Numerical computation of rare events via large deviation theory

Tobias Grafke[1] and Eric Vanden-Eijnden[2]

[1] *Mathematics Institute, University of Warwick, Coventry CV4 7AL, United Kingdom*[a]

[2] *Courant Institute, New York University, 251 Mercer Street, New York, NY 10012, USA*[b]

An overview of rare events algorithms based on large deviation theory (LDT) is presented. It covers a range of numerical schemes to compute the large deviation minimizer in various setups, and discusses best practices, common pitfalls, and implementation trade-offs. Generalizations, extensions, and improvements of the minimum action methods are proposed. These algorithms are tested on example problems which illustrate several common difficulties which arise e.g. when the forcing is degenerate or multiplicative, or the systems are infinite-dimensional. Generalizations to processes driven by non-Gaussian noises or random initial data and parameters are also discussed, along with the connection between the LDT-based approach reviewed here and other methods, such as stochastic field theory and optimal control. Finally, the integration of this approach in importance sampling methods using e.g. genealogical algorithms is explored.

Rare events often have a drastic impact despite their low frequency of occurrence. Examples include hurricanes, financial crises, heat waves, or tsunamis, that are few and far between but have devastating consequences. Other important phenomena such as phase transitions, chemical reactions, or conformational changes of biomolecules also involve rare events. The accurate description of these events is complicated, since their low rate of occurrence makes them hard to observe both in experiments and in simulations. In many cases, when a rare event occurs, it does so in its least unlikely form, the instanton, rendering all other realizations of the same event negligible in comparison. Whenever such a situation holds, a large deviation principle (LDP) quantifies this concentration phenomenon. The LDP specifies a deterministic optimization problem to identify the instanton, and allows the estimation of its probability. In this review, we discuss numerical algorithms to solve the large deviation optimization problem, compare their associated trade-offs, and present best practices, pitfalls, improvements, and generalizations.

I. INTRODUCTION

Rare but important events are by definition difficult to observe, both in experiments and in simulations. In order to design efficient schemes for the numerical computation of these events one therefore typically resorts to one of the following two strategies: either manipulation of the system in a controlled way that makes rare events more likely and can be corrected *a posteriori*; or computation of a single dominant event characterizing the possible ways the rare event happens. The first approach can be categorized as importance sampling; the second can be justified within sample path large deviation theory (LDT) and leads to an action minimization problem to be solved. In this review, we focus mainly on algorithmic developments in the second class, and discuss the interplay between this LDT-based approach and importance sampling towards the end of our paper. The techniques covered focus solely on the small noise (or low temperature) limit of stochastic processes covered by Freidlin-Wentzell theory[1] and generalizations thereof. Specifically, in section II we discuss rare event algorithms based on the global minimization of LDT action functionals, suitable for computing paths by which infrequent transitions between two prescribed states occur. Subsequently, in section III, we explain how to calculate large deviation minimizers in the context of the estimation of rare expectations dominated by tail statistics. In section IV, we generalize these two approaches to the non-Gaussian case. In section V, we demonstrate a generalization to arbitrary dynamical systems with random initial conditions and parameters. Section VI suggests possibilities to use the minimizing trajectories obtained by the earlier algorithms as input for importance sampling algorithms. Finally, some concluding remarks are presented in section VII. Many of the covered techniques and algorithms are not new, and more in-depth discussions exist in the literature; further reading is indicated on a per-case basis. Some of the example applications, notably the application to extreme concentrations of prey in the Lotka-Volterra model (section III B), extreme amplitudes in the Korteweg-de Vries equation (section III C), optimal excitations of the Fitzhugh-Nagumo model (section V B), and extreme number of infections in an epidemiology model with vaccination (section VI C) are new to this review.

In the remainder of the present section we review the aspects of LDT relevant to our purpose, with focus on Freidlin-Wentzell theory for dynamical systems subject to random noise of low amplitude.

---

[a] Electronic mail: T.Grafke@warwick.ac.uk

[b] Electronic mail: eve2@cims.nyu.edu

A. Freidlin-Wentzell theory of large deviations

Consider a dynamical system with variables X_t^ε in $\mathbb{R}^d$, subject to small random perturbations that are additive Gaussian and white in time. Assuming that the noise amplitude scales with the smallness-parameter ε, the evolution of the stochastic variables X_t^ε is described by the stochastic differential equation (SDE)

$$dX_t^\varepsilon = b(X_t^\varepsilon)\,dt + \sqrt{\varepsilon}\sigma\,dW_t, \quad X_0^\varepsilon = x, \ \ t \geq 0, \quad (1)$$

with deterministic drift $b : \mathbb{R}^d \to \mathbb{R}^d$, noise covariance $a = \sigma\sigma^\top$ with $\sigma \in \mathbb{R}^{d\times d}$, and where W_t is a d-dimensional Wiener process – for simplicity, we assume here that a is independent of the system's position (i.e. the noise is additive) and invertible: the generalization to multiplicative noise and degenerate noise will be discussed through examples. We are interested in situations where the stochastic process (1) realizes a certain event, for example when the trajectory ends at time T in a given set $A \subset \mathbb{R}^d$, so that $X_T^\varepsilon \in A$. Even if these events are impossible in the deterministic system ($\varepsilon = 0$), they will in general occur in the presence of noise ($\varepsilon > 0$) but they become rarer and rarer in the low noise limit, $\varepsilon \to 0$.

Large deviation theory (LDT) gives a precise characterization of this decay of probability: The probability of observing any sample path close to a given function $\phi(t)$ can be estimated as (see chpt 3 of[1])

$$P\Big\{ \sup_{t\in[0,T]} \|X_t^\varepsilon - \phi(t)\| < \delta \Big\} \asymp \exp\left(-\varepsilon^{-1} S_T(\phi)\right), \quad (2)$$

for small enough $\delta > 0$, where $\asymp$ denotes log-asymptotic equivalence (i.e. for $\varepsilon \to 0$, the ratio of the logarithms of both sides converges to 1). The functional $S_T(\phi)$ is called the *rate function* or *action functional*, and it is generally given by

$$S_T(\phi) = \begin{cases} \int_0^T L(\phi,\dot{\phi})\,dt & \text{if the integral converges,} \\ \infty & \text{otherwise.} \end{cases} \quad (3)$$

Here we defined the *Lagrangian* $L(\phi,\dot{\phi})$, which for the concrete example of equation (1) is given by:

$$L(\phi,\dot{\phi}) = \tfrac{1}{2}\|\dot{\phi} - b(\phi)\|_a^2 \quad (4)$$

via the a-metric induced by the inner product $\|v\|_a^2 = \langle v, a^{-1}v\rangle$. The corresponding action functional is then termed the *Freidlin-Wentzell action functional*.

The probability of observing the event $X_T^\varepsilon \in A$ consists of contributions of the sample paths close to all the possible absolutely continuous $\phi(t) \in \mathscr{C} = \{\phi(t) \in AC([0,T],\mathbb{R}^d) \mid \phi(0) = x, \phi(T) \in A\}$, and each of these contributions scales according to equation (2). Consequently, in the limit $\varepsilon \to 0$, the only contribution that matters is that coming from the trajectory $\phi^*(t)$ with the smallest action $S_T(\phi^*)$. We call

$$\phi^*(t) = \operatorname*{argmin}_{\phi\in\mathscr{C}} S_T(\phi) \quad (5)$$

the *maximum likelihood pathway* (MLP) or *instanton*. It constitutes the *least unlikely* trajectory to realize the rare event, in the sense that almost surely all sample paths conditioned on the rare event will be arbitrarily close to $\phi^*(t)$. More precisely (compare section 3.1 of[1]), for all $\delta > 0$ sufficiently small, we have

$$\lim_{\varepsilon\to 0} P\Big\{ \sup_{t\in[0,T]} \|X_t^\varepsilon - \phi^*(t)\| < \delta \ \big|\ X_T^\varepsilon \in A \Big\} = 1. \quad (6)$$

The efficient numerical solution of the minimization problem (5) for different rare events (and therefore different sets of trajectories $\mathscr{C}$ to minimize over) lies at the core of this work.

If the action at the instanton, $S_T(\phi^*)$ is zero, the corresponding trajectory fulfills $\dot{\phi} = b(\phi)$ and can be considered *deterministic*, i.e. the corresponding evolution is the one selected by the deterministic dynamics ($\varepsilon = 0$). If, on the other hand, the action at the instanton is finite, the probability of observing the corresponding event in a given time frame T decays to zero as indicated by equation (2).

LDT additionally permits the analysis of the effect of infinitesimal perturbations over an infinite time interval, $T \to \infty$, on which these rare events almost surely happen. The central object in this context is the *quasipotential*, defined as

$$V(x,y) = \inf_{T>0}\min_{\phi\in\mathscr{C}_{x,y}} S_T(\phi), \quad (7)$$

where $\mathscr{C}_{x,y} = \{\phi \in AC([0,T],\mathbb{R}^d) \mid \phi(0) = x, \phi(T) = y\}$. The quasipotential characterizes the long time behavior of the system. For example, if the deterministic system $\dot{X} = b(X)$ possesses only one single stable fixed point $\bar{x}$, with basin of attraction $\mathbb{R}^d$, then the density $\rho(x)$ associated with the invariant measure of equation (1) can be written in the limit $\varepsilon \to 0$ as

$$\rho(x) \asymp \exp\left(-\varepsilon^{-1} V(\bar{x},x)\right) \quad (8)$$

(compare chapter 4, theorem 4.3 of[1]). Similarly, in situations where the deterministic system has multiple fixed points $\bar{x}_i$, the mean first passage time $\tau_{i,j}$ between the basins of attraction of neighboring $\bar{x}_i$ and $\bar{x}_j$,

$$\tau_{i,j} = \mathbb{E}\inf\{t > 0 \mid X(0) = \bar{x}_i, \|X(t) - \bar{x}_j\| < \delta\} \quad (9)$$

with $\delta > 0$ small enough, can be estimated in the small noise limit as

$$\tau_{i,j} \asymp \exp\left(\varepsilon^{-1} V(\bar{x}_i,\bar{x}_j)\right) \quad (10)$$

(see section 4.4 of[1]). This result also allows the investigation of the long time dynamics of the system by mapping it onto a Markov jump process whose states are the fixed points $\bar{x}_i$, $\bar{x}_j$, etc. and whose transition rates are $k_{i,j} \asymp \tau_{i,j}^{-1}$.

All these examples show that it is useful to have access to the minimizing trajectory ϕ^* to describe rare events:

First it gives the typical way a rare event is observed, enabling us to identify their mechanism. Second it allows the estimation of their probability of occurrence, and their expected recurrence time. Third it gives the relative probability of multiple typical (i.e. deterministically stable) states, and the most likely way by which transitions between them occur.

B. Hamiltonian principle and connections to classical mechanics and field theory

The minimization problem in equation (5) to find the instanton precisely corresponds to Hamilton's principle from classical mechanics, $\delta S_T(\phi)/\delta\phi = 0$. As a consequence, the methods and ideas from classical mechanics are transferable to our situation. In particular, the variational problem can be solved by seeking solutions of the corresponding Euler-Lagrange equation,

$$\frac{\partial L}{\partial \phi} - \frac{d}{dt}\frac{\partial L}{\partial \dot{\phi}} = 0\,, \tag{11}$$

which, for a system of the type (1), gives

$$a^{-1}\ddot{\phi} + \left(a^{-1}\nabla b(\phi) - \nabla b(\phi)^{\top}a^{-1}\right)\dot{\phi} + \nabla\langle b(\phi), a^{-1}b(\phi)\rangle = 0\,. \tag{12}$$

Several algorithms presented below aim at the numerical solution of the second order equation (12).

Similarly inspired by classical mechanics, we can define a *conjugate momentum*

$$\theta = \frac{\partial L(\phi, \dot{\phi})}{\partial \dot{\phi}}\,, \tag{13}$$

and a *Hamiltonian* as Fenchel-Legendre transform of the Lagrangian,

$$H(\phi, \theta) = \sup_{y}\left(\langle \theta, y\rangle - L(\phi, y)\right)\,, \tag{14}$$

such that, assuming convexity of $L(\phi, \dot{\phi})$ in $\dot{\phi}$,

$$L(\phi, \dot{\phi}) = \sup_{\theta}\left(\langle \dot{\phi}, \theta\rangle - H(\phi, \theta)\right)\,. \tag{15}$$

The minimization (5) is then equivalent to solving Hamilton's equations of motion, or *instanton equations*,

$$\begin{cases} \dot{\phi} = \nabla_{\theta}H(\phi, \theta) \\ \dot{\theta} = -\nabla_{\phi}H(\phi, \theta)\,. \end{cases} \tag{16}$$

For the system (1), the Hamiltonian is given by

$$H(\phi, \theta) = \langle b(\phi), \theta\rangle + \tfrac{1}{2}\langle \theta, a\theta\rangle\,, \tag{17}$$

so that the instanton equations read

$$\begin{cases} \dot{\phi} = b(\phi) + a\theta \\ \dot{\theta} = -(\nabla b(\phi))^{\top}\theta\,. \end{cases} \tag{18}$$

Solving the instanton equations (16) constitutes another possible approach to solving the minimization problem (5), but care has to be taken to obtain the correct boundary conditions for (16), depending on the rare event under consideration—this point will be discussed at length below and we will see that these boundary conditions make working with (12) more appropriate in some cases and with (18) in others. Notably, neither (17) nor (18) necessitate inverting a, which we can exploit in the case of degenerate forcing with non-invertible a.

The Hamiltonian $H(\phi, \theta)$ is a conserved quantity along the minimizing trajectory, since

$$dH/dt = \langle \nabla_{\phi}H, \dot{\phi}\rangle + \langle \nabla_{\theta}H, \dot{\theta}\rangle = 0\,. \tag{19}$$

Additional simplifications apply in the special case that the minimizing trajectory starts at rest at a fixed point $\bar{x}$ of the deterministic dynamics, in which case individually $b(\bar{x}) = 0$ and $\theta = 0$. This necessitates at the same time that the transition time T diverges to ∞, and furthermore that the Hamiltonian vanishes, $H(\phi, \theta) = 0$. This property can in turn be used to rewrite the action functional (3) as

$$S(\phi) = \int_0^{\infty} L(\phi, \dot{\phi})\,dt = \int_0^{\infty}(\langle \dot{\phi}, \theta\rangle - H(\phi, \theta))\,dt = \int \langle \theta, d\phi\rangle\,. \tag{20}$$

Writing the action in this form is an instance of the Maupertuis principle in mechanics, which minimizes over paths of a given energy. Since it omits any explicit time parametrization, it offers an approach at solving the double minimization problem to calculate the quasipotential in (7).

Interestingly, there is a parallel between the LDT discussed above and concepts from field theory applied to stochastic systems, first established by Onsager and Machlup [2]. Later, the Janssen-de Dominicis formalism [3,4] based on the Martin-Siggia-Rose path integral [5], considers computing expectations as path-integrals over all possible noise realizations, and performs a change of variables to the field variable itself. The constraint of the dynamics is embedded as Lagrange multiplier, which gives rise to an additional *auxiliary field*, corresponding to the conjugate momentum. Similarly, the minimization problem (5) then amounts to finding a semiclassical trajectory as saddle-point approximation of the action functional. It is this correspondence which is the root of the terms "action functional" and "instanton" for the rate function and its minimizer. Noteworthy in this context is also the Doi-Peliti formalism [6,7], which follows a similar route for dominant reaction pathways.

The same results discussed in this section can be obtained by instead applying a Wentzel-Kramers-Brillouin (WKB) approximation to the Fokker-Planck equation associated to the stochastic process [8,9].

