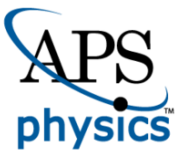

Dynamical Computation of the Density of States and Bayes Factors Using Nonequilibrium Importance Sampling

Grant M. Rotskoff and Eric Vanden-Eijnden

Phys. Rev. Lett. **122**, 150602 — Published 16 April 2019

DOI: 10.1103/PhysRevLett.122.150602

Dynamical computation of the density of states and Bayes factors using nonequilibrium importance sampling

Grant M. Rotskoff* and Eric Vanden-Eijnden†
Courant Institute, New York University, 251 Mercer Street, New York, New York 10012, USA

Nonequilibrium sampling is potentially much more versatile than its equilibrium counterpart, but it comes with challenges because the invariant distribution is not typically known when the dynamics breaks detailed balance. Here, we derive a generic importance sampling technique that leverages the statistical power of configurations transported by nonequilibrium trajectories, and can be used to compute averages with respect to arbitrary target distributions. As a dissipative reweighting scheme, the method can be viewed in relation to the annealed importance sampling (AIS) method and the related Jarzynski equality. Unlike AIS, our approach gives an unbiased estimator, with provably lower variance than directly estimating the average of an observable. We also establish a direct relation between a dynamical quantity, the dissipation, and the volume of phase space, from which we can compute quantities such as the density of states and Bayes factors. We illustrate the properties of estimators relying on this sampling technique in the context of density of state calculations, showing that it scales favorable with dimensionality—in particular, we show that it can be used to compute the phase diagram of the mean-field Ising model from a single nonequilibrium trajectory. We also demonstrate the robustness and efficiency of the approach with an application to a Bayesian model comparison problem of the type encountered in astrophysics and machine learning.

Keywords: Bayes factor; Bayesian evidence; partition function

Statistical estimation using averages over a dynamical process typically relies on the principle of detailed balance. Consider a dynamical system

$$\dot{\boldsymbol{X}}(t,\boldsymbol{x}) = \boldsymbol{b}(\boldsymbol{X}(t,\boldsymbol{x})) \qquad \boldsymbol{X}(0,\boldsymbol{x}) = \boldsymbol{x} \tag{1}$$

where $\boldsymbol{x} \in \Omega \subset \mathbb{R}^n$ is a state that is propagated in time to $\boldsymbol{X}(t,\boldsymbol{x})$ via the vector field $\boldsymbol{b} : \Omega \to \mathbb{R}^n$. If the dynamics is microscopically reversible with respect to some target density $\rho(\boldsymbol{x})$, then this density is preserved under time evolution. Practically, this means that the expectation of an observable $\phi(\boldsymbol{x})$ with respect to $\rho(\boldsymbol{x})$, which we denote by $\langle\phi\rangle$, can be computed as a time average along an equilibrium trajectory generated from (1), provided that the dynamics is ergodic. This direct sampling scheme becomes inefficient if the expectation $\langle\phi\rangle$ is dominated by values of $\boldsymbol{x}$ that are rare under $\rho(\boldsymbol{x})$ and therefore infrequently visited by the dynamics (1).

Importance sampling estimates relying on nonequilibrium dynamics have shown success in a variety of applications, from statistical physics to machine learning [1–7]. Here, we derive a class of estimators based on an exact reweighting of the samples gathered during a nonequilibrium process with a stationary density. Similar to the annealed importance sampling (AIS) method [6] and estimators based on the Jarzynski equality [1], our scheme accelerates the transport of density to rare regions of phase space which may make substantial contributions to equilibrium averages. This basic idea is exploited by many different enhanced sampling techniques [8–11]. Physically, the statistical weight of the transported density can be interpreted through the fluctuation theorem as a dissipative reweighting whose value can be derived explicitly. As we show, the resulting estimator is unbiased, unlike the AIS estimator, which requires computing a ratio of sample means (cf. Eq. 12 of Ref. [6]). Our estimator always has lower variance than the direct estimator, a reduction that comes at the nontrivial cost of generating trajectories. That said, nonequilibrium transport enables us to access states that are exponentially rare in original density but which may dominate expectation values; direct sampling fails dramatically in such settings.

