T(s) = m(s) + \rho(s,T)[\rho(T,T)^{-1}(q - m(T))] \qquad (0 \leqslant s \leqslant T)$$
$$\lambda(h^T) = \frac{1}{2}[\rho(T,T)^{-1}(q - m(T))] \cdot [q - m(T)],$$

where $\rho(s,t)$ are the covariance matrices of the process Z_t. Furthermore, the mean $m(t)$ and covariance $\rho(s,t)$ both verify linear first-order delay differential equations, enabling fast computation of optimal paths and associated energies. Finally, a second minimization over T of the energy $\lambda(h^T)$ provides the most likely time the transition occurred as well as the quasi-potential V.

2.2. **Bacterial Evolutionary Dynamics.** Populations of *bacteria* or *viruses* exhibit strong genetic adaptivity through emergence and fixation of beneficial mutations. Predictive studies of these evolutions have strong potential impact on questions such as bacterial resistance to antibiotics or emergence of viral strains transferable from animal to humans. However, existing stochastic dynamic models for these populations still lack applicable algorithmic tools to quantify genetic evolution trajectories in the fitness landscape. Current analysis tends to rely on intensive simulations, which are not efficient to evaluate key rare events such as specific chains of beneficial genotype fixations. Therefore, we developed an applicable large deviations framework to solve difficult numerical and mathematical questions of high biological interest, such as computing the most likely evolutionary path linking two given population states in the fitness landscape and evaluating transition probabilities between successive genotype fixations.

Based on experiments used to model the main features of random bacterial evolutions, the following class of Markov chains has often been utilized [7, 12, 13, 15]. These Markov chains are called "locked box" models here. The finite set of distinct genotypes is denoted $\Gamma = \{1, 2, \ldots, g\}$. Cells of genotype j (called *j-cells* here) have rate of exponential growth $f_j > 0$, called the *fitness* of genotype j. Genetic evolution is modeled as a sequence of cycles called "daily" cycles as is the case for many experimental contexts. The n^{th} cycle starts with a population pop_n of fixed large size N and involves three successive steps to generate pop_{n+1}:

(1) **Deterministic growth with no mutations:** The size of each j-cell colony in pop_n is multiplied by the *growth factor* $F_j = \exp(\tau f_j)$ where τ is a fixed growth duration parameter.

(2) **Independent random mutations:** For any two distinct genotypes (j,k) a random number $R_{j,k}$ of j-cells mutate into k-cells. Let $siz_n(j)$ be the j-cells colony size after growth in step 1. Then, $R_{j,k}$ has a Poisson distribution with mean $siz_n(j)M_{j,k}$ where the mutation rates $M_{j,k}$ are fixed and very small.

(3) **Random Selection:** After step 2, the population has reached a saturation size N_{sat} much larger than N, and one extracts a random sample of fixed size N, which constitutes pop_{n+1}. These daily random selections have a strong *bottleneck effect* [12, 13, 15] on the emergence and persistence of new genotypes, since very few new mutants born during the n^{th} cycle are transferred to pop_{n+1}.

Thus, if the population histogram H_n at the start of the n^{th} cycle is given by the histogram H, one can directly calculate the probability $Prob(H_{n+1} = G | H_n = H)$ of transitioning from H to G at the end of the n^{th} cycle as well as the conditional expectation $\mathbb{E}[H_{n+1} | H_n = H]$, which are key in developing a large deviations framework for this Markov chain.

Large deviations principles have been established in various settings (see [1–3, 8, 14]). Using these principles, one can associate a "cost" $\Lambda(A) \geq 0$ to sets of trajectories $A \subset \Omega(T)$ where $\Omega(T)$ is the set of all trajectories of length T such that under a weak condition on A, one has for large N

$$Prob(H(0,T) \in A) \approx \exp\left(-N\Lambda(A)\right)$$

where $H(0,T) = [H_0\ H_1\ \ldots\ H_T]$ is the discrete Markov chain trajectory consisting of population histograms $H_n \in [0,1]^g$ on day n for $0 \leq n \leq T$. This large deviations principle applied to our setting so that the *rate functional* $\Lambda(A)$ became key in obtaining probabilistic estimates. Calculation of $\Lambda(A)$ first involved the explicit computation of a *one-step cost* function $C(H,G)$, which is achieved by directly calculating the logarithm of the probability of transitioning from one histogram H to another histogram G over one daily cycle. This calculation involved the use of Stirling's formula. Consequently, the one-step cost function $C(H,G)$ is only valid for histograms $H, G \in \mathbb{R}^g$ whose coordinates are at least $50/N$, creating a small

boundary in the space of histograms where $C(H, G)$ is invalid. The one-step cost function immediately yielded an explicit *cost function* $\lambda(w)$ for histogram trajectories w given as a sum of one-step cost functions. Finally, the rate functional $\Lambda(A)$ is obtained by minimizing $\lambda(w)$ over all paths $w \in A$.