C. Detailed balance and gradient flows

A special case of interest is when the dynamics is a *diffusion in a potential*, with the drift given by the negative gradient of a potential $U : \mathbb{R}^d \to \mathbb{R}$ and $\sigma = \sqrt{2}\text{Id}$, where Id is the identity on $\mathbb{R}^{d\times d}$, such that equation (1) becomes

$$dX_t^\varepsilon = -\nabla U(X_t^\varepsilon)\,dt + \sqrt{2\varepsilon}\,dW_t\,. \qquad (21)$$

Suppose that we look at the calculation of the quasipotential (7) between two local minima of U located at x_a and x_b and with adjacent basins of attraction. In this case, the minimum of (7) is approached by taking either $\dot\phi = -\nabla U(\phi)$ (in which case the action is zero), or we realize that (see chpt 3 of[1])

$$\begin{aligned}
S_T(\phi) &= \tfrac{1}{4}\int_0^T |\dot\phi + \nabla U(\phi)|^2\,dt\\
&= \int_0^T |\dot\phi - \nabla U(\phi)|^2\,dt + \int_0^T \langle\dot\phi, \nabla U(\phi)\rangle\,dt\\
&= \int_0^T |\dot\phi - \nabla U(\phi)|^2\,dt + \phi(T) - \phi(0)\,.
\end{aligned}$$

Now, since the last terms depend only on the trajectory end-points, we are free to choose $\dot\phi = \nabla U(\phi)$ to make the first integral disappear. As a consequence, for a diffusion in a potential landscape of the form (21), to calculate the minimum involved in the quasipotential, we can patch together the solutions of $\dot\phi = \pm\nabla U(\phi)$ that connects x_a and x_b via the saddle point x_s of minimum potential. We can interpret that to say that the minimum is achieved by following either the deterministic dynamics $\dot\phi = -\nabla U(\phi)$, or its *time-reversed* version. This is nothing but a manifestation of *time-reversal symmetry* that is the consequence of the random process defined by equation (21) being in *detailed balance* with respect to the stationary distribution.

This simple relationship between the tangential vector $\dot\phi$ and the deterministic drift ∇U simplifies the computation of the minimizers significantly. In particular, we realize that minimizers for gradient flows are *heteroclinic orbits* of the dynamical system defined by the deterministic drift. As such, they are numerically accessible by the *string method*[10,11].

Similar simplifications as the above can be realized for any system in detailed balance with respect to its stationary distribution, and as a results, its large deviation minimizers are always heteroclinic orbits of the associated *generalized* gradient flow (but not necessarily of a traditional gradient flow)[12].

II. RARE EVENT ALGORITHMS FOR NOISE-INDUCED TRANSITIONS

In this section, we want to consider a particular subclass of problems of the form discussed in section I A:

The computation of the optimal noise-induced transition trajectory from a basin of attraction around one fixed point of the deterministic dynamics to a neighboring one. Specifically, we are not considering more complicated attracting structures such as limit cylces, and only consider transitions between neighboring basins. For these assumptions, the minimization problem (5) becomes

$$\phi^*(t) = \operatorname*{argmin}_{\phi\in\mathscr{C}_{x,y}} S_T(\phi)\,, \qquad (22)$$

where $\mathscr{C}_{x,y} = \{\phi \in AC([0,T],\mathbb{R}^d) \mid \phi(0) = x, \phi(T) = y\}$, and the instanton constitutes the maximum likelihood transition trajectory between the two deterministically stable fixed points x and y. By additionally minimizing over the transition time T, the resulting instanton can be used to compute the quasipotential (7), at least along the instanton trajectory. Alternatives, such as fast marching techniques[13], are viable only in low dimensions. Here, instead, we perform the computation of the quasipotential by applying the minimum action method[14], that discretizes the action functional (3), and considers the discrete (finite dimensional) gradient as descent direction for numerical minimization algorithms, such as gradient descent or quasi-Newton methods. In section II A, we present a simplified version of the minimum action method, and discuss its implementation details. In section II B, this method is then illustrated by applying it to a simplified metastable atmosphere dynamics model. Finally in section II C we discuss generalizations to stochastic partial differential equations, and consider the example of the stochastic Burgers-Huxley model in section II D.

A. A simplified geometric minimum action method

One obvious disadvantage of a straightforward discretization of the Freidlin-Wentzell action functional (3) is its inability to treat infinite transition times. In the context of the quasipotential, we are looking for transition trajectories of arbitrary transition time T, which generally diverges, $T \to \infty$, since the trajectory contains fixed points. The minimum of the outer minimization in the computation of the quasipotential,

$$V(x,y) = \inf_{T>0} \min_{\phi\in\mathscr{C}_{x,y}} S_T(\phi)\,, \qquad (23)$$

is simply not attained for any finite T in these cases. This complication was successfully addressed with the *geometric* minimum action method[15], which instead considers a minimization over the space of arc-length parametrized curves that may remain finite even for diverging transition time. In this section, we want to introduce a simplified version of this geometric picture, allowing us to formulate an algorithm to compute of the geometric minimizer with a lower number of derivatives of the Hamiltonian.

Based on the Maupertuis principle (20), the minimizing trajectory ϕ between two fixed points x and y, when additionally minimizing over the transition time T, fulfills $H(\phi, \theta) = 0$, and the corresponding action (20) is given by

$$S(\phi) = \int_x^y \langle \theta, d\phi \rangle. \tag{24}$$

This form of the action makes it obvious that the action is invariant under reparametrization: The total action is a line-integral along the minimizer, and we are free to choose any parametrization to describe it. This enables us to treat infinite time-intervals with finitely many discretization points, for example by parametrizing (24) by normalized arc-length.

The minimization problem (5) can be rewritten as a nested optimization problem,

$$\phi^* = \underset{\phi \in \mathscr{C}_{x,y}}{\operatorname{argmin}} \; \underset{\theta : H(\phi,\theta)=0}{\sup} \; E(\phi, \theta), \tag{25}$$

with

$$E(\phi, \theta) = \int_0^1 \langle \theta, \phi' \rangle \, ds. \tag{26}$$

Here the prime denotes differentiation with respect to the parametrization s we choose for ϕ and θ, and we impose $\|\phi'\|_\sim = L = \text{const}$, with L the length of the curve. Note that the algorithm works independently of the choice of the norm, and we will discuss appropriate norms at the end of this section. Therefore, in the following, the norm $\| \cdot \|_\sim$ and corresponding inner product $\langle \cdot, \cdot \rangle_\sim$ are to be seen as a placeholder for our preferred choice.

Let

$$E^*(\phi) = \underset{\theta : H(\phi,\theta)=0}{\sup} E(\phi, \theta) \tag{27}$$

and $\theta^*(\phi)$ be the solution of the inner optimization problem (27), such that $E(\phi, \theta^*(\phi)) = E^*(\phi)$. Then, equivalently, $\theta^*(\phi)$ fulfills the Euler-Lagrange equation for the constrained maximization problem (27). Using $\delta E(\phi, \theta)/\delta\theta = \phi'$, this Euler-Lagrange equation reads

$$\phi' = \mu \nabla_\theta H(\phi, \theta), \tag{28}$$

where $\mu(s)$, $s \in [0, 1]$, is a Lagrange multiplier to enforce the constraint of a vanishing Hamiltonian. This Lagrange multiplier is explicitly computable by multiplying equation (28) by ϕ' and solving for μ, i.e.

$$\mu = \frac{\|\phi'\|_\sim^2}{\langle \phi', \nabla_\theta H \rangle_\sim}. \tag{29}$$

Similarly, using $\delta E(\phi, \phi)/\delta\phi = -\theta'$ the functional derivative of $E^*(\phi)$ with respect to ϕ can be expressed as

$$\frac{\delta E^*(\phi)}{\delta\phi} = -\theta^{*'}(\phi) + \mu \nabla_\theta H(\phi, \theta^*(\phi)) \nabla_\phi \theta^*(\phi)$$
$$= -\theta^{*'}(\phi) - \mu \nabla_\phi H(\phi, \theta^*(\phi)), \tag{30}$$

where the last step makes use of

$$\nabla_\phi H(\phi, \theta^*(\phi)) = -\nabla_\theta H(\phi, \theta^*(\phi)) \nabla_\phi \theta^*(\phi),$$

which holds by definition due to $H(\phi, \theta^*(\phi)) = 0$.

Note how in this formulation the reparametrization into arc-length emerges naturally as Lagrange multiplier μ to enforce the Hamiltonian constraint. In particular, comparing equation (28) with Hamilton's equation with respect to physical time, $d\phi/dt = \nabla_\theta H(\phi, \theta)$ shows that the Lagrange multiplier μ is nothing but the change of parametrization, $\mu = dt/ds$ from physical time to arc-length parametrization.

Taking these equivalences, the nested optimization problem (25) can now be solved in an iterative manner. Starting from the k-th guess ϕ^k for the transition trajectory,

(i) solve the inner constrained optimization problem

$$\theta^k = \theta^*(\phi^k) = \underset{\theta : H(\phi^k,\theta)=0}{\operatorname{argmax}} E(\phi^k, \theta),$$

(ii) compute a descent direction for the outer optimization problem,

$$d^k = \delta E^*(\phi^k)/\delta\phi^k = \dot{\theta}^k + \nabla_\phi H(\phi^k, \theta^k), \tag{31}$$

(iii) descent along the descent direction, for example by gradient descent, pre-conditioned with μ^{-1}, and step-length α,

$$\phi^{k+1} = \phi^k + \alpha \mu^{-1} d^k, \tag{32}$$

to obtain the next guess ϕ^{k+1}, and finally

(iv) iterate until convergence.

In step (iii), with pre-conditioning we refer to the fact that the direction d^k specified by the gradient is only one of many possible directions in which the cost function decreases. In fact, any direction $\tilde{d}^k$ such that the inner product between $\tilde{d}^k$ and d^k remains a valid search direction.

For the specific case of the small-noise Gaussian SDE, equation (1), this algorithm can be even more simplified. In particular, the inner constrained optimization problem to find $\theta^*(\phi)$ can be solved analytically, instead of relying on numerical optimization. Taking the Euler-Lagrange equation (28) for the inner optimization problem, together with the specific form of the Hamiltonian (17), yields

$$\phi' = \mu \nabla_\theta H(\phi, \theta^*(\phi)) = \mu(b(\phi) + a\theta^*(\phi)),$$

so that

$$\theta^*(\phi) = a^{-1}(\mu^{-1}\phi' - b(\phi)). \tag{33}$$

On the other hand, the Lagrange multiplier μ is directly available without knowledge of θ^*: Since

$$\|\nabla_\theta H\|_a^2 = \|b + a\theta\|_a^2 = \|b\|_a^2 + 2\langle b, \theta \rangle + \|a\theta\|_a^2 = \|b\|_a^2 + 2H = \|b\|_a^2, \tag{34}$$

we conclude that

$$\mu = \frac{\|\nabla_\theta E(\phi, \theta^*(\phi))\|_a}{\|\nabla_\phi H(\phi, \theta^*(\phi))\|_a} = \frac{\|\phi'\|_a}{\|b + \theta\|_a} = \frac{\|\phi'\|_a}{\|b\|_a} . \quad (35)$$

(35) naturally leads to the choice $\| \cdot \|_\sim = \| \cdot \|_a$. The descent direction is then immediately available as

$$d^k = \theta^{k\prime} + (\nabla b(\phi^k))^\top \theta^k ,$$

with θ^k and μ given by (33) and (35), respectively.

We want to make a few points about possible pitfalls and best practices.

- Even though any parametrization $s(t)$ is permissible, as discussed above it is natural to choose arclength, such that $\|\phi'\|_\sim = \mathrm{const}$. This parametrization can be enforced, as in the improved string method[11] and the original geometric minimum action method[15], by interpolation along the trajectory. This avoids stiff terms enforcing the parametrization constraint.

- Pre-conditioning is necessary to obtain good convergence. Pre-conditioning with μ^{-1} is necessary to ensure convergence around fixed points. Additionally, pre-conditioning with $\nabla_\theta \nabla_\theta H$ is often beneficial. This corresponds to the noise covariance a in the SDE case. For general Hamiltonians, this comes at the cost of needing to compute higher derivatives of the Hamiltonian, which one might want to avoid. Details about these considerations are discussed in[16].

- The choice of norm has to be taken with care as well. For the additive Gaussian case, as pointed out above, it is natural to use $\| \cdot \|_\sim = \| \cdot \|_a$. This generalizes to $\langle \cdot, (\nabla_\theta \nabla_\theta H)^{-1} \cdot \rangle$, which is the choice of the traditional geometric minimum action method (gMAM[15]). For simplicity, the Euclidean norm might be preferred in the general case to avoid the computation of higher order derivatives of H.

B. Example: Metastability in a simple atmosphere dynamics model

We want to demonstrate the effectiveness of the algorithm introduced in section II A by applying it to a problem motivated by meta-stability in a simplified atmosphere dynamics model introduced by Charney and DeVore[17]. Starting from the two-dimensional barotropic vorticity equation for the atmospheric flow, a projection of the stream function $\psi(x, y)$ on the 6 dominant spatial

Fourier modes is performed, resulting in an SDE system

$$
\begin{aligned}
dx_1 &= (\tilde{\gamma}_1 x_3 - C(x_1 - x_1^*))\, dt + \sqrt{2\varepsilon}\, dW_1 , \\
dx_2 &= (-(\alpha_1 x_1 - \beta_1)x_3 - Cx_2 - \delta_1 x_4 x_6)\, dt + \sqrt{2\varepsilon}\, dW_2 , \\
dx_3 &= ((\alpha_1 x_1 - \beta_1)x_2 - \gamma_1 x_1 - Cx_3 + \delta_1 x_4 x_5)\, dt + \sqrt{2\varepsilon}\, dW_3 , \\
dx_4 &= (\tilde{\gamma}_2 x_6 - C(x_4 - x_4^*) + \eta(x_2 x_6 - x_3 x_5))\, dt + \sqrt{2\varepsilon}\, dW_4 , \\
dx_5 &= (-(\alpha_2 x_1 - \beta_2)x_6 - Cx_5 - \delta_2 x_3 x_4)\, dt + \sqrt{2\varepsilon}\, dW_5 , \\
dx_6 &= ((\alpha_2 x_1 - \beta_2)x_5 - \gamma_2 x_4 - Cx_6 + \delta_2 x_2 x_4)\, dt + \sqrt{2\varepsilon}\, dW_6 ,
\end{aligned}
$$
$$(36)$$

where, for $m \in \{1, 2\}$,

$$
\begin{aligned}
\alpha_m &= \frac{8\sqrt{2}}{\pi} \frac{m^2}{4m^2 - 1} \frac{b^2 + m^2 - 1}{b^2 + m^2} , \\
\beta_m &= \frac{\beta b^2}{b^2 + m^2} , \\
\gamma_m &= \gamma \frac{\sqrt{2}b}{\pi} \frac{4m^3}{(4m^2 - 1)(b^2 + m^2)} , \\
\tilde{\gamma}_m &= \gamma \frac{\sqrt{2}b}{\pi} \frac{4m}{4m^2 - 1} , \\
\delta_m &= \frac{64\sqrt{2}}{15\pi} \frac{b^2 - m^2 + 1}{b^2 + m^2} , \\
\eta &= \frac{16\sqrt{2}}{5\pi} .
\end{aligned}
$$
$$(37)$$

The original model is detailed in[17], and was modified in (36) to add additive Gaussian noise to each degree of freedom. The model (36) allows for two metastable states, the so-called "zonal" state, and the "blocked" state, alluding to the atmospheric blocking phenomena observed in meteorology.