Nonequilibrium estimators.—A generic importance sampling scheme to compute the average with respect to some target density $\rho(\boldsymbol{x})$ reweights samples drawn from another density $\rho_{\rm ne}(\boldsymbol{x})$

$$\langle\phi\rangle = \langle\phi\rho/\rho_{\rm ne}\rangle_{\rm ne}. \tag{2}$$

Our samplers use for $\rho_{\rm ne}(\boldsymbol{x})$ the non-equilibrium stationary density of a dynamical system based on generating trajectories by an initiate-then-propagate procedure: We draw points $\boldsymbol{x}$ from the density $\rho(\boldsymbol{x})$ that we then propagate forward and backward in time using the dynamics (1) until the trajectories $\boldsymbol{X}(t,\boldsymbol{x})$ hit some fixed target set. Concrete applications determine the appropriate choice of target sets: they could for example be the boundary of Ω or the fixed points of (1) in Ω [12]. Using the set of trajectories generated this way we define the

nonequilibrium average $\langle\cdot\rangle_{\rm ne}$ as

$$\langle\phi\rangle_{\rm ne} = \frac{1}{\langle\tau\rangle}\int_\Omega \int_{\tau^-(\boldsymbol{x})}^{\tau^+(\boldsymbol{x})} \phi(\boldsymbol{X}(t,\boldsymbol{x}))dt\,\rho(\boldsymbol{x})d\boldsymbol{x} \tag{3}$$

where $\tau^+(\boldsymbol{x}) \geq 0$ and $\tau^-(\boldsymbol{x}) \leq 0$ are the first times at which $\boldsymbol{X}(t,\boldsymbol{x}) \in \partial\Omega$ in the future or in the past, respectively, and $\langle\tau\rangle = \langle\tau^+\rangle - \langle\tau^-\rangle$ [13]. By changing integration variables using $X(t,\boldsymbol{x}) \to \boldsymbol{x}$ and $t \to -t$ we can express (3) as

$$\langle\phi\rangle_{\rm ne} = \frac{1}{\langle\tau\rangle}\int_\Omega \phi(\boldsymbol{x}) \int_{\tau^-(\boldsymbol{x})}^{\tau^+(\boldsymbol{x})} J(t,\boldsymbol{x})\rho(\boldsymbol{X}(t,\boldsymbol{x}))dt\,d\boldsymbol{x} \tag{4}$$

where $J(t,\boldsymbol{x})$ is the Jacobian of the transformation:

$$J(t,\boldsymbol{x}) = \exp\left(\int_0^t \nabla\cdot\boldsymbol{b}(\boldsymbol{X}(s,\boldsymbol{x}))ds\right). \tag{5}$$

Physically, the Jacobian corresponds to the total energy dissipation upto time t along the trajectory. This derivation is described in detail in the supplementary material (SM). Now, (4) can be interpreted as an expectation with respect to a nonequilibrium density, $\langle\phi\rangle_{\rm ne} = \int_\Omega \phi(\boldsymbol{x})\rho_{\rm ne}(\boldsymbol{x})d\boldsymbol{x}$, with $\rho_{\rm ne}(\boldsymbol{x})$ given by

$$\rho_{\rm ne}(\boldsymbol{x}) = \frac{1}{\langle\tau\rangle}\int_{\tau^-(\boldsymbol{x})}^{\tau^+(\boldsymbol{x})} J(t,\boldsymbol{x})\rho(\boldsymbol{X}(t,\boldsymbol{x}))dt. \tag{6}$$