Since we were interested in generating most likely evolutionary paths linking an initial trajectory H to a target histogram G, the above minimization was taken over all possible transition pathways linking H to G. The optimization was completed using Lagrange multipliers and Taylor expansions. This minimization revealed that for the final three points on any trajectory segment $[x\ y\ G]$, the histogram x is uniquely determined by y and G. One can then generate trajectories inductively in reverse time by fixing the desired target G and penultimate point y. However, there is currently no way to find the penultimate point y that generates the optimal trajectory. Consequently, reducing the number of penultimate points to consider becomes crucial when developing efficient numerical algorithms.

The complicated formulas involved in the reverse trajectories and cost function λ created many difficulties in proving rigorous results directly. Therefore, we relied on numerical evidence using realistic parameter sets in order to choose good sets of penultimate points and sensible discretizations for the space of histograms. We also relied on numerical evidence to illustrate the general nature of optimal trajectories. The key function used to choose good penultimate points and discretizations is the norm of the gradient $\|\nabla_y C(y, G)\|$ with respect to the penultimate point for a fixed target G. We claimed the following:

(1) For targets away from the boundary, the penultimate point that approximately minimizes $\|\nabla_y C(y, G)\|$ is a solution to a system of g linear equations independent of the choice of discretization.

(2) For any given target, the best discretization for the choice of penultimate point will be of adaptive size; namely, the discretization cell centered around a penultimate point y should be of size $c/\|\nabla_y C(y, G)\|$ where c is a proportionality constant. This can then be used to choose a neighborhood centered at the solution to the linear system in (1).

(3) When N is very large (10^8), the mutation rate m is very small (10^{-8}), and the target G is close to the boundary ($G(j) \approx 100/N$ for at least one genotype j), the norm of the gradient will tend to large values when the penultimate point y is close to the boundary. Consequently, in asymptotic situations with $N \to \infty$ and $m \to 0$, a trajectory that bounces back from the boundary will have a large cost and cannot be optimal.

In the case of three genotypes ($g = 3$), it is possible to let the set of penultimate points be all histograms away from the boundary and still have relatively efficient computations. However, when $g \geq 4$, this quickly becomes inefficient and causes major computer memory issues. Therefore, we used the case $g = 3$ to support the claims above where we let the set of possible penultimate points be all histograms away from the boundary. We also compared the two strategies, which showed that using our strategy yielded more efficient computations. Finally, we demonstrated the effectiveness of our strategy using an example in the case when $g = 4$.

3. **Research Directions**

During my postdoctoral tenure, I plan to continue research in applications of large deviations. This will enable collaboration with other departments (biology, biochemistry, etc.) and provide grant funding opportunities for the mathematics department. Working with different groups and models will also extend the theory of large deviations to models that are used to describe natural phenomena. We will extend mathematical knowledge while providing a service to others by addressing many computational problems that cannot be addressed by direct simulation or experimentation. I am willing to work on any problem suggested by my advisor as well. Below are a few areas I believe contain interesting open problems as well as provide opportunities for grant funding.

Applications of Large Deviations to Stochastic Models. Similar to my research in bacterial populations of E.Coli which utilized experiments from Tim Cooper's lab at the University of Houston, I would like to collaborate with local groups and labs who experiment with random models in order to provide concrete information about their models using large deviations. The areas are not limited to biology as large deviations has applications in biochemistry, physics, and finance to name a few. For instance, we could

collaborate with a biology group who studies and models bacteria other than E. Coli to establish a large deviations principle and compute most likely evolutionary trajectories under desired constraints of interest. We could also collaborate with a biochemistry group who models protein production in cellular populations. We could even collaborate with local finance companies who model the price of a stock option over time under certain assumptions.

Accuracy of Linear-Noise Approximations. One significant advantage to studying Gaussian diffusions with delay is that we can model a nonlinear delay SDE locally as a linear delay SDE and then apply our framework to this approximation. In our application to the co-repressive toggle switch, we calculated optimal escape trajectories from a small neighborhood of a metastable state using a linear approximation. The optimal trajectories were reasonable based on the model behavior of the co-repressive toggle switch. However, the true accuracy of these linear noise approximations is largely unknown. I would like to study the accuracy of linear noise approximations in modeling the behavior of nonlinear delay SDEs locally. I conjecture is a good approximation that is valid in a sufficiently small neighborhood of a metastable state.