Application of the action minimization algorithm introduced in section II A allows us to compute the most likely transition trajectories in the small noise limit, $\varepsilon \to 0$, and deduce the relative stability of the states. The results are shown in figure 1: Starting from the zonal state in the upper left corner, snapshots of the stream function $\psi(x, y)$ are shown along the transition trajectory in reading order, arriving at the blocked state at the bottom right. For comparison, the right collection of plots in figure 1 shows the corresponding backward transition from blocked to zonal state. Note that the backward transition is not merely the time-reversed forward transition, implying (as expected) a breaking of time-reversal symmetry and thus demonstrating the non-equilibrium nature of the transition. The relative stability of the two configurations can be quantified via figure 2: The action to transition towards the blocked state is far larger than the action to transition towards the zonal state, meaning that the zonal state is exponentially preferred in the low-noise limit.
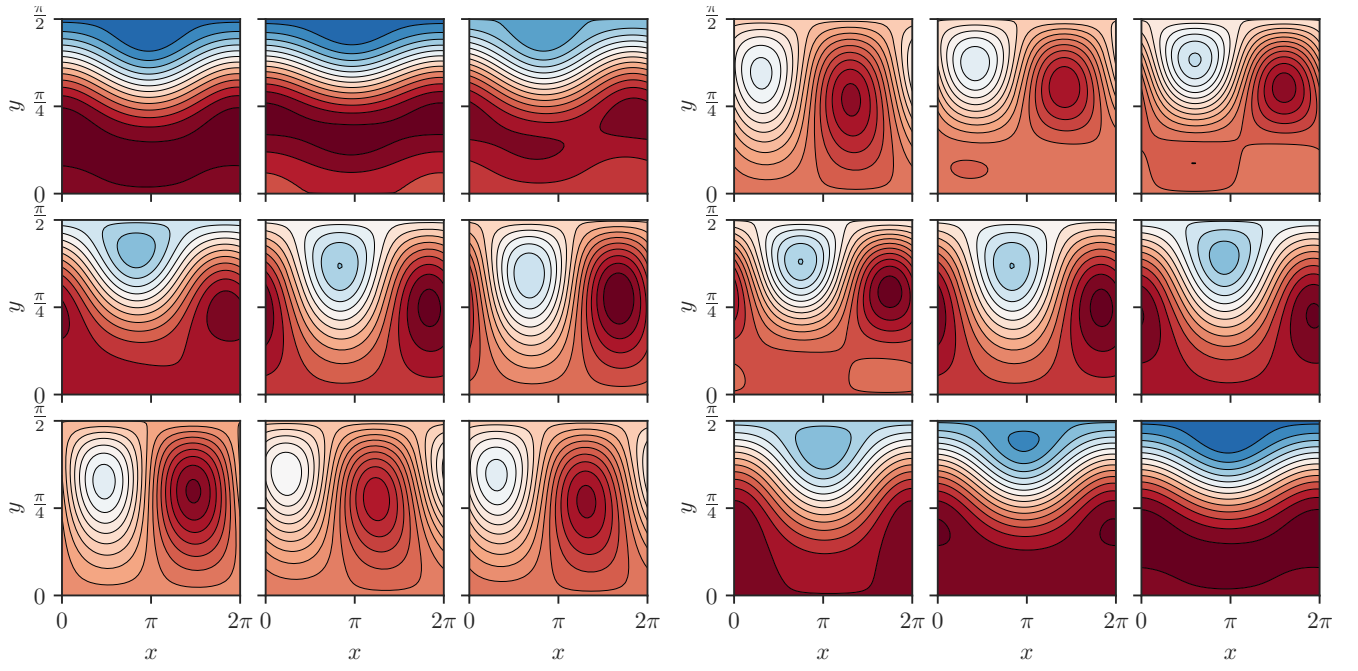
FIG. 1. Left: Stream function $\psi(x,y)$ along the transition from zonal to blocked configuration, where the arclength-parameter is increased left-to-right, top-to-bottom. The central configuration is the unstable saddle configuration on the separatrix between the basins of attraction of zonal and blocked configuration. Right: Stream function $\psi(x,y)$ along the transition from blocked to zonal configuration. Notably, this backward transition is not identical to the time-reversal of the forward transition depicted on the left, but again the same saddle is visited, as visible in the center field.
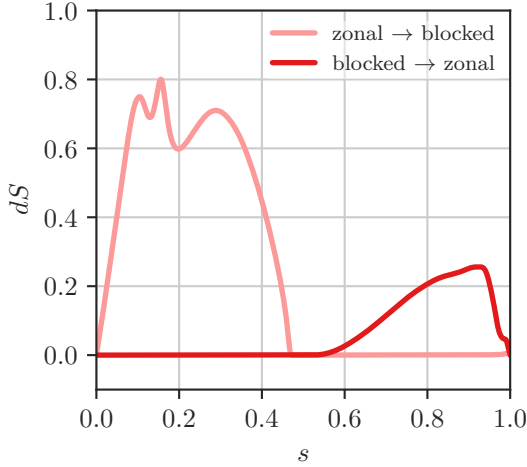

FIG. 2. Action density along the transition trajectories between the zonal and the blocked configuration, where s is the arclength parameter along the transition. As clearly visible, the transition towards the blocked state occurs at higher action, making the blocked state relatively more stable.

C. Instantons for stochastic partial differential equations

Many systems of interest in physical applications have continuous spatial variables, i.e. do not fit the framework of equation (1). Instead, they are stochastic *partial* differential equations (SPDEs). Applying the algorithm of section II A to stochastic processes in infinite-dimensional spaces is nevertheless largely done in practice. The mathematical foundation is less clear in this case, though, and a few comments are in order.

A stochastic partial differential equation, even in the simplest case of additive Gaussian noise, is possibly ill-posed. Consider for example

$$\partial_t U = B(U) + \sqrt{\varepsilon}\,\eta(x,t)\,, \qquad (38)$$

where $U : [0,T] \times \mathbb{R}^d \to \mathbb{R}^m$ and η denotes temporal white noise. If the noise is also not smooth in space, for example if it is white-in-space as well, $\mathbb{E}\eta(x,t)\eta(x',t') = \delta(t-t')\delta(x-x;)$, it is a non-trivial undertaking to make sense of possible non-linear terms in the drift, $B(U)$, especially if the spatial dimension is higher than one. Recent mathematical breakthroughs[18] specify a rigorous renormalization procedure in specific cases. In regards to LDT, the main concern is whether this renormalization procedure subsists in the limit $\varepsilon \to 0$. For example, in[19], it was discussed for the stochastic Allen-Cahn equation (e.g. $B(U) = U - U^3 + \kappa\partial_x^2 U$) in 2 or 3 spatial dimensions, that indeed the rate function corresponds to the naively

assumed one,

$$S_T(\phi) = \begin{cases} \int_0^T \|\partial_t \phi - B(\phi)\|_{L^2}^2 \, dt, & \text{if the integral converges,} \\ \infty & \text{otherwise,} \end{cases}$$
(39)

where $\|\cdot\|_{L^2}$ denotes the L^2-norm in the spatial components. In the following, we will consider SPDE examples, but no longer dwell upon the mathematical intricacies, instead assuming that (39) is valid.

If the rate function takes the form (39), all arguments put forward in the finite dimensional case can be transferred to the SPDE situation, and a corresponding algorithm can be constructed. In particular, a gradient descent of the form introduced in section II is still feasible, with gradients of vectors replaced by functional derivatives of the corresponding operators. Therefore, equation (31) to compute the descent direction for the SPDE (38) becomes

$$d^k = \theta^{k'} + (D_\phi B(\phi))^\top \theta^k \,,$$
(40)

where $A^\top$ is the L^2-adjoint of the (differential) operator A and D_ϕ the functional derivative. Consider for example Burgers equation with periodic boundary conditions, where

$$B(U) = \nu \partial_x^2 U - U \partial_x U \,.$$
(41)

Then $D_\phi B(\phi)$ is the operator

$$D_\phi B(\phi) = \nu \partial_x^2 - (\partial_x U) - U \partial_x$$
(42)

such that

$$(D_\phi B(\phi))^\top = \nu \partial_x^2 + U \partial_x \,.$$
(43)

Recall that we can still compute θ^*, i.e. the minimizer of the inner constrained optimization problem (27) via

$$\theta^*(\phi) = \mu^{-1} \phi' - B(\phi) = \mu^{-1} \phi' - \nu \partial_x^2 \phi + \phi \partial_x \phi \,, \quad (44)$$

where the last equality holds for the Burgers example with spatio-temporal white noise.

In practice, equation (40) has to be rewritten in order to be practical for the SPDE case, because the involved high spatial derivatives come with Courant-Friedrichs-Levy (CFL[20]) stability conditions that limit the time-step of the descent, and therefore the convergence rate of the scheme. A detailed discussion of tricks and optimizations for the SPDE case is given in[16].

D. Example: The stochastic Burgers-Huxley equation

As example for a nonlinear SPDE, we consider the stochastic Burgers-Huxley model,

$$\partial_t u + \alpha u \partial_x u - \kappa \partial_x^2 u = f(u) + \sqrt{\varepsilon} \eta(x,t) \,, x \in [0,1] \,, \quad (45)$$

where $\alpha > 0$ determines the strength of the nonlinear advection term, $\kappa > 0$ is the diffusion constant, and
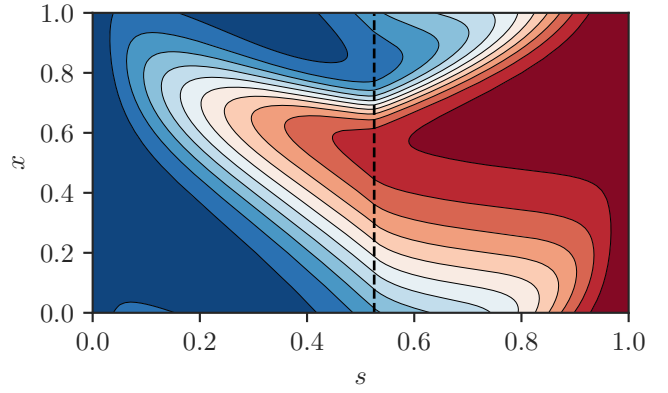

FIG. 3. Maximum likelihood transition pathway of the bi-stable Burgers-Huxley model, transitioning from $u = -1$ to $u = 1$. The transition happens as Allen-Cahn like nucleation, but the critical nucleus forms as steepening, asymmetric shock-wave. The saddle-point, denoting the critical nucleus of the transition, is marked by a dashed line.

the boundary conditions are periodic. The field $\eta(x,t)$ is spatio-temporal white noise. For $f(u) = 0$, this equation is the stochastic Burgers equation, arising in compressible gas dynamics, traffic flows, and as test-bed for turbulence. With the inclusion of a double-well reaction term $f(u) = u - u^3$, the equation becomes metastable, with two spatially homogeneous stable fixed points at $u = -1$ and $u = 1$. The spatially homogeneous solution $u = 0$ is a fixed point as well, but depending on the size of κ might not be a saddle point with a single unstable direction. Instead, for small enough κ, we expect Allen-Cahn like nucleation dynamics, but the nucleation must happen as a Burgers-like steepening shock wave. Indeed, as figure 3 shows, this intuition is confirmed by the numerics: The creation of the nucleus happens in a spatially asymmetric way, and the nucleating seed travels in space. The spatial resolution for the numerical computation is $N_x = 256$, while the temporal resolution is $N_s = 100$.

III. RARE EVENT ALGORITHMS FOR EXPECTATIONS AND EXTREME EVENTS

In section II we discussed how to compute noise-induced transition trajectories in bi-stable systems and thereby estimate the rate of transitions between the two metastable states and their relative likelihood. We chose to implement a global minimization procedure based on the Maupertuis principle form of the action functional to use the independence of the choice of parametrization to our numerical advantage. Nevertheless, because both the initial *and* the final conditions of the transition trajectory are fixed, we were unable to harness the Hamilton's equations of motion directly (these would have to be solved by shooting methods, which are inefficient or even ill-posed).

In this section, instead, we will concentrate on situations where it is indeed feasible to solve the minimization problem by integrating the coupled pair of equations of motion, or instanton equations, to obtain the large deviation minimizer. As we will see, if applicable, this approach comes with a couple of advantages of both theoretical and numerical nature. In this section we will therefore first review the class of algorithms based on solving Hamilton's equations in section III A, that can be used to compute instantons for expectations dominated by extreme events. Examples are shown in section III B applying this algorithm to a system with multiplicative Gaussian noise, and furthermore in section III C demonstrating the use in an infinite dimensional system, with the additional complication of degenerate forcing (i.e. non-invertible noise covariance matrix). We discuss connections to other fields in section III D and numerical details in section III E. A geometric variant of the numerical scheme is introduced in section III F, and implemented for an example case in section III G.

A. Instantons for expectations and extreme events

For the stochastic process X_t^ε of equation (1),

$$dX_t^\varepsilon = b(X_t^\varepsilon)\, dt + \sqrt{\varepsilon}\sigma\, dW_t\,,$$

consider the random variable $F(X_T^\varepsilon)$, where $F : \mathbb{R}^d \to \mathbb{R}$. This random variable, also termed the "observable", acts only on the final configuration of the process. We are interested in estimating the tail scaling of its probability density, i.e. in quantifying the likelihood of extreme values of the observable. For example, assume that X_t^ε is a stochastic model describing the interaction of predator and prey in a habitat (cf. section III B). We set out to find the probability of observing an abundance of prey. We might additionally be interested in the most likely amount of predators at this unusual configuration, and the historic development into this event. Or we have a stochastic description of waves (cf. section III C), and are interested in the probability of observing high amplitude waves. Additionally we might ask for the most likely shape of the wave at the moment of extreme elevation, or possibly identify the evolution into the extreme wave event to analyze it for possible mechanisms leading to the amplification.

From the discussion in section I we understand that in the limit $\varepsilon \to 0$, the probability of observing the event $F(X_T^\varepsilon) = z$, subject to $X_0^\varepsilon = x$, fulfills

$$P(z) \asymp \exp(-\varepsilon^{-1} \inf_{\phi \in \mathscr{C}_z} S_T(\phi))\,, \qquad (46)$$

where $\mathscr{C}_z = \{\phi \in AC([0,T],\mathbb{R}^d) \mid \phi(0) = x, F(\phi(T)) = z\}$, i.e. the set of continuous trajectories starting and x that observe the event. Let

$$I(z) = \inf_{\phi \in \mathscr{C}_z} S_T(\phi)\,, \qquad (47)$$

and define

$$I^*(\lambda) = \inf_{\phi \in \mathscr{C}} \left(S_T(\phi) - \lambda F(\phi(T)) \right)\,, \qquad (48)$$

with $\mathscr{C} = \{\phi \in AC([0,T],\mathbb{R}^d) \mid \phi(0) = x\}$. Here, minimization is not constrained at the final point, i.e. $\mathscr{C}$ describes the set of continuous trajectories starting at x regardless of their final point. We can then rewrite

$$\begin{aligned} I^*(\lambda) &= \inf_{\phi \in \mathscr{C}} \left(S_T(\phi) - \lambda F(\phi(T)) \right) \\ &= \inf_{z \in \mathbb{R}} \inf_{\phi \in \mathscr{C}_z} \left(S_T(\phi) - \lambda F(\phi(T)) \right) \\ &= \inf_{z \in \mathbb{R}} \left(\inf_{\phi \in \mathscr{C}_z} S_T(\phi) - \lambda z \right) \\ &= \inf_{z \in \mathbb{R}} \left(I(z) - \lambda z \right)\,. \end{aligned}$$

so that $I^*(\lambda)$ is the Fenchel-Legendre transform of $I(z)$. This connection allows us to solve the minimization problem (48) instead of the original problem (47). The same form can be obtained by considering λ as a Lagrange multiplier to enforce the constraint on the final point.

In terms of Hamilton's principle, the variations of the argument of the infimum in equation (48) with respect to ϕ gets one additional term that only applies at the final point, so that the Hamilton's equations become

$$\begin{cases} \dot{\phi} = \nabla_\theta H(\phi,\theta) & \phi(0) = x \\ \dot{\theta} = -\nabla_\phi H(\phi,\theta) & \theta(T) = -\lambda \nabla F(\phi(T))\,. \end{cases} \qquad (49)$$

The difference with Hamilton's equations of the problem discussed in section II

$$\begin{cases} \dot{\phi} = \nabla_\theta H(\phi,\theta) & \phi(0) = x,\ F(\phi(T)) = z \\ \dot{\theta} = -\nabla_\phi H(\phi,\theta) & \text{(no boundary conditions)} \end{cases} \qquad (50)$$

appears minuscule, but is profound: The ϕ-equation in (50) has to be solved with initial *and* final condition, and therefore necessitates shooting methods which are inefficient in high dimension (hence the alternative approach we took in section II). For (49), on the other hand, the equations for both ϕ and θ have exactly one boundary condition each. It is natural to integrate the ϕ-equation forward in time, starting at x, while integrating the θ-equation backward in time, starting at $-\lambda \nabla F(\phi(T))$. This direction of integration is the only sensible one in the first place: Due to the conjugate momentum equation containing the term $-(\nabla b(\phi))^\top$, a numerical integration forward in time would be numerically unstable or even ill-posed. An algorithm to find the instanton in this case then consists of the following steps: Starting from the k-th guess $\phi^k(t)$ for the instanton trajectory,

(i) solve the equation

$$\dot{\theta} = -\nabla_\phi H(\phi^k, \theta), \quad \theta(T) = -\lambda F(\phi^k(T)) \qquad (51)$$

backward in time,

(ii) solve the equation

$$\dot{\phi} = \nabla_\theta(\phi, \theta), \quad \phi(0) = x \qquad (52)$$

forward in time to obtain the next guess ϕ^{k+1},

(iii) iterate until convergence.