We can now use this expression for $\rho_{\rm ne}(\boldsymbol{x})$ in (2) for reweighting. As shown in the SM this gives

$$\langle\phi\rangle = \left\langle \frac{\int_{\tau^-}^{\tau^+} \phi(\boldsymbol{X}(t))J(t)\rho(\boldsymbol{X}(t))dt}{\int_{\tau^-}^{\tau^+} J(t)\rho(\boldsymbol{X}(t))dt}\right\rangle \tag{7}$$

which yields the estimator, one of our main results,

$$\langle\phi\rangle = \lim_{N\to\infty}\langle\phi\rangle_N \quad \text{where} \quad \langle\phi\rangle_N = \frac{1}{N}\sum_{i=1}^N \frac{\int_{\tau^-(\boldsymbol{x}_i)}^{\tau^+(\boldsymbol{x}_i)} \phi(\boldsymbol{X}(t,\boldsymbol{x}_i))J(t,\boldsymbol{x}_i)\rho(\boldsymbol{X}(t,\boldsymbol{x}_i))dt}{\int_{\tau^-(\boldsymbol{x}_i)}^{\tau^+(\boldsymbol{x}_i)} J(t,\boldsymbol{x}_i)\rho(\boldsymbol{X}(t,\boldsymbol{x}_i))dt}, \tag{8}$$

provided that the points $\boldsymbol{x}_i$ are drawn (not necessarily independently) from $\rho(\boldsymbol{x})$. This equation reweights points sampled along a nonequilibrium trajectory according to their dissipation, a physically analogous strategy to that in AIS.

Unlike other dissipative reweighting strategies, the estimator $\langle\phi\rangle_N$ is unbiased, valid for any dynamics (1) and any target density $\rho(\boldsymbol{x})$. Like standard Metropolis Monte-Carlo, it only requires knowing the density up to a normalization factor. It has lower variance than the direct estimator $N^{-1}\sum_{i=1}^N \phi(\boldsymbol{x}_i)$ [14] since the variance of this direct estimator is $N^{-1}(\langle\phi^2\rangle - \langle\phi\rangle^2)$ whereas the variance of $\langle\phi\rangle_N$ is $N^{-1}(A - \langle\phi\rangle^2)$ with $A \leq \langle\phi^2\rangle$ since Jensen's inequality implies

$$\begin{aligned}\left\langle\left|\frac{\int_{\tau^-}^{\tau^+} \phi(\boldsymbol{X}(t))J(t)\rho(\boldsymbol{X}(t))dt}{\int_{\tau^-}^{\tau^+} J(t)\rho(\boldsymbol{X}(t))dt}\right|^2\right\rangle &= A \\ \leq \left\langle\frac{\int_{\tau^-}^{\tau^+} |\phi(\boldsymbol{X}(t))|^2 J(t)\rho(\boldsymbol{X}(t))dt}{\int_{\tau^-}^{\tau^+} J(t)\rho(\boldsymbol{X}(t))dt}\right\rangle &= \langle\phi^2\rangle.\end{aligned} \tag{9}$$

With a proper choice of $\boldsymbol{b}(\boldsymbol{x})$ the estimator $\langle\phi\rangle_N$ in (8) has the potential to significantly outperform the direct estimator. It does so by transporting points drawn naively from $\rho(\boldsymbol{x})$ towards regions that statistically dominate the expectation of $\phi(\boldsymbol{x})$. In practice, it is also simple to use: (i) sample points $\boldsymbol{x}_i$ from $\rho(\boldsymbol{x})$ using e.g. standard Monte Carlo; (ii) compute the trajectory $\boldsymbol{X}(t,\boldsymbol{x}_i)$ passing through each of these points by integrating (1) forward and backward in time until $\boldsymbol{X}(\tau^\pm(\boldsymbol{x}_i),\boldsymbol{x}_i) \in \partial\Omega$ (which also gives $\tau^\pm(\boldsymbol{x}_i)$); (iii) use these data to calculate $J(t,\boldsymbol{x}_i)$ from (5) first, then the integrals in (8); (iv) average the results to get $\langle\phi\rangle_N$. Note that the operations in (ii) and (iii) can be performed in parallel, and we can monitor the value of the running average $\langle\phi\rangle_N$ as N increases to check convergence.