Impact of Delay on Optimal Transitions. We saw the necessity of delay in modeling many biological processes. Numerous characteristics of lagged systems, including stability, are heavily dependent on the delay term. For a Gaussian diffusion with delay, varying the time delay while holding all other parameters constant can drastically affect the behavior of the mean and covariance of the process by introducing oscillatory behavior or unbounded solutions as time progresses. Consequently, optimal transitions and corresponding energy estimates would be affected since these formulas are explicitly written in terms of the mean and covariance. Therefore, I would like to investigate the influence of the delay on optimal transitions of generic lagged systems. Since optimal transitions have been established for Gaussian diffusions with delay, I would start my investigation with these systems.

Applying Large Deviations to Nonlinear SDEs. As stated when we applied large deviations to Gaussian diffusions with delay, a benefit in working with local linear approximations of nonlinear delay SDEs is that computations can be executed easily with linear models. In addition, while large deviations principles exist for many nonlinear SDEs, applications and calculations for these nonlinear models is difficult and largely unexplored since the relevant operators involved are highly nonlinear and implicit as opposed to the Gaussian case. Therefore, I would like to study the computational aspects of these nonlinear problems so that applications of large deviations principles to nonlinear models can be executed directly to the global nonlinear model. The goal would be to avoid using Hamilton-Jacobi equations, if possible, so that calculations can be executed efficiently.

References

[1] R. Azencott, *Grandes déviations et applications*, in Eighth Saint Flour Probability Summer School—1978 (Saint Flour, 1978), vol. 774 of Lecture Notes in Math., Springer, Berlin, 1980, pp. 1–176.

[2] R. Azencott, M. I. Freidlin, and S. R. S. Varadhan, *Large deviations at Saint-Flour*, Probability at Saint-Flour, Springer, Heidelberg, 2013.

[3] R. Azencott and G. Ruget, *Mélanges d'équations différentielles et grands écarts à la loi des grands nombres*, Z. Wahrscheinlichkeitstheorie und Verw. Gebiete, 38 (1977), pp. 1–54.

[4] F. Bouchet, T. Grafke, T. Tangarife, and E. Vanden-Eijnden, *Large deviations in fast-slow systems*, J. Stat. Phys., 162 (2016), pp. 793–812.

[5] F. Bouchet, J. Laurie, and O. Zaboronski, *Langevin dynamics, large deviations and instantons for the quasi-geostrophic model and two-dimensional Euler equations*, J. Stat. Phys., 156 (2014), pp. 1066–1092.

[6] T. Brett and T. Galla, *Stochastic processes with distributed delays: Chemical langevin equation and linear-noise approximation*, Physical Review Letters, 110 (2013).

[7] T. F. Cooper, D. E. Rozen, and R. E. Lenski, *Parallel changes in gene expression after 20,000 generations of evolution in escherichia coli*, Proceedings of the National Academy of Sciences, 100 (2003), pp. 1072–1077.

[8] A. Dembo and O. Zeitouni, *Large deviations techniques and applications*, vol. 38 of Applications of Mathematics (New York), Springer-Verlag, New York, second ed., 1998.

[9] A. Eldar and M. Elowitz, *Functional roles for noise in genetic circuits*, Nature, 467 (2010), pp. 167–173.

[10] M. I. Freidlin and A. D. Wentzell, *Random perturbations of dynamical systems*, vol. 260 of Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], Springer, Heidelberg, third ed., 2012. Translated from the 1979 Russian original by Joseph Szücs.

[11] C. Gupta, J. M. López, R. Azencott, M. R. Bennett, K. Josić, and W. Ott, *Modeling delay in genetic networks: From delay birth-death processes to delay stochastic differential equations*, J Chem Phys, 140 (2014), p. 204108.

[12] M. Hegreness, N. Shoresh, D. Hartl, and R. Kishony, *An equivalence principle for the incorporation of favorable mutations in asexual populations*, Science, 311 (2006), pp. 1615–1617.
[13] S. H. Rice, *Evolutionary Theory: Mathematical and Conceptual Foundations*, Sinauer Associates, Inc., 2001.
[14] S. R. S. Varadhan, *Large deviations and applications*, vol. 46 of CBMS-NSF Regional Conference Series in Applied Mathematics, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1984.
[15] W. Zhang, V. Sehgal, D. Dinh, R. Azevedo, T. Cooper, and R. Azencott, *Estimation of the rate and effect of new beneficial mutations in asexual populations*, Theoretical population biology, 81 (2012), pp. 168–178.