The convergence properties, stability and possible improvements of this algorithm are discussed in section III D. Considering the dual problem (48) instead of the original one (47) comes at a price: Instead of choosing directly the value z of the observable, instead we prescribe its dual λ, and obtain the corresponding value of z *a posteriori*. In other words, we loose the capability of computing the instanton for a specific observable z. In practice, this is usually not a problem, even though the map $z(\lambda)$ is not available in general: Typically one is interested in the complete distribution of $P(z)$, and therefore producing instantons for a whole range of λ similarly covers a whole range of z. Alternatively, a self-correcting version of the algorithm is easily implemented, where λ is adjusted on-the-fly to achieve the desired outcome z.

Note that $I^*(\lambda)$ is nothing but the limit of the scaled cumulant generating function of the random variable $F(X_T^\varepsilon)$, i.e.

$$I^*(\lambda) = \lim_{\varepsilon \to 0} \varepsilon \log \mathbb{E} \exp(\varepsilon^{-1} \lambda F(X_T^\varepsilon)). \qquad (53)$$

In this interpretation, we could call the instanton solving (49) also the instanton corresponding to the expectation

$$\mathbb{E} \exp(\varepsilon^{-1} \lambda F(X_T^\varepsilon)) \qquad (54)$$

in the limit $\varepsilon \to 0$. It is similarly possible to define observables not only on the final point of the trajectory, but for example of the form

$$F(\{X_t^\varepsilon\}) = \int_0^T f(X_t^\varepsilon)\, dt \quad \text{or} \quad F(\{X_t^\varepsilon\}) = \int_0^T \langle g(X_t^\varepsilon), dW \rangle \qquad (55)$$

and perform similar arguments, leading to additional drift terms in the conjugate momentum equation.

Finally, while the above arguments rigorously hold under suitable conditions in the limit $\varepsilon \to 0$, it is common to loosen conditions on the stochastic process and consider the case ε fixed, but $\lambda \to \infty$. The intuition is that for large λ only extreme events of the process are considered, and a large deviation principle might hold for the observable even for finite noise. One can then write down an *a priori* large deviation principle for the random variable $F(X_T^\varepsilon)$ and compute the instanton for large values of λ to probe the tail of the probability to observe the event. It is in this sense that this approach can be considered as instantons for *extreme events*. They are commonly used in practice, for example in fluid dynamics, where an equivalent algorithm has been introduced by Chernykh and Stepanov[21], which was applied to compute instantons for the Burgers[22,23], Navier-Stokes[24], and KPZ equations[25] as well as the integrated current of the periodic diffusion equation[26,27].
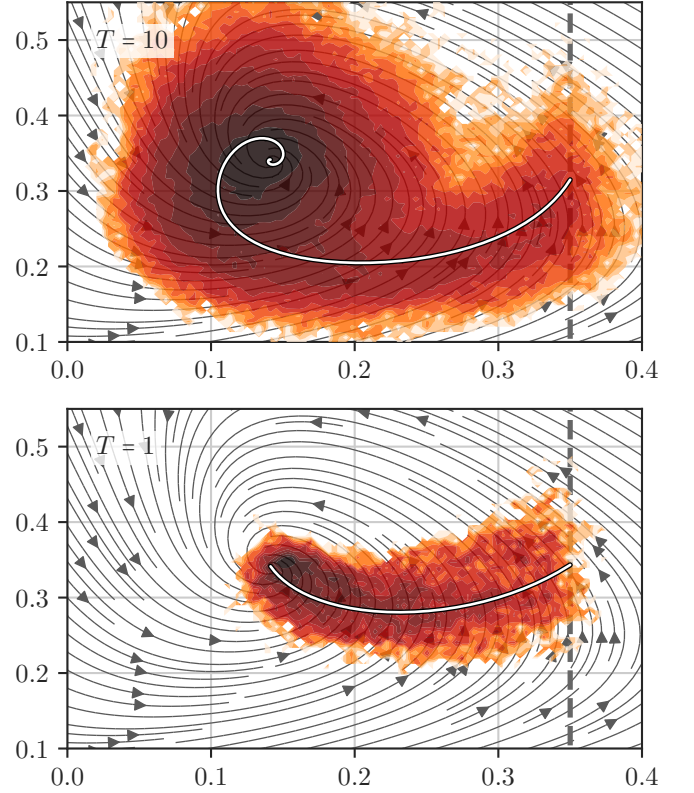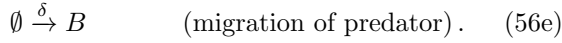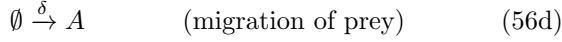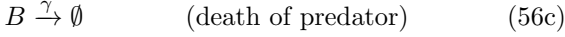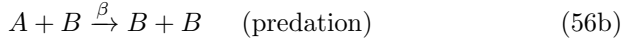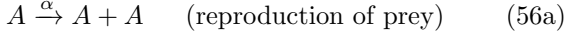

FIG. 4. Occurrence of extreme concentration of prey in the Lotka-Volterra model. The streamlines are showing the deterministic flow field. The heat-map shows the logarithm of a histogram of all trajectories starting at the fixed point that reach $a(T) = 0.35$ (regardless of $b(T)$). The white line depicts the instanton for the expectation $\mathbb{E} \exp(-\lambda a(T))$. Even for finite ε, the sample trajectories clearly cluster around the instanton. Shown are two different event times, $T = 10$ (top) and $T = 1$ (bottom).

B. Example: Extreme concentration of prey in the Lotka-Volterra model

The Lotka-Volterra system, or predator-prey system, is frequently used in biology as the simplest description of the interaction of two species, one of which preys on the other. In its typical form, it is considered without any fluctuations, but as it can be understood as continuous limit of a network of reactions, it is clear how a noise term can be derived as chemical Langevin equation.

To this end, consider a habitat with two species, the prey A and the predator B, where $A, B \in \mathbb{Z}^+$ denotes the number of individuals of the respective species. We want to consider interactions between individuals, modeled by

the stoichiometric reaction network

$$A \xrightarrow{\alpha} A + A \quad \text{(reproduction of prey)} \quad (56a)$$

$$A + B \xrightarrow{\beta} B + B \quad \text{(predation)} \quad (56b)$$

$$B \xrightarrow{\gamma} \emptyset \quad \text{(death of predator)} \quad (56c)$$

$$\emptyset \xrightarrow{\delta} A \quad \text{(migration of prey)} \quad (56d)$$

$$\emptyset \xrightarrow{\delta} B \quad \text{(migration of predator)}. \quad (56e)$$

Each of these is to be understood as a Poisson process with rates α to δ. The first three are standard in Lotka-Volterra, the last two are added to model migration of both species from neighboring habitats towards the considered location. This prevents degeneracy of the forcing at extinct population levels and the difficulty of absorbing boundary conditions.

Under the assumption that the typical populations N are sufficiently large, $\varepsilon = N^{-1} \to 0$, one can obtain a limiting behavior of the mean concentrations, and supplement it with Gaussian noise consistent with the central limit theorem around this mean. The resulting model, often termed *chemical Langevin equation*[28], that corresponds to the reaction network (56), then is

$$\begin{cases} da = (-\beta ab + \alpha a + \delta)\, dt + \sqrt{\varepsilon}\sqrt{\beta ab + \alpha a + \delta}\, dW_a\,, \\ db = (\beta ab - \gamma b + \delta)\, dt + \sqrt{\varepsilon}\sqrt{\beta ab + \gamma b + \delta}\, dW_b\,, \end{cases}$$
$$(57)$$

where a, b are functions from $[0, T]$ into $\mathbb{R}^+$, denoting the concentration of predator and prey in the habitat. The stochastic fluctuations are white-in-time Gaussian and zero mean, but notably multiplicative. Note that while this noise term chosen to be consistent with a central limit theorem for $N \to \infty$, it is actually not true that this approximation is valid for large deviations as well. In general, the LDP is sensitive to the non-Gaussian nature of the stochastic process defined in (56), which has Poisson statistics. We explain how to treat such non-Gaussian systems correctly in section IV, while here, for simplicity, we are considering the multiplicative Gaussian SDE (57) as given.

We choose the interaction rates $\alpha, \beta, \gamma,$ and δ in a way that there exists a unique fixed point $(\bar{a}, \bar{b})$ at which concentrations of predators and prey are in equilibrium. Concretely, we take $\alpha = 1$, $\beta = 5$, $\gamma = 1$, $\delta = 0.1$. Changing these parameters can produce more complicated attractors, such as limit cycles, which we will not investigate here. Instead, we are interested in the question of how unlikely high concentrations of prey develop on different time frames T when the system starts at the fixed point $(\bar{a}, \bar{b})$. For that reason, we choose $F(a, b) = a(T)$, i.e. condition on high values of $a(T)$, regardless of $b(T)$.

Since this is the first time we encounter multiplicative noise, a few comments are in order. For a system of the form

$$dX_t^\varepsilon = b(X_t^\varepsilon)\, dt + \sqrt{\varepsilon}\sigma(X_t^\varepsilon)\, dW_t\,, \quad (58)$$

with $a(x) = (\sigma\sigma^\top)(x)$, the Hamiltonian is

$$H(\phi, \theta) = \langle b(\phi), \theta \rangle + \tfrac{1}{2}\langle \theta, a(\phi)\theta \rangle\,, \quad (59)$$

so that an additional term enters the equation for the conjugate momentum,

$$\dot{\theta} = -\nabla_\phi H(\phi, \theta) = -(\nabla b(\phi))^\top \theta + \langle \theta, \nabla_\phi a(\phi)\theta \rangle\,, \quad (60)$$

where the last term is to be understood as $(\langle \theta, \nabla_\phi a(\phi)\theta \rangle)_i = \sum_{j,k} \theta_j \nabla_{\phi_i} a_{jk} \theta_k$. Consequently, the instanton equations for the (stochastic) Lotka-Volterra model are

$$\begin{cases} \dot{a} = -\beta ab + \alpha a + \delta + (\beta ab + \alpha a + \delta)\theta_a \\ \dot{b} = \beta ab - \gamma b + \delta + (\beta ab + \gamma b + \delta)\theta_b \\ \dot{\theta}_a = -(\alpha - \beta b)\theta_a - \beta b\theta_b + \tfrac{1}{2}((\alpha + \beta b)\theta_a^2 + \beta b\theta_b^2) \\ \dot{\theta}_b = \beta a\theta_a - (-\gamma + \beta a)\theta_b + \tfrac{1}{2}(\beta a\theta_a^2 + (\gamma + \beta a)\theta_b^2)\,, \end{cases}$$
$$(61)$$

which have to be solved with the boundary conditions $(a(0), b(0)) = (\bar{a}, \bar{b})$ and $(\theta_a(T), \theta_b(T)) = -\lambda \nabla F(a, b) = (-\lambda, 0)$.

Figure 4 shows the result of applying the algorithm of section III A to this system, and comparing to Monte-Carlo sampling. Here, two different transition times are chosen, $T = 1$ and $T = 10$. For $T = 10$, the system has enough time to explore around the fixed point, but it is obvious that the last portion of the excursion, before it hits $a(T) = 0.35$, clusters around the instanton trajectory. In particular, the b-coordinate at which $a = 0.35$ is attained seems to be predicted reasonably well. For $T = 1$, instead, the transition trajectory needs to follow a different route, and the endpoint $a = 0.35$ will most likely be attained at higher concentration of predators. Some points, such as $(a, b) = (0.25, 0.3)$, are almost never visited for $T = 10$, but are very likely under $T = 1$, which is correctly predicted by the instanton computation. Note that the heat-map depicting the empiric probability density in figure 4 has a logarithmic color-map for the tails to remain visible. Deviations from the optimal path are therefore very unlikely indeed.

Parameters for $T = 1$ are $\Delta t = 10^{-2}$, $\varepsilon = 0.005$ and $\lambda = 0.4209$, and for $T = 10$ are $\Delta t = 10^{-2}$, $\varepsilon = 0.004$, and $\lambda = 0.2106$. Roughly $5 \cdot 10^6$ trajectories were sampled for the Monte-Carlo estimate.

C. Example: Extreme amplitudes in the Korteweg-de Vries equation

Consider for the field $u(x, t) : [0, 2\pi] \times [0, T] \to \mathbb{R}$ the stochastic partial differential equation

$$\partial_t u + u\partial_x u + \kappa\partial_x^3 u - \nu\partial_x^2 u = \eta(x, t)\,, \quad u(x, t=0) = 0\,, \quad (62)$$

with periodic boundary conditions in space, and $x \in [0, 2\pi]$. This is a modification of the standard Korteweg-de Vries equation that describes the evolution of shallow water surface waves. To this, we added energy input
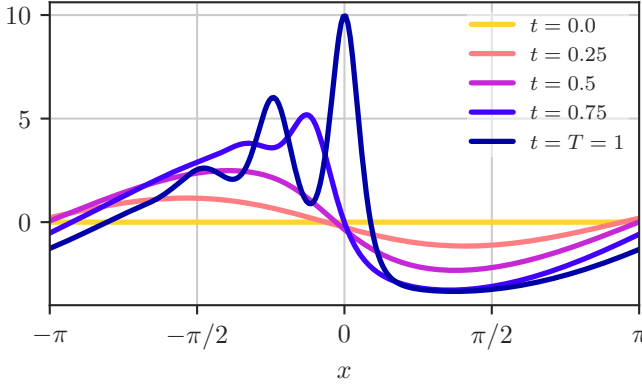
FIG. 5. Evolution of the Korteweg-de Vries instanton into a large amplitude at the final time $T = 1$, starting from rest and forcing only the largest Fourier mode of the system.

through the forcing η and energy dissipation through a diffusion term with viscosity ν. For the forcing, we demand that, in Fourier space,

$$\mathbb{E}\hat{\eta}_k(t)\hat{\eta}_q(t') = \varepsilon\delta(t - t')\hat{\chi}_{q-k}\,, \qquad (63)$$

where $\hat{\chi} : \mathbb{Z} \to \mathbb{R}$ is the forcing spectrum and $\hat{\eta}_k$ is the k-th mode of the Fourier transform of η.

Intuitively, for a $\hat{\chi}_k$ with compact support only for small k, the forcing $\eta(x, t)$ inserts energy on large scales, and the nonlinearity transfers those to smaller scales, on which dispersion and dissipation act on them. We are interested how this nonlinear cascading effect produces waves of extreme amplitude. To this effect, we choose an observable

$$F(u(x, T)) = (\phi_\Delta \star u)(x, T)\,, \qquad (64)$$

with $\phi_\Delta(x) = A\exp\left(-x^2/\Delta^2\right)$, and $\star$ denoting spatial convolution. For small Δ, this observable selects high amplitudes in close proximity to $x = 0$, i.e. at the center of the domain, and therefore generates high wave elevations at this position. As forcing spectrum, we want

$$\chi_k = \begin{cases} 1 & \text{if } |k| = 1 \\ 0 & \text{otherwise}\,, \end{cases} \qquad (65)$$

which inserts energy only into the largest mode of the system. This is the first time we consider *degenerate* forcing, in that only a subset of the available degrees of freedom are forced, or, equivalently, the noise covariance matrix a of (1) is not invertible. This poses practical problems for algorithms based on global minimization discussed in section II, where heavy use is made of either the a-norm, or θ is expressed as $\theta = a^{-1}(\dot{\phi} - b(\phi))$. For these algorithms, the inverse has to be replaced by the pseudo-inverse (see section 3.4 of[1]), and the degenerate forcing introduces additional stiff constraints for the unforced modes, as those effectively behave deterministically (and thus are attained with infinite action if they deviate from the deterministic behavior). For the Hamilton's equations, and

the algorithm discussed in this section, the noise correlation is never inverted, and degenerate forcing can be treated without extra effort.