Density of states (DOS).— Consider a d-dimensional system with position $\boldsymbol{q} \in \mathbb{R}^d$, momenta $\boldsymbol{p} \in \mathbb{R}^d$, and Hamiltonian $\mathcal{H}(\boldsymbol{q},\boldsymbol{p}) = \frac{1}{2}|\boldsymbol{p}|^2 + U(\boldsymbol{q})$ where $U(\boldsymbol{q})$ is some potential bounded from below. Let $V(E)$ be the volume of phase space below some threshold energy E,

$$V(E) = \int_{\mathcal{H}(\boldsymbol{q},\boldsymbol{p})<E} d\boldsymbol{q}d\boldsymbol{p}, \tag{10}$$

From $V(E)$ one can compute the DOS, $D(E) = V'(E)$, or the canonical partition function, $Z(\beta) = \int_{\mathbb{R}} e^{-\beta E}D(E)dE = \beta\int_{\mathbb{R}} e^{-\beta E}V(E)dE$.

To calculate (10) with our estimator (8), we set $\boldsymbol{x} = (\boldsymbol{q},\boldsymbol{p})$, define $\Omega = \{(\boldsymbol{q},\boldsymbol{p}) : \mathcal{H}(\boldsymbol{q},\boldsymbol{p}) < E_{\rm max}\}$ for some $E_{\rm max} < \infty$, and use dissipative Langevin dynamics with $\boldsymbol{b}(\boldsymbol{x}) = (\boldsymbol{p}, -\nabla U(\boldsymbol{q}) - \gamma\boldsymbol{p})$ in (1)

$$\dot{\boldsymbol{q}} = \boldsymbol{p}, \qquad \dot{\boldsymbol{p}} = -\nabla U(\boldsymbol{q}) - \gamma\boldsymbol{p}, \tag{11}$$

for some friction coefficient $\gamma > 0$. With this choice, the dissipative term in the estimator (8) takes the simple form:

$$J(t, \boldsymbol{x}) = e^{-d\gamma t}. \tag{12}$$

If we also choose the target density $\rho(\boldsymbol{x})$ to be uniform in Ω, the estimator further simplifies due to cancelation of the two ρ terms in (8). We arrive at

$$\frac{V(E)}{V(E_{\max})} = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} e^{-d\gamma\left(\tau^E(\boldsymbol{x}_i) - \tau^-(\boldsymbol{x}_i)\right)} \tag{13}$$

where $\tau^E(\boldsymbol{x})$ denotes the positive (and possibly infinite) or negative time for a trajectory initiated from $\boldsymbol{x} = (\boldsymbol{q}, \boldsymbol{p})$ to reach energy $E \le E_{\max}$ under the dynamics (11). Eq. (13) is our second main result: it establishes a dictionary between a nonequilibrium dynamical quantity and a purely static, global property of the energy landscape, $V(E)$. This result asserts that the rate of decrease of the volume of phase space can be measured by computing an average of the total dissipation of nonequilibrium descent trajectories. We do not know of an analogous result in the literature.

The $\tau^+(\boldsymbol{x})$ terms vanish in this dynamics because the time to reach a local minimum diverges [15]. In practice we halt the forward trajectories when the norm of the gradient is below some tolerance. To compute an unnormalized volume, we can estimate $V(E_{\max})$ with standard Monte Carlo integration.