The instanton equations corresponding to the posed problem are

$$\begin{cases} \partial_t u + u\partial_x u + \kappa\partial_x^3 u - \nu\partial_x^2 u = \chi \star \theta\,, & u(x, t=0) = 0\,, \\ \partial_t\theta + u\partial_x\theta + \kappa\partial_x^3\theta + \nu\partial_x^2\theta = 0\,, & \theta(x, t=T) = -\lambda\phi_\Delta(x)\,, \end{cases}$$
$$(66)$$

where $\chi(x)$ is the inverse Fourier transform of the forcing spectrum. The instanton computed by solving equations (66) is depicted in figure 5. It is clearly visible that a high final amplitude around $x = 0$ is achieved by a combination of non-linear advection and dispersion. Additionally, the final configuration clearly contains Fourier modes different from $|k| = 1$, implying that indeed the non-linearity cascaded energy into higher modes in a way to optimize the final amplitude. Note also that because of $\Delta \ll 1$, we are merely demanding a large amplitude at $x = 0$, but leave the rest of the wave form unconstrained. The elevation profile in the rest of the domain is chosen in a most likely manner, and the curves shown in figure 5 can be interpreted as the prototypical way of forming the considered amplitude. The model parameters are $T = 1$, $\Delta = 10^{-1}$, $\alpha = \kappa = 4 \cdot 10^{-2}$, $\lambda = 1$, and $A = 0.25$. The numerical parameters are $N_x = 256$, $N_t = 1000$, and $\Delta t = 10^{-3}$.

D. Connections to optimal control

It is instructive to formulate the optimization problem (46) in the language of optimal control[29]: We are interested in finding the *optimal control* $p : [0, T] \to \mathbb{R}^d$ such that for $X \in \mathbb{R}^d$, the system

$$\dot{X}(t) = b(X(t)) + p(t)\,, \quad X(0) = x\,, \qquad (67)$$

has the desired outcome, $F(X(T)) = z$. We penalize large values of p by choosing

$$J(p) = \tfrac{1}{2}\int_0^T |p(t)|^2\,dt \qquad (68)$$

as cost function. In other words, we are searching for the optimal noise realization p to drive the system into a final state where $F(X(T)) = z$. To obtain a minimization procedure that honors the constraints given by the observable and equation (67), we introduce Lagrange multipliers $\xi \in [0, T] \times \mathbb{R}^d$ and $\lambda \in \mathbb{R}$, such that we attempt to minimize

$$E(p) = \frac{1}{2}\int_0^T |p(t)|^2\,dt + \lambda F(X(T))$$
$$+ \int_0^T \langle\xi, \dot{X} - b(X) - p\rangle\,dt + \xi(0)(X(0) - x)\,.$$

Its total variation is given by

$$\delta E(p) = \langle p - \xi, \delta p \rangle + \langle \dot{X} - b(X) - p, \delta\xi \rangle + \langle X(0) - X, \delta\xi(0) \rangle$$
$$+ \lambda \langle \nabla F(X(T)), \delta X(T) \rangle + \langle -\dot{\xi} - (\nabla b(X))^\top \xi, \delta X \rangle$$
$$+ \langle \xi(T), \delta X(T) \rangle - \langle \xi(0), \delta X(0) \rangle \,.$$

We can read of the desired conditions to fulfill the constraints as

$$\begin{cases} \dot{X} = b(X) + p, & X(0) = x \\ \dot{\xi} = -(b(X))^\top \xi, & \xi(T) = -\lambda \nabla F(X(T)) \,, \end{cases} \quad (69)$$

and the gradient of the cost functional $E(p)$ with respect to the control p is then given as

$$\frac{\delta E(p)}{\delta p} = p - \xi \,. \quad (70)$$

We immediately identify that the conjugate momentum θ is the variable ξ in optimal control, often termed the *adjoint* variable. Second we realize that the forward and adjoint equations are identical to the instanton equations. Therefore, the iterative algorithm given in section III A is nothing but a gradient descent for the cost functional $E(p)$, with step length 1. This not only answer questions about (local) convergence of the algorithm of section II A, but furthermore allows to improve stability and order of convergence of the algorithm. First, it is almost always necessary to adjust the step size for each iteration according to a line search strategy to achieve convergence. Second, one might consider pre-conditioning, to allow for faster convergence. Lastly, the computation of the descent direction $-\delta E / \delta p$ from equation (70) allows to construct higher order optimization algorithms, such as nonlinear conjugate gradient or quasi-Newton methods.

Note that, similar to the argument above, for practical reasons we choose to not consider variations with respect to λ, and instead consider $\lambda \in \mathbb{R}$ given *a priori* to establish a mapping $\lambda(z)$ from $z(\lambda) = F(X^*(T))$, where $X^*(T)$ depends on λ through the boundary condition of the adjoint equation (69).

E. Improvements and implementation considerations

A few remarks are in order to point out possible improvements and implementation concerns when solving Hamilton's equations.

- While solving a global minimization problem as introduced in section II necessitates a complicated procedure to compute descent directions, the solution of Hamilton's equations put forward in this section usually comes at a much lower *implementation* cost: Given a stochastic problem at hand, one likely has already available an efficient solver of the forward equation, just replacing stochastic noise with a function of the conjugate momentum. The backward equation (auxiliary equation, adjoint

equation), on the other hand, is often available as well for professional software packages, usually from automatic differentiation, in order to quantify the uncertainty from the adjoint field ξ. In this case, a computation of the instanton might be achieved in a truly "black-box" form, where the iterative solution of the Hamilton's equations can be achieved without modifying the original software.

- The mutual dependency of the forward and backward equations necessitates in principle that the whole trajectory is stored in its entirety. While this is usually feasible for finite-dimensional problems, it quickly becomes prohibitive in terms of memory requirement when talking about SPDEs in many dimensions, where usually the storage requirements are chosen to be of the order of magnitude of a single state of the field variables, and not a continuous trajectory. In optimal control, this restriction is usually overcome via *checkpointing* mechanisms, where one does not save to memory a complete trajectory, but instead only retains checkpoints, from which subsequent states can be deduced. In its most efficient form, this checkpointing can be implemented in a recursive manner, leading generally to memory requirements of $O(\log N_t)$ instead of the naive $O(N_t)$, where N_t is the number of discretization points in time. Some details of the application of this method to instanton computations is laid out in[30].

- The nature of the conjugate momentum as adjoint variable, as discussed in section III D, highlights its *adjointness* to the main field variable. Would one discretize the action, and consider forward and backward equation as their discrete variation, this state of affairs would necessitate the usage of a special temporal integration scheme for the numerical solution of the adjoint equation, namely a temporal integration scheme that is adjoint to the forward scheme. Some time integration schemes (such as the forward Euler scheme) have the property of being self-adjoint, and thus can be used for both equations. The failure to use a correct pair of integration schemes generally results in the failure to converge to the minimum of the cost function.

A more detailed discussion of implementation concerns of this technique in the infinite-dimensional case, specifically with the application to fluid mechanics, is given in[24,30].

F. Geometric version of the Hamilton's equations

As discussed in section I, many questions, including the computation of the quasi-potential, necessitate a minimization not only over all possible paths $\phi(t)$, but also over all possible time intervals $T > 0$—for example,
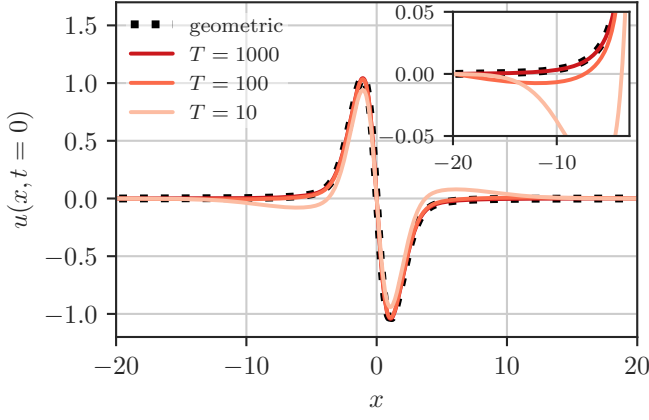
FIG. 6. Comparison of the final condition of the instanton, $u(x, t = 0)$, conditioning on extreme gradients in the origin, for the geometric parametrization and physical time parametrizations for varying T. As visible in the inset, only for $T = 1000$, secondary extrema disappear.

this is needed to calculate expectations with respect to the invariant measure of the process, assuming it exists. For algorithms based on the *minimum action method* (MAM)[14], the additional complication the minimization over $T > 0$ introduces is resolved by invoking Maupertuis principle and focusing on the computation of the *geometric* minimizer, i.e. realizing that the minimizing trajectory can be computed without explicit reference to its parametrization.

For algorithms based on the Hamilton's equation, similar ideas and extensions exists[31]: Instead of solving the original Hamilton's equations (16), we can choose a reparametrization $s(t)$, and consider Hamilton's equations in this parameter,

$$\begin{cases} \phi' = \mu \nabla_\theta H(\phi, \theta) \\ \theta' = \mu \nabla_\phi H(\phi, \theta), \end{cases} \quad (71)$$

where the prime denotes derivatives with respect to s and $\mu = dt/ds$. Now, as pointed out in section II A, μ can be interpreted as Lagrange multiplier enforcing the Hamiltonian constraint, and is available as

$$\mu = \frac{\|\phi'\|_\sim^2}{\langle \phi', \nabla_\theta H \rangle_\sim} = \frac{\|\phi'\|_a}{\|b(\phi)\|_a}, \quad (72)$$

where the much simpler second form only holds for Gaussian additive noise of the form (1) (compare section II A).

G. Example: Extreme gradients of the stochastic Burgers equation

As an example, following[31], consider for the field $u(x, t) : [-L/2, L/2] \times [-T, 0] \to \mathbb{R}$ the stochastic Burger's equation,

$$\partial_t u + u \partial_x u - \nu \partial_x^2 u = \eta(x, t), \quad u(x, t = -T) = 0, \quad (73)$$

with periodic boundary conditions in space. Here, we consider a noise term η that is white in time, but has a finite correlation length in space,

$$\mathbb{E}\eta(x, t)\eta(x', t') = \varepsilon \delta(t - t')\chi(x - x'), \quad (74)$$

and we prescribe the specific correlation in Fourier space of

$$\hat{\chi}(k) = k^2 \exp(-k^2/2)\mathcal{H}(k_c - |k|),$$

where $\mathcal{H}$ denotes the Heaviside step function. In effect, the forcing correlation has the shape of a "Mexican hat" function, with cut-off wave number k_c. Effectively, equation (73) can be considered as a test-bed for turbulence, where energy is inserted on large scales due to the forcing, and then cascades to smaller scales via the nonlinearity, where it dissipates. We focus on events that lead to a strong negative gradient at the final time, $t = 0$, and therefore choose an observable

$$F(u(x, 0)) = (\phi_\Delta \star \partial_x u)(x, 0), \quad (75)$$

where, identically to the KdV-case in (64), ϕ_Δ mollifies on scales Δ, so that here we concentrate on high gradients at the origin.

In order to probe for events on the invariant measure of (73), we want to consider the limit $T \to \infty$, and therefore need to either consider extremely large time intervals T, or alternatively employ the geometric variant of the Hamiltonian formalism as proposed in (71), where the norm is induced by the covariance $\chi(x)$, i.e. $\|v\| = \langle v, \chi^{-1}(x)v \rangle^{1/2}$ on its support. For technical details, see[31]. Given this setup, figure 6 compares the final condition of the instanton between finite times T and the limit $T \to \infty$ obtained in the geometric formalism. It demonstrates how, for choices $T = 10$ or $T = 100$, unphysical secondary maxima are present (compare inset of figure 6), that disappear in the infinite time case, and with $T = 1000$. Similarly, one can look at the value of the Hamiltonian $H(u, p)$, which necessarily disappears for large T, $\lim_{T \to \infty} H = 0$, and consider the quantity $\max(|H|)$ along the instanton trajectory as measure of the numerical error of the discretization. In this metric, the geometric variant for the same number of discretization points in time, is roughly 10^4 times more accurate than the naive parametrization with physical time[31].

IV. GENERALIZATIONS TO THE NON-GAUSSIAN CASE

In all above considerations and examples, we always considered the presence of an LDP for an SDE, either with additive or with multiplicative Gaussian noise that is temporally white. Generalizations to colored noise are considered for example in[32,33]. Here, instead, we consider the case of temporally white noise that is not necessarily Gaussian. It is treatable with algorithms of the

above form, as can be guessed from the fact that both MAM-style global minimization of section II and the algorithms based on the solution of Hamilton's equations of section III are written in terms of a generic large deviation Hamiltonian. In this section, we therefore intend to broaden the scope by demonstrating how more generic large deviation Hamiltonians are obtained, and corresponding instantons can be computed with the above algorithms. In particular, the Gaussianity of the underlying stochastic process is reflected in the fact that the Hamiltonian is *quadratic* in its conjugate momentum. Other processes, most notably those that result from limits of continuous time Markov jump processes (MJPs), generally lead to a non-quadratic Hamiltonian.

A. Large deviation principles as WKB approximation

Consider a homogeneous continuous time Markov jump process X_t, $t \in [0,T]$, with state space $\mathcal{E}$. The process is completely characterized by its generator $\mathcal{L}$, which allows us to write down its backward Kolmogorov equation (BKE) as

$$\partial_t f + \mathcal{L}f = 0, f(T) = \phi \tag{76}$$

for

$$f(T - t, n) = \mathbb{E}^n \phi(X_t), \tag{77}$$

where $n \in \mathcal{E}$, $f : [0,T] \times \mathcal{E} \to \mathbb{R}$, and $\phi : \mathcal{E} \to \mathbb{R}$. For concreteness, consider as a state space a counting space $\mathcal{E} = \mathbb{Z}_+^N$ for $N \in \mathbb{N}$ different species, where individuals of each species can interact via $R \in \mathbb{N}$ independent Poisson processes, manipulating the number of individuals of each species. For each of the R different interactions, the number of individuals changes from n to $n + \nu_r$, where $\nu_r \in \mathbb{Z}^N$, with a rate $\tilde{a}_r(n)$, $r \in \{1, \dots, R\}$. Then, the corresponding generator reads

$$(\mathcal{L}f)(n) = \sum_{r=1}^{R} \tilde{a}_r(n)\left(f(n + \nu_r) - f(n)\right). \tag{78}$$

Rescaling this into new variables $x = n/M$, where M is a typical number of individuals, we can expand in $\varepsilon = M^{-1}$ to obtain a large deviation principle in the limit of many individuals. To this end, consider the generator in the rescaled variables,

$$(\mathcal{L}^\varepsilon f)(x) = \sum_{r=1}^{R} a_r(x)\left(f(x + \varepsilon\nu_r) - f(x)\right), \tag{79}$$

where a_r is defined on the rescaled variables. We now evaluate this rescaled generator onto a function of the form $\exp(\varepsilon^{-1}g(x))$ and rescale time with ε appropriately. This corresponds to a WKB approximation of the BKE, or equivalently to the method of Feng and Kurtz[34] (compare also[35]), and yields

$$\partial_t g(x) + \sum_{r=1}^{R} a_r(x)\left(e^{\varepsilon^{-1}(g(x+\varepsilon\nu_r)-g(x))}\right) = 0, \tag{80}$$
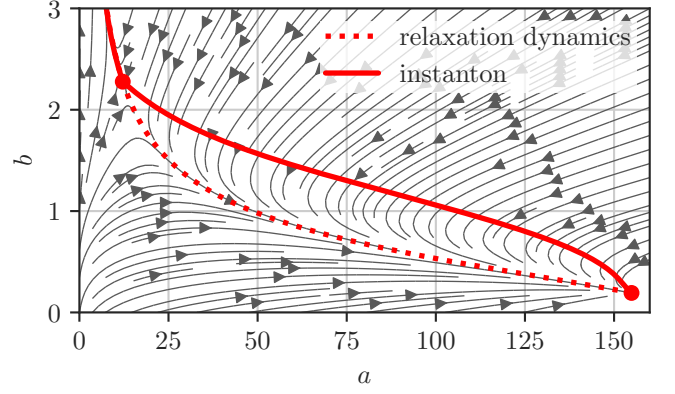

FIG. 7. Instanton for the non-Gaussian genetic switch. The arrows denote the direction of the deterministic flow, the red solid line depicts the minimizer, the red dashed line the relaxation path from the saddle. Red dots are located at the fixed points (stable and unstable). The whole figure is a zoom into the uphill region, the other stable fixed point is far up the upper left corner.

which can be expanded, to leading order in ε, into

$$\partial_t g(x) + \sum_{r=1}^{R} a_r(x)\left(e^{\langle \nu_r, \nabla g(x)\rangle} - 1\right) = 0. \tag{81}$$

Equation (81) can be interpreted as Hamilton-Jacobi equation

$$\partial_t g(x) + H(x, \nabla g(x)) = 0, \tag{82}$$

with

$$H(x, \theta) = \sum_{r=1}^{R} a_r(x)\left(e^{\langle \nu_r, \theta\rangle} - 1\right). \tag{83}$$

The Hamiltonian (83) is precisely the large deviation Hamiltonian in that the large deviation rate function is the time integral of its Fenchel-Legendre transform. The Hamiltonian (83) is furthermore a prime example of a non-quadratic Hamiltonian.

Note that applying the same method to the generator of the SDE (1),

$$(\mathcal{L}^\varepsilon f)(x) = \langle b(x), \nabla\rangle f(x) + \tfrac{\varepsilon}{2}\sum_{i,j=1}^{d} a_{ij}\nabla_i\nabla_j f(x) \tag{84}$$
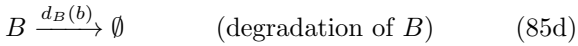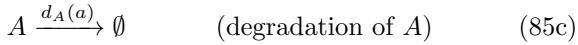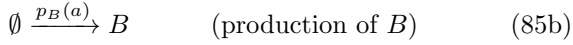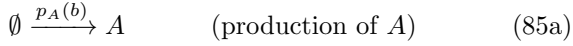
recovers exactly the expected Hamiltonian (17) for the leading order in ε.

B. Example: Genetic switch

As an example for instantons of a non-Gaussian LDP for a continuous time MJP of the form of section IV A, we

want to consider a simplified model of a genetic switch based on a model discussed in[36]: Inside a bacterium, plasmids contain genes that encode two different proteins, A and B. Each protein is able to form polymers that inhibit the production of the other protein, respectively. In the emerging situation, the cell may exist in a state close to one of two fixed points: Either protein A dominates, and inhibits the production of protein B, while the production of A remains high. Alternatively, protein B dominates, and inhibits the production of A. Rarely, fluctuations may arise that push the system from one fixed point to the other.

We choose to model this system by describing the concentrations of proteins A and B by $a \in \mathbb{R}_+$ and $b \in \mathbb{R}_+$, respectively. There are four reactions in total, namely production and degradation of A and B, leading to the reaction network

$$\emptyset \xrightarrow{p_A(b)} A \qquad \text{(production of } A) \tag{85a}$$

$$\emptyset \xrightarrow{p_B(a)} B \qquad \text{(production of } B) \tag{85b}$$

$$A \xrightarrow{d_A(a)} \emptyset \qquad \text{(degradation of } A) \tag{85c}$$

$$B \xrightarrow{d_B(b)} \emptyset \qquad \text{(degradation of } B) \tag{85d}$$

with rates

$$\begin{aligned} p_A(b) &= C/(1+b^3), \quad p_B(a) = D/(1+a), \\ d_A(a) &= a, \qquad\qquad d_B(b) = b\,. \end{aligned} \tag{86}$$

The corresponding large deviation Hamiltonian, using equation (81), is thus

$$\begin{aligned} H(a,b,\theta_a,\theta_b) &= \frac{C}{1+b^3}(e^{\theta_a}-1) + a(e^{-\theta_a}-1) \\ &+ \frac{D}{1+a}(e^{\theta_b}-1) + b(e^{-\theta_b}-1)\,. \end{aligned} \tag{87}$$

The minimizer for this setup, as well as the relaxation paths from the saddle, are depicted in figure 7. They are computed by implementing the algorithm presented in section II A for the Hamiltonian (87) for the transition between the two fixed points. The model parameters here are chosen to be $C = 156$ and $D = 30$.

V. SYSTEMS WITH RANDOM PARAMETERS AND EXTREME EVENTS

Up to now, all discussions in the previous sections concerned sample path large deviations, where a stochastic process realizes a rare event almost surely by following a trajectory that minimizes the corresponding action functional. In this section, we are focusing on a related, but different setup of a dynamical system with random parameters. Given a distribution of the random parameters, we want to reason about probabilities to observe certain events, and again characterize the rare ones by

dominating configurations of parameters. To this effect, consider for $u : [0, T] \to \mathbb{R}^d$ the dynamical system

$$\partial_t u = b(u, \theta), \quad u(t=0) = u_0(\theta)\,, \tag{88}$$

where $\theta \in \Omega \subseteq \mathbb{R}^M$ is the set of M random real parameters, distributed according to a measure μ. Since the initial conditions and the drift of equation (88) depend on the random parameters, the solution is a random variable, denoted by $u(\cdot, \theta)$. We can then try to quantify the probability of an observable exceeding a threshold $z \in \mathbb{R}$, for example at the final time, or integrated over time, or as temporal maximum, i.e.

$$P_T(z) = P(F(\theta) \geq z)\,, F(\theta) = \begin{cases} f(u(T,\theta)) & \text{or} \\ \int_0^T f(u(t,\theta))\,dt & \text{or} \\ \max_{0 \leq t \leq T} f(u(t,\theta))\,. \end{cases} \tag{89}$$

A. Large deviations for systems with random parameters

Indeed, if in the limit of large z this probability becomes small, $\lim_{z\to\infty} P_T(z) = 0$, then under some additional assumptions on $F : \Omega \to \mathbb{R}$ we have a large deviation principle as a contraction principle of the form

$$P_T(z) \asymp \exp(-\min_{\theta \in \Omega(z)} I(\theta))\,, \tag{90}$$

in the limit $z \to \infty$, where the set $\Omega(z) \subseteq \Omega$ is the set of permissible random parameters,

$$\Omega(z) = \{\theta \in \Omega | F(\theta) \geq z\}\,, \tag{91}$$

and the rate function $I(\theta)$ is obtained as the Legendre transform of the cumulant generating function of θ,

$$I(\theta) = \max_\eta (\langle \eta, \theta \rangle - S(\eta))\,, \tag{92}$$

for

$$S(\eta) = \log \mathbb{E} \exp\langle \eta, \theta \rangle = \log \int_\Omega \exp\langle \eta, \theta \rangle d\mu(\theta)\,. \tag{93}$$

The minimizer of $I(\theta)$ in $\Omega(z)$, i.e.

$$\theta^*(z) = \underset{\theta \in \Omega(z)}{\operatorname{argmin}} I(\theta)\,, \tag{94}$$

dominates the occurrence of the event, and is termed the *dominating point*[37]. Since we are considering only finite dimensional Ω, the corresponding optimization problem (94) has to be solved in a generally smaller search space, compared to the pathwise LDPs considered in the earlier sections. Equivalently, here, the instanton is not a preferred trajectory of the system, but instead the maximum likelihood set of parameters that lead to the event. Of course, given any $\theta \in \Omega$, there is a unique trajectory

$u(\theta)$ solving (88) associated to it, so that $u(\theta^*)$ represents the most likely trajectory of the system to realize the rare event. The proof of the large deviation principle (90) is carried out in[38], where sufficient assumptions on the behaviour of the function F and the geometry of the sets $\Omega(z)$ are specified, which is based on the notion of dominating points considered in[37].

From a numerical perspective, we can again solve the constrained optimization problem (94) by instead considering a Lagrange multiplier $\lambda \in \mathbb{R}$. For example, for the first of the three cases in (89), we attempt to minimize the objective function

$$E(u(T,\theta),\theta) = I(\theta) - \lambda f(u(T,\theta)). \qquad (95)$$

Via the Jacobian $J_{ij}(t) = \partial u_i(t)/\partial \theta_j$, i.e. the variation of the current configuration with respect to the random parameters, one can express the gradient as

$$\nabla_\theta E = \nabla_\theta I - \lambda J^\top(T,\theta)\partial_u f(u(T,\theta)), \qquad (96)$$

where the Jacobian is available through

$$\partial_t J = (\partial_u b)J + \partial_\theta b, \quad J(0) = \partial_\theta u_0. \qquad (97)$$

While integrating the forward equation (88) and the Jacobian (97) allows us to evaluate the gradient for a given θ, note that $J : [0,T] \to \mathbb{R}^{d\times d}$ is quite costly to compute. Instead, we can again fall back to an adjoint formulation to overcome this limitation. To this effect, consider the adjoint field $\mu : [0,T] \to \mathbb{R}^d$, subject to the adjoint equation

$$\partial_t \mu = -(\partial_u b)^\top \mu, \ \mu(T) = \lambda \partial_u f(u(T,\theta)). \qquad (98)$$

Since, using equations (97) and (98), we have

$$\partial_t(J^\top \mu) = (\partial_\theta b)^\top \mu,$$

it follows that

$$\int_0^T (\partial_\theta b)^\top \mu \, dt = (J^\top \mu)\big|_0^T = \lambda J^\top(T)\partial_u f(u(T)) - (\partial_\theta u_0)^\top \mu(0),$$

and thus, the gradient (96) is computable without referring to the Jacobian as

$$\nabla_\theta E = \nabla_\theta I - (\partial_\theta u_0)^\top \mu(0) - \int_0^T (\partial_\theta b)^\top \mu \, dt. \qquad (99)$$

In total, the gradient of the objective function (95) can be computed at a given θ in three steps:

(i) Integrate the forward equation,

$$\partial_u = b(u,\theta), \quad u(0) = u_0(\theta),$$

(ii) Compute the adjoint field μ by integrating

$$\partial_t \mu = -(\partial_u b)^\top \mu, \quad \mu(T) = \lambda \partial_u f(u(T)),$$
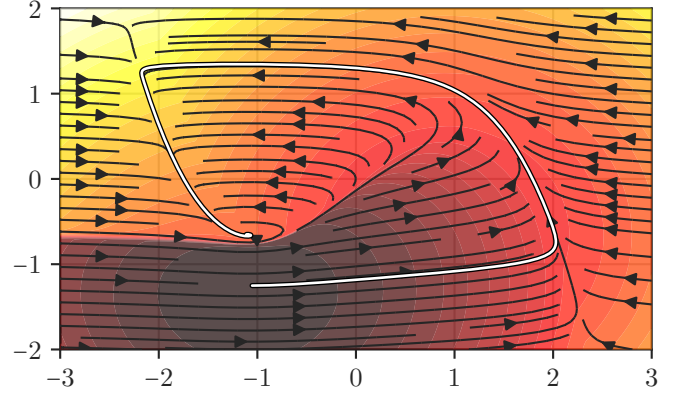

FIG. 8. Optimal perturbation of the initial condition to achieve an extreme excursion in the Fitzhugh-Nagumo model (100). The flow field denotes the drift term, and the color denotes the value of the objective function. The trajectory realizing the maximal excursion is indicated as white line.

(iii) and finally, compute the gradient

$$\nabla_\theta E(\theta) = \nabla_\theta I - (\partial_\theta u_0)^\top \mu(0) - \int_0^T (\partial_\theta b)^\top \mu \, dt.$$

A few comments are of note:

- In contrast to the sections before, in the current setup we are not considering the case of small noise. Instead, the fluctuations are held at fixed amplitude, but we consider events in the limit of infinite threshold, $z \to \infty$. It is in this limit that the LDP in (90) is obtained, which might lead to a non-standard large deviation speed as a consequence. Therefore, the LDT computation discussed in this section can truly be considered for an *extreme event instanton.*

- Again, the formulation in the form of an adjoint equation (98) is often beneficial from an implementation perspective as well: Many software packages for complex systems contain the computation of the adjoint field. Therefore, the computation of the gradient can possibly be achieved in a black-box manner.

B. Example: Optimal excitation of the Fitzhugh-Nagumo model

As an example, consider the following version of the deterministic Fitzhugh-Nagumo model,

$$\begin{cases} \dot{x} = \nu^{-1}(x - \frac{1}{3}x^3 - y) \\ \dot{y} = x + a. \end{cases} \qquad (100)$$

If we consider the case $a > 1$, this model is an excitable system, in that there is a unique fixed point $(\bar{x}, \bar{y}) = (-a, \frac{1}{3}a^3 - a)$, but small perturbations out of this fixed point potentially lead to large excursions until the system returns to its steady state. Here, we are interested in the optimal perturbation of the initial condition away from the fixed point to achieve a large excursion. For $\theta \in \Omega = \mathbb{R}^2$, we define the distribution of initial conditions as Gaussian centered around the fixed point,

$$(x_0(\theta), y_0(\theta)) = \theta \sim \exp\left(\frac{1}{2\Delta}\left((x_0 - \bar{x})^2 + (y_0 - \bar{y})^2\right)\right),$$

and take as observable

$$F(\theta) = \max_{t \in [0,T]} x(t, \theta), \tag{101}$$

i.e. the maximal excursion in the x-component of the trajectory. We want to know $P(z) = P(F(\theta) \geq z)$, and the corresponding most likely initial conditions (and trajectory) that realize this extreme event. Note that (101) is an observable of the third form of (89), and the algorithm lined out above has to be modified slightly. In particular, for $u(t) = (x(t), y(t))$ we obtain the gradient $\nabla_\theta E$ from forward and backward equations, which are, respectively,

$$\begin{cases} \partial_t u = b(u), & u(0) = (x_0, y_0) \\ \partial_t \mu = -(\partial_u b)^\top \mu, & \mu(t^*) = \partial_u f(u(t^*)) = (1, 0), \end{cases} \tag{102}$$

where $t \in [0, t^*]$ and t^* is the time at which the maximum is reached. We are considering only the first local maximum in time, but the dynamics of (100) are such that the first local maximum necessarily is the global maximum. In this situation, there is no additional term coming from the dependence of t^* on the random parameters, since

$$\nabla_\theta f(u(t^*(\theta), \theta)) = \partial_u f(u(t^*, \theta))\partial_\theta u(t^*, \theta) \\ + \partial_u f(u(t^*, \theta))\partial_t u(t^*)\partial_\theta t^*,$$

and the second term disappears because at t^* we have $\partial_t u(t^*) = 0$. The gradient can then be computed as

$$\nabla_\theta E = \frac{1}{\Delta}(\theta - \bar{u}) - \lambda \mu(0)$$

for $u_0(\theta) = \theta \in \mathbb{R}^2$ and $\bar{u} = (\bar{x}, \bar{y})$.

Figure 8 shows the result of the minimization procedure: For $\Delta = 2, \lambda = 1, a = 1.1, \nu = 10^{-1}$, the shading indicates the objective function (95) for every initial condition. One can clearly make out the jump in the objective function across the separatrix, where trajectories start exhibiting large excursions. The trajectory starting at the minimum is the one that maximizes the excursion in x-direction, before it decays to the fixed point. The streamlines show the dynamics of the Fitzhugh-Nagumo model (100).

VI. INSTANTONS AS PART OF OTHER RARE EVENT ALGORITHMS

While instantons as prototypical realizations of rare events can be used for their own sake to estimate probabilities, relative stability, and transition mechanisms, they can also be helpful as ingredient to increase efficiency of other types of rare event algorithms. Most notably, whenever rare events are sampled numerically by *tilting* a given stochastic process to facilitate a rare event in an importance sampling setup, the instanton can be considered as the *optimal tilt* in the small-noise limit. Equivalently, this tilting can be interepreted as a (generalized) Doob transform[39], yielding a process conditioned on a rare outcome. In this section, we want to discuss how this can be achieved in practice, and common pitfalls of this strategy: First, in section VIA, we will show how to use instantons to perform importance sampling for Monte-Carlo methods. In section VIB, we use instantons to construct weighting functions for genealogical particle algorithms. This will be accompanied by an example computing the probability of infection rates in a stochastic model for epidemiology.

A. Instantons for importance sampling

Consider an expectation of the form

$$A^\varepsilon = \mathbb{E}\exp(\varepsilon^{-1}F(X_T^\varepsilon)) \tag{103}$$

for a random process $X_t^\varepsilon \in \mathbb{R}^d$, for example the one obeying an SDE like (1). We saw in section III how to compute the corresponding instanton and get the dominating contribution in the limit $\varepsilon \to 0$. In order to get hold of a proper quantitative estimate of (103), though, one would naively consider a Monte Carlo method with estimator

$$\delta_\varepsilon = \frac{1}{M}\sum_{i=1}^{M}\exp(\varepsilon^{-1}F(X_T^{\varepsilon,i})), \tag{104}$$

where $\{X_t^{\varepsilon,i}\}_{i=1}^N$ are N independent realizations of the process. This estimator is unbiased, meaning that

$$\mathbb{E}\delta_\varepsilon = A^\varepsilon.$$

The relative error of this estimator,

$$e(\delta_\varepsilon) = \frac{\text{std}(\delta_\varepsilon)}{\text{mean}(\delta_\varepsilon)} = \frac{1}{\sqrt{M}}\sqrt{\frac{\mathbb{E}\exp(-2\varepsilon^{-1}F(X_T^\varepsilon))}{\left(\mathbb{E}\exp(-\varepsilon^{-1}F(X_T^\varepsilon))\right)^2} - 1} \tag{105}$$

describes the relative variance of the estimator. For example, for $e(\delta_\varepsilon) = 1$, the typical fluctuations of the estimate are of the size of the estimated value itself. The goal is to achieve a small relative error. In practice, for rare events, one often struggles to even keep $e(\delta_\varepsilon)$ bounded for $\varepsilon \to 0$: Even though increasing the number of samples improves the quality of the estimate, $e(\delta_\varepsilon) \to 0$ for

$M \to \infty$, the relative error increases exponentially for fixed M as $\varepsilon \to 0$. As a consequence, estimating rare events with the naive estimator (104) is impractical as the relative variance blows up. The standard answer to this problem is to employ *importance sampling*, i.e. introducing a new process Y_t^ε under which the rare event becomes typical, but accounting for this change of probability measure by correcting with the proper Girsanov-factor. Indeed, considering

$$dY_t^\varepsilon = (b(Y_t^\varepsilon) + \sigma v(t, Y_t^\varepsilon)) \, dt + \sqrt{\varepsilon}\sigma \, dW_t \,, \qquad (106)$$

for some function $v : [0, T] \times \mathbb{R}^d \to \mathbb{R}^d$, we can express the expectation (103) as

$$A^\varepsilon = \mathbb{E}^x \exp(-\varepsilon^{-1} F(Y_T^\varepsilon)) M_T^\varepsilon \,, \qquad (107)$$

where

$$M_T^\varepsilon = \exp\left(-\frac{1}{\sqrt{\varepsilon}} \int_0^T \langle v(s, Y_s^\varepsilon), dW_s \rangle - \frac{1}{2\varepsilon} \int_0^T |v(s, Y_s^\varepsilon)|^2 \, ds\right). \qquad (108)$$

This identity can be used to construct an unbiased estimator of A^ε by replacing the expectation in (107) by an empirical expectation over M independent copies of Y_t^ε, similar to what was done to obtain the estimator (104). The question is how to best choose the importance sampling bias $v(t, x)$ to lower the variance of this new estimator. An intuitive idea would be to use the instanton to do so[40,41]. For example, it has been suggested to take

$$v(t, x) = \sigma^{-1}(\dot{\phi}(t) - b(x)) \qquad (109)$$

or alternatively

$$v(t, x) = \sigma^T \theta(t) \,, \qquad (110)$$

where $(\phi(t), \theta(t))$ is the instanton position and momentum corresponding to the expectation (103), i.e. taking Y_t^ε to be, respectively, the stochastic process

$$\begin{aligned} dY_t^\varepsilon &= \dot{\phi}(t) \, dt + \sqrt{\varepsilon}\sigma \, dW_t \qquad \text{or} \\ dY_t^\varepsilon &= b(Y_t^\varepsilon) \, dt + a\theta(t) \, dt + \sqrt{\varepsilon}\sigma \, dW_t \,. \end{aligned} \qquad (111)$$

The intuition is that using either one of the processes in (111) biases the dynamics towards the dominating path $\phi(t)$. The technique to consider biased dynamics, by the instanton or otherwise, to increase the frequency of rare events, was considered already in[42], and is frequently used in the literature to create efficient rare event algorithms[43,44]. Similar ideas, inspired from lattice quantum chromodynamics, have entered through stochastic field theory to bias Monte Carlo methods with the knowledge of the instanton. For example in[45] the instanton for the stochastic Burgers equation is used precisely in the way of (111) to sample a modified process describing the fluctuations around it, getting improved statistics in the rare event regime.

Although it has been pointed out that this strategy does not succeed to decrease variance in general, or might even perform worse than the naive one in the limit as $\varepsilon \to 0$[46], it can be modified[47] to achieve optimal variance decay by recomputing the instanton trajectory on-the-fly. Notably, though, counter-examples that lead to worse relative error are specifically crafted, while in general performing importance sampling guided by the instanton trajectory works well in practical applications.

B. Instantons for cloning algorithms

There is another way to incorporate knowledge of the instanton within importance sampling, namely through algorithms of genealogical type[48–51]. In these methods, an ensemble of trajectories (aka particles, copies, or clones) is integrated, and particles are removed or duplicated according to some rating that selects behaviors favorable to the event at hand. To explain how this can be done in the context of rare event algorithms, let us focus on the choice (110) for $v(t, x)$, since out of the two it is the one that requires the least modification of the drift[52]. The second equation in (111) reads

$$dY_t^\varepsilon = b(Y_t^\varepsilon) \, dt + a\theta(t) \, dt + \sqrt{\varepsilon}\sigma \, dW_t \qquad (112)$$

along with the estimator for (103)

$$A^\varepsilon = \mathbb{E}^x \exp\left(-\varepsilon^{-1} F(Y_T^\varepsilon) - \frac{1}{\sqrt{\varepsilon}} \int_0^T \langle \theta(t), \sigma dW_t \rangle - \frac{1}{2\varepsilon} \int_0^T \langle \theta(t), a\theta(t) \rangle \, dt\right). \qquad (113)$$

We begin by rewriting this last formula in a form that is more convenient for resampling. To this end, let us integrate the following identity

$$d\langle \theta(t), Y_t^\varepsilon - \phi(t) \rangle = \langle \dot{\theta}(t), Y_t^\varepsilon - \phi(t) \rangle dt + \langle \theta(t), b(Y_t^\varepsilon) - b(\phi(t)) \rangle dt + \sqrt{\varepsilon}\langle \theta(t), \sigma dW_t \rangle \qquad (114)$$

to get

$$-\frac{1}{\sqrt{\varepsilon}} \int_0^T \langle \sigma^T \theta(t), dW_t \rangle = -\frac{1}{\varepsilon} \langle \theta(T), Y_T^\varepsilon - \phi(T) \rangle + \frac{1}{\varepsilon} \int_0^T \alpha(t, Y_t^\varepsilon) \, dt \,, \qquad (115)$$

where we defined

$$\alpha(t, x) = \langle \dot{\theta}(t), x - \phi(t) \rangle + \langle \theta(t), b(x) - b(\phi(t)) \rangle \,. \qquad (116)$$

These manipulations allow us to write the expectation (113) as

$$A^\varepsilon = \mathbb{E}^x \exp\left(-\varepsilon^{-1} F(Y_T^\varepsilon) - \varepsilon^{-1} \langle \theta(T), Y_T^\varepsilon - \phi(T) \rangle - \tfrac{1}{2}\varepsilon^{-1} \int_0^T \langle \theta(t), a\theta(t) \rangle dt\right) W_T^\varepsilon \,, \qquad (117)$$

where

$$W_t^\varepsilon = \exp\left(\frac{1}{\varepsilon}\int_0^t \alpha(s, Y_s^\varepsilon)ds\right). \qquad (118)$$

The expression (117) can be used to design a genealogical algorithm in which W_t^ε is viewed as a weight that each of the particles carries and according to which they are periodically resampled. More concretely, consider M copies of the system, all evolving according to the SDE (112). Denote by Y_t^i the position of the i-th copy at time t and by W_t^i its weight. The particle positions and weights are evolved independently on intervals (t_{k-1}, t_k), where t_k, $k \in \mathbb{N}_0$ with $t_{k-1} < t_k$ are *selection steps* when the resampling occurs. It proceeds as follows: Denoting by ΔW_k^i the weight accumulated by particle i on the interval (t_{k-1}, t_k), i.e.

$$\Delta W_k^i = \exp\left(\frac{1}{\varepsilon}\int_{t_{k-1}}^{t_k} \alpha(s, Y_s^i)ds\right), \qquad (119)$$

we compute

$$p_k^i = \frac{\Delta W_k^i}{\sum_{j=1}^N \Delta W_k^j}, \quad \Delta\bar{W}_k = \frac{1}{M}\sum_{i=1}^M \Delta W_k^i, \qquad (120)$$

and choose independently (with replacement) M copies in the set $\{Y_{t_k}^i\}_{i=1}^M$, using probability p_k^i to pick copy $Y_{t_k}^i$. We then use the resulting copies as new set $\{Y_{t_k}^i\}_{i=1}^M$, assign to each the same weight $W_{t_k}^i = \prod_{l=1}^k \Delta\bar{W}_l$, and repeat on the next interval (t_k, t_{k+1}).

As a result of this procedure, we have at any time a set of M copies with nearly uniform weights (since they only diverge from one another during the intervals (t_k, t_{k+1})), that provides us with the following expression for A^ε (compare with (117))

$$A^\varepsilon = \exp\left(-\tfrac{1}{2}\varepsilon^{-1}\int_0^T \langle\theta(t), a\theta(t)\rangle dt\right)\mathbb{E}^x\zeta_M \quad \text{with}$$

$$\zeta_M = \frac{1}{M}\sum_{i=1}^M \exp\left(-\varepsilon^{-1}F(Y_T^i) - \varepsilon^{-1}\langle\theta(T), Y_T^i - \phi(T)\rangle\right)W_T^i,$$

$$(121)$$

where $\mathbb{E}^x$ denotes expectation over both the noise term in (112) and the resampling steps. If M is large enough we can simply build an unbiased estimator for A^ε by using ζ_M directly (i.e, removing the expectation); or we can repeat the estimation R times with M copies and replace the expectation in (121) with an empirical average over the values of ζ_M calculated in these R runs.

The variance of the estimator based on (121) has been analyzed e.g. in[48,49], where other variants of the algorithm (e.g. in terms of the resampling step) are also discussed. Let us simply mention here that this estimator may not be better behaved than the one directly based on (117) (i.e. the one using no resampling based on the values of the weights), but it offers multiple possibilities of modifications that can systematically improve its variance—we refer the reader to[53] for more details.

Similar approaches are possible with adaptive multilevel splitting[49,54], where again a rating function ("reaction coordinate") has to be found to evaluate the performance of multiple copies, and for which the instanton dynamics can be taken as input.

C. Example: Epidemiology and vaccination at birth

To illustrate the scheme above, we consider the following compartmental model inspired from epidemiology, where the spread of a disease is modeled in the presence of vaccination. The total population of individuals, denoted by N, is comprised of individuals susceptible to the disease (S), individuals that are infected (I), individuals that are recovered and thus immune (R) and individuals that are vaccinated and thus immune (V). Individuals are born and die with the same rate μ, so that the total population remains constant. In this simple rendition of disease spread with vaccination, the vaccine is administered at birth, and with a vaccination rate of q. In other words, a child is born vaccinated with probability q or susceptible otherwise. The disease is transmitted by contact between infected and susceptible individuals, with contact rate β, while recovery is associated with the recovery rate γ. In total, the model therefore reads

$$\begin{cases} \dot{S} = \mu N(1-q) - \mu S - \beta N^{-1}IS \\ \dot{I} = \beta N^{-1}IS - (\mu + \gamma)I \\ \dot{V} = \mu Nq - \mu V \\ \dot{R} = \gamma I - \mu R. \end{cases} \qquad (122)$$

Interestingly, depending on the vaccination rate, the model (122) results in either total eradication of the disease after transient dynamics (disease free equilibrium), or a fixed point where the disease is still present (endemic equilibrium). More precisely, the reproduction number

$$R_0 = \frac{\beta}{\mu + \gamma},$$

describes the average number of contacts per infected individual (i.e. the ratio between contact frequency and the frequencies associated with recovery or death). If the vaccination rate q exceeds a threshold q^*,

$$q \geq q^* = 1 - \frac{1}{R_0},$$

then the disease will be eradicated eventually. Note that, since the dynamics of S and I are independent of V and R, it is enough to consider the first two equations of (122) to establish whether the disease is eradicated in the long-time limit. Furthermore, we will normalize the quantities to ratios in $[0, 1]$.

In order to produce estimates of probabilities in this system, we furthermore need to make assumptions about stochasticity present in the quantities S and I. It is natural to interpret the rate equations (122) as the law
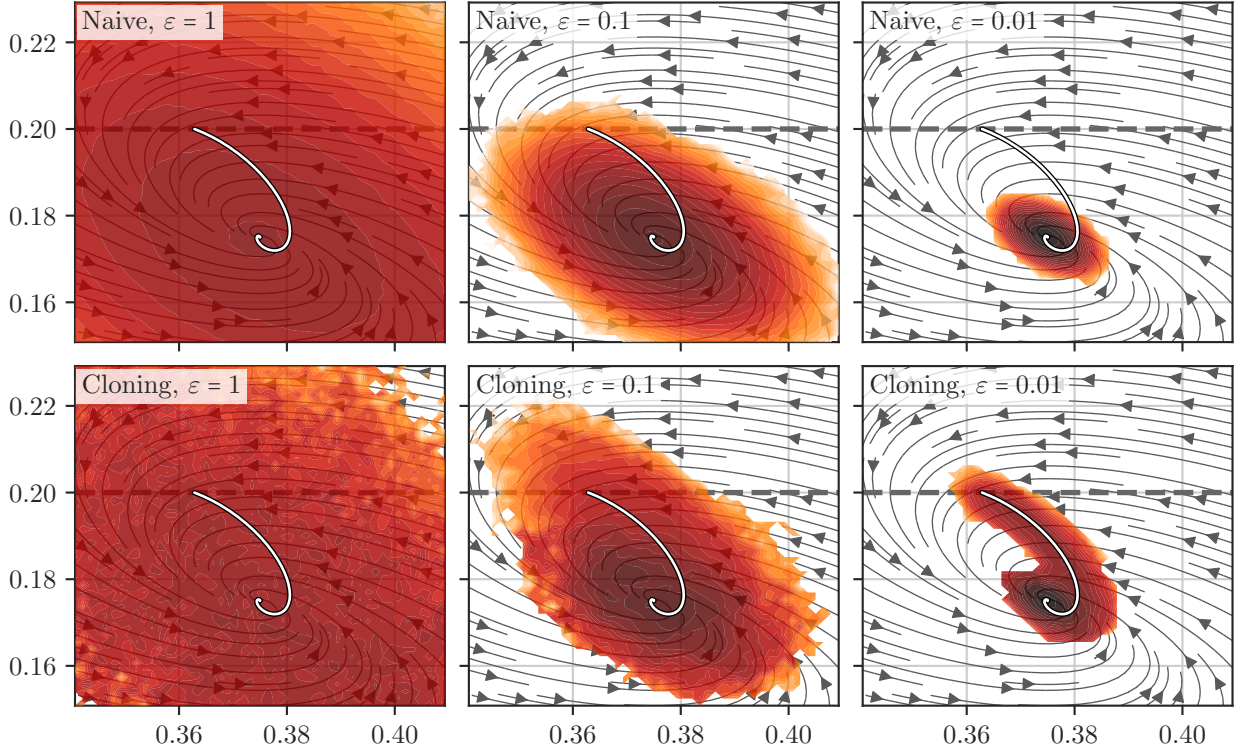
FIG. 9. Comparison of estimating the probability to reach an infection rate of 20% with the naive and the cloning estimator. The heatmap depicts the logarithm of an $(S(t), I(t))$ histogram, the streamlines represent the deterministic dynamics (122). The white line represents the instanton trajectory, and the dashed line the infection threshold. All measurements are obtained with 10000 copies.

ε	1.0	0.5	0.2	0.1	0.09	0.08	0.07	0.06	0.05	0.01
e_{naive}	0.0859	0.1399	0.4222	2.4839	3.3331	6.0061	9.4868	18.2391	31.6228	—
e_{cloning}	0.1487	0.1590	0.1685	0.1729	0.1849	0.1816	0.1836	0.1928	0.1969	0.2749

TABLE I. Relative error of the naive estimator and the cloning estimator for different values of ε. Each value is generated from 1000 experiments with 1000 copies each. The relative variance of the naive estimator diverges for $\varepsilon \to 0$.

of mass action of a reaction network, transforming the species into each other, which would lead to Poisson noise terms as encountered in section IV. For large population sizes, one could also consider a multiplicative Gaussian noise, consistent with the central limit theorem, similar to the discussion of the stochastic Lotka-Volterra model in section III B. Using this approximation, the stochastic system reads

$$
\begin{cases}
dS & = \left(\mu N(1-q) - \mu S - \beta N^{-1} IS \right) dt \\
& \quad + \sqrt{\mu N(1-q) + \mu S + \beta N^{-1} IS} \, dW_S \\
dI & = \left(\beta N^{-1} IS - (\mu + \gamma)I \right) dt \\
& \quad + \sqrt{\beta N^{-1} IS + (\mu + \gamma)I} \, dW_I,
\end{cases} \tag{123}
$$

where W_S and W_I are independent Wiener processes.

As observable, we want to estimate the probability that after time T we have reached an unusually high ratio $z \in [0, 1]$ of infected individuals, $P(I(T) \geq z)$, which we

can write as expectation via

$$
P(I(T) \geq z) = \mathbb{E}\Theta(I(T) - z),
$$

for the Heaviside step function Θ. Using this observable, we can compare the naive estimator with the cloning estimator.

We are choosing parameters $\mu = 0.1$, $\beta = 0.8$, $\gamma = 0.2$ and $N = 1500$, which result in a critical vaccination rate of $q^* = 0.625$. Here, we set $q = 0.1$ instead, resulting in an endemic equilibrium $(\bar{S}, \bar{I}) = (0.375, 0.175)$. The threshold is set to $z = 0.2$, i.e. we want to estimate the probability to have a ration of infected individuals above 20% after $T = 100$. The results for three different values of ε are shown in figure 9, depicting the logarithm of a histogram of final trajectories $(S(t), I(t))$. The streamlines describe the deterministic drifts given in equation (122), while the white line is the instanton to reach the threshold. Indeed, we observe that for $\varepsilon = 1$, reaching the threshold infected rate is not a rare event,

the instanton has no predictive power and the cloning algorithm performs equally or worse to naive sampling. For $\varepsilon = 10^{-1}$, the event is rarely observed for naive sampling, resulting in larger variance estimates of the probability. The configurations resulting from the cloning algorithm instead show a higher prevalence of increased infection rates. This effect is especially pronounced for $\varepsilon = 10^{-2}$, where we do not observe any sample reaching the threshold in the naive case, and where the cloning samples clearly track the instanton trajectory towards the threshold. The relative errors of the two estimators are summarized in table I. With decreasing ε (and therefore exponentially decreasing probability of the rare event), the variance of the naive estimator blows up, while the variance of the cloning estimator remains largely unchanged.

VII. CONCLUSION

Summarizing, in this review we presented a collection of algorithms to estimate rare event probabilities and properties by computing the large deviation minimizer (instanton) for the small-noise limit. They are largely divided into two categories:

In the first category, one minimizes the rate function globally, by discretizing it and then employing numerical minimization techniques. Traditional members of this category are the minimum action method (MAM)[14] and the geometric MAM (gMAM)[15]. Here, we provide a simplified and optimized version of the second, the simplified gMAM, that allows for carrying out the optimization in the space of arc-length parametrized curves with a minimal number of necessary derivatives of the large deviation Hamiltonian. Effectively, this translates into gains in either run-time or implementation complexity over traditional variants. Methods in this category are particularly suited for computing transition trajectories between two sets or points.

In the second category, one instead solves the Hamilton's equations (or instanton equations) associated with the large deviation Hamiltonian. In this category are the Chernykh-Stepanov algorithm[21] and its geometric variant[31]. Here, we provide an interpretation of these algorithms in form of the adjoint formulation of the optimization problem. Methods in this category are effectively employed when the intention is to compute expectations along sample paths, or loosely speaking most likely realizations of extreme events.

Even though these formalisms constitute dual approaches to the same problem, we conclude that they are drastically different in terms of applicability: For example, degenerate forcing is easily incorporated into Hamilton's equations, but constitutes a numerical difficulty in the form of stiff constraints for MAM-type algorithms. Conversely, traversing a saddle point or crossing a separatrix is readily achieved in MAM-type schemes, but leads to loss of convergence in the equations of motion formulation.

Nevertheless, both approaches can be generalized to treat SDEs driven by multiplicative noise as well as stochastic processes driven by non-Gaussian noise. They can also handle, at least formally, infinite dimensional processes, like the solutions of SPDEs. These approaches can also be extended in multiple ways. Here we discussed how related considerations apply to the case of dynamical systems with generic random parameters and we also showed how instantons can be used as input in importance sampling algorithms.

REFERENCES

[1]M. I. Freidlin and A. D. Wentzell, *Random perturbations of dynamical systems*, Vol. 260 (Springer, 2012).

[2]L. Onsager and S. Machlup, "Fluctuations and irreversible processes," Phys. Rev. **91**, 1505–1512 (1953).

[3]H.-K. Janssen, "On a Lagrangean for classical field dynamics and renormalization group calculations of dynamical critical properties," Zeitschrift für Physik B Condensed Matter **23**, 377–380 (1976).

[4]C. de Dominicis, "Techniques de renormalisation de la théorie des champs et dynamique des phénomènes critiques," J. Phys. C **1**, 247 (1976).

[5]P. C. Martin, E. D. Siggia, and H. A. Rose, "Statistical Dynamics of Classical Systems," Physical Review A **8**, 423–437 (1973).

[6]M. Doi, "Second quantization representation for classical many-particle system," Journal of Physics A: Mathematical and General **9**, 1465 (1976).

[7]L. Peliti, "Renormalisation of fluctuation effects in the A+A to A reaction," Journal of Physics A: Mathematical and General **19**, L365 (1986).

[8]R. Graham, "Macroscopic potentials, bifurcations and noise in dissipative systems," in *Fluctuations and Stochastic Phenomena in Condensed Matter*, Lecture Notes in Physics, edited by L. Garrido (Springer Berlin Heidelberg, 1987) pp. 1–34.

[9]D. G. Luchinsky, P. V. E. McClintock, and M. I. Dykman, "Analogue studies of nonlinear systems," Reports on Progress in Physics **61**, 889–997 (1998).

[10]W. E, W. Ren, and E. Vanden-Eijnden, "String method for the study of rare events," Physical Review B **66**, 052301 (2002).

[11]W. E, W. Ren, and E. Vanden-Eijnden, "Simplified and improved string method for computing the minimum energy paths in barrier-crossing events," J. Chem. Phys. **126**, 164103 (2007).

[12]T. Grafke, "String method for generalized gradient flows: computation of rare events in reversible stochastic processes," Journal of Statistical Mechanics: Theory and Experiment **2019**, 043206 (2019).

[13]M. K. Cameron, "Finding the quasipotential for nongradient SDEs," Physica D: Nonlinear Phenomena **241**, 1532–1550 (2012).

[14] W. E, W. Ren, and E. Vanden-Eijnden, "Minimum action method for the study of rare events," Communications on Pure and Applied Mathematics **57**, 637–656 (2004).

[15] M. Heymann and E. Vanden-Eijnden, "The geometric minimum action method: A least action principle on the space of curves," Communications on Pure and Applied Mathematics **61**, 1052–1117 (2008).

[16] T. Grafke, T. Schäfer, and E. Vanden-Eijnden, "Long Term Effects of Small Random Perturbations on Dynamical Systems: Theoretical and Computational Tools," in *Recent Progress and Modern Challenges in Applied Mathematics, Modeling and Computational Science*, Fields Institute Communications (Springer, New York, NY, 2017) pp. 17–55.

[17] J. G. Charney and J. G. DeVore, "Multiple Flow Equilibria in the Atmosphere and Blocking," Journal of the Atmospheric Sciences **36**, 1205–1216 (1979).

[18] M. Hairer, "A theory of regularity structures," Inventiones mathematicae **198**, 269–504 (2014).

[19] M. Hairer and H. Weber, "Large deviations for white-noise driven, nonlinear stochastic PDEs in two and three dimensions," Annales de la Facult des sciences de Toulouse : Mathmatiques **24**, 55–92 (2015).

[20] R. Courant, K. Friedrichs, and H. Lewy, "ber die partiellen Differenzengleichungen der mathematischen Physik," Mathematische Annalen **100**, 32–74 (1928).

[21] A. I. Chernykh and M. G. Stepanov, "Large negative velocity gradients in Burgers turbulence," Physical Review E **64**, 026306 (2001).

[22] T. Grafke, R. Grauer, and T. Schäfer, "Instanton filtering for the stochastic Burgers equation," Journal of Physics A: Mathematical and Theoretical **46**, 062002 (2013).

[23] T. Grafke, R. Grauer, T. Schäfer, and E. Vanden-Eijnden, "Relevance of instantons in Burgers turbulence," EPL (Europhysics Letters) **109**, 34003 (2015).

[24] T. Grafke, R. Grauer, and T. Schäfer, "The instanton method and its numerical implementation in fluid mechanics," Journal of Physics A: Mathematical and Theoretical **48**, 333001 (2015).

[25] L. Zarfaty and B. Meerson, "Statistics of large currents in the Kipnis-Marchioro-Presutti model in a ring geometry," Journal of Statistical Mechanics-Theory and Experiment , 033304 (2016).

[26] T. Speck, A. Engel, and U. Seifert, "The large deviation function for entropy production: the optimal trajectory and the role of fluctuations," Journal of Statistical Mechanics: Theory and Experiment **2012**, P12001 (2012).

[27] P. Tsobgni Nyawo and H. Touchette, "Large deviations of the current for driven periodic diffusions," Physical Review E **94**, 032101 (2016).

[28] D. T. Gillespie, "The chemical Langevin equation," The Journal of Chemical Physics **113**, 297–306 (2000).

[29] W. H. Fleming, "Exit probabilities and optimal stochastic control," Applied Mathematics and Optimization **4**, 329–346 (1977).

[30] T. Grafke, R. Grauer, and S. Schindel, "Efficient Computation of Instantons for Multi-Dimensional Turbulent Flows with Large Scale Forcing," Communications in Computational Physics **18**, 577–592 (2015).

[31] T. Grafke, R. Grauer, T. Schäfer, and E. Vanden-Eijnden, "Arclength Parametrized Hamilton's Equations for the Calculation of Instantons," Multiscale Modeling & Simulation **12**, 566–580 (2014).

[32] A. J. Bray and A. J. McKane, "Instanton Calculation of the Escape Rate for Activation over a Potential Barrier Driven by Colored Noise," Physical Review Letters **62**, 493–496 (1989).

[33] F. Moss and P. V. E. McClintock, "Noise in Nonlinear Dynamical Systems," Noise in Nonlinear Dynamical Systems, by Frank Moss , P. V. E. McClintock, Cambridge, UK: Cambridge University Press, 2009 (2009).

[34] J. Feng and T. G. Kurtz, *Large deviations for stochastic processes*, Mathematical surveys and monographs No. 131 (American Mathematical Society, Providence, RI, 2006) oCLC: 634887801.

[35] A. Shwartz and A. Weiss, *Large Deviations For Performance Analysis: Queues, Communication and Computing* (CRC Press, 1995).

[36] D. M. Roma, R. A. OFlanagan, A. E. Ruckenstein, A. M. Sengupta, and R. Mukhopadhyay, "Optimal path to epigenetic switching," Physical Review E **71**, 011902 (2005).

[37] P. Ney, "Dominating Points and the Asymptotics of Large Deviations for Random Walk on $\mathbb{R}^d$," The Annals of Probability **11**, 158–167 (1983).

[38] G. Dematteis, T. Grafke, and E. Vanden-Eijnden, "Extreme event quantification in dynamical systems with random components," (2018), https://arxiv.org/abs/1808.10764.

[39] R. Chetrite and H. Touchette, "Nonequilibrium Markov Processes Conditioned on Large Deviations," Annales Henri Poincaré **16**, 2005–2057 (2015).

[40] J. Bucklew, *Introduction to Rare Event Simulation*, Springer Series in Statistics (Springer-Verlag, New York, 2004).

[41] A. Pelissetto and F. Ricci-Tersenghi, "Large Deviations in Monte Carlo Methods," Large Deviations in Physics , 161–191 (2014).

[42] M. Cottrell, J.-C. Fort, and G. Malgouyres, "Large deviations and rare events in the study of stochastic algorithms," IEEE Transactions on Automatic Control **28**, 907–920 (1983).

[43] C. Hartmann and C. Schütte, "Efficient rare event simulation by optimal nonequilibrium forcing," Journal of Statistical Mechanics: Theory and Experiment **2012**, P11004 (2012).

[44] G. Ferré and H. Touchette, "Adaptive Sampling of Large Deviations," Journal of Statistical Physics **172**, 1525–1544 (2018).

[45] L. Ebener, G. Margazoglou, J. Friedrich, L. Biferal, and R. Grauer, "Instantons and fluctuations," in this issue (2018).

[46] P. Glasserman and Y. Wang, "Counterexamples in importance sampling for large deviations probabilities," The Annals of Applied Probability **7**, 731–746 (1997).

[47] E. Vanden-Eijnden and J. Weare, "Rare Event Simulation of Small Noise Diffusions," Communications on Pure and Applied Mathematics **65**, 1770–1803 (2012).

[48] P. Del Moral, "Feynman-Kac Formulae," in *Feynman-Kac Formulae: Genealogical and Interacting Particle Systems with Applications*, Probability and its Applications, edited by P. Del Moral (Springer New York, New York, NY, 2004) pp. 47–93.

[49] F. Cerou and A. Guyader, "Adaptive multilevel splitting for rare event analysis," Stochastic Analysis and Applications **25**, 417–443 (2007), wOS:000245237500008.

[50] C. Giardina, J. Kurchan, V. Lecomte, and J. Tailleur, "Simulating Rare Events in Dynamical Processes," Journal of Statistical Physics **145**, 787–811 (2011).

[51] J. Wouters and F. Bouchet, "Rare event computation in deterministic chaotic systems using genealogical particle analysis," Journal of Physics A: Mathematical and Theoretical **49**, 374002 (2016).

[52] In complex situations, such as applications in meteorology or climate where only a block-box type solver is available, explicit modification of the drift should be kept at a minimum.

[53] G. Ferré, T. Grafke, and E. Vanden-Eijnden, In preparation.

[54] C.-E. Bréhier, M. Gazeau, L. Goudenège, T. Lelièvre, and M. Rousset, "Unbiasedness of some generalized adaptive multilevel splitting algorithms," The Annals of Applied Probability **26**, 3559–3601 (2016).