The power of the procedure we have described comes from the fact that the forward trajectories are guaranteed to visit regions of low energy around local minima of $U(\boldsymbol{q})$ that would otherwise be difficult to sample by drawing points uniformly in $\{\mathcal{H}(\boldsymbol{q}, \boldsymbol{p}) < E_{\max}\}$. In this regard our approach is also similar to nested sampling [16–20]. Like nested sampling, we do not require an *a priori* stratification of the energy shells, which is the way the DOS is typically calculated via thermodynamic integration [21, 22] or simulated tempering [23, 24]. Our method also offers several advantages over nested sampling. First, the depth of energies reached in nested sampling is determined by the initial number of points used in a computation. If too few points are used, the calculation must be repeated in full with a larger number of initial points. Here, the accuracy of the calculation improves and explores deeper minima simply by running additional ascent/descent trajectories. In addition, our approach does not require uniform sampling below *every* energy level, which is required in nested sampling and is a difficult condition to implement [19]. We must only generate points uniformly below the highest energy level, $E_{\max}$, which is usually much easier. Computationally, we also benefit from the fact that every trajectory contributes independently to our estimator, meaning that the implementation is trivially parallelizable.

Variance estimation in the small γ limit.— We know from (9) that the variance of our estimator is lower than that of the direct estimator. In the specific context of a DOS calculation using (11), we can analyze the variance more explicitly in the limit of small γ, in which the descent dynamics in (11) reduces to a closed equation for the energy $E = \mathcal{H}(\boldsymbol{q}, \boldsymbol{p})$ on the rescaled time $t' = \gamma t$. This dynamics evolves on the disconnectivity (or Reeb) graph [25], which branches at every energy level at which a basin where $\mathcal{H}(\boldsymbol{q}, \boldsymbol{p}) \le E$ splits into more than one connected component. In the simplest case when the potential $U(\boldsymbol{q})$ has a single well, the graph has only one branch and the value of $\gamma(\tau^E(\boldsymbol{x}_i) - \tau^-(\boldsymbol{x}_i))$ becomes the same along every trajectory when $\gamma \to 0$. Therefore the estimator (13) has zero variance—a single trajectory gives the exact value for $V(E)/V(E_{\max})$. If the disconnectivity graph has several branches, we can count all the paths along the graph starting at $E = E_{\max}$ which end at a given branch. Assuming that the number of such paths is $M \ge 1$, we can associate a deterministic time $\Delta\tau_j^E > 0$, possibly infinite, along each path. We define $\Delta\tau_j^E$ with $j = 1, \ldots, M$ as the total rescaled time the trajectory takes to go from $\mathcal{H}(\boldsymbol{q}, \boldsymbol{p}) = E_{\max}$ to $\mathcal{H}(\boldsymbol{q}, \boldsymbol{p}) = E$ by the effective dynamics for E along the path with index j and $\Delta\tau_j^E = \infty$ if the path terminates at an energy $E' > E$. For any initial condition, $\lim_{\gamma\to 0} \gamma(\tau^E(\boldsymbol{x}_i) - \tau^-(\boldsymbol{x}_i)) = \Delta\tau_j^E$ for some index j, meaning the only random component in the procedure is which path is picked if the trajectory starts at $\boldsymbol{x}_i$. We denote by p_j the probability, computed over all initial conditions drawn uniformly in $\{\mathcal{H}(\boldsymbol{q}, \boldsymbol{p}) < E_{\max}\}$, that the path with index j is taken. Then in the small γ limit the mean and variance of the estimator (13) are

$$\begin{aligned} \text{mean} &= \frac{V(E)}{V(E_{\max})} = \sum_{j=1}^{M} p_j e^{-d\Delta\tau_j^E}, \\ \text{variance} &= \sum_{j=1}^{M} p_j e^{-2d\Delta\tau_j^E} - \text{mean}^2. \end{aligned} \tag{14}$$

The specific values of p_j, $\Delta\tau_j^E$, and M which determine the quality of the estimator (13) depend on both the structure of the disconnectivity graph and the effective equation for the energy on this graph. What is remarkable, however, is that p_j, $\Delta\tau_j^E$, and M depend on the dimensionality of the system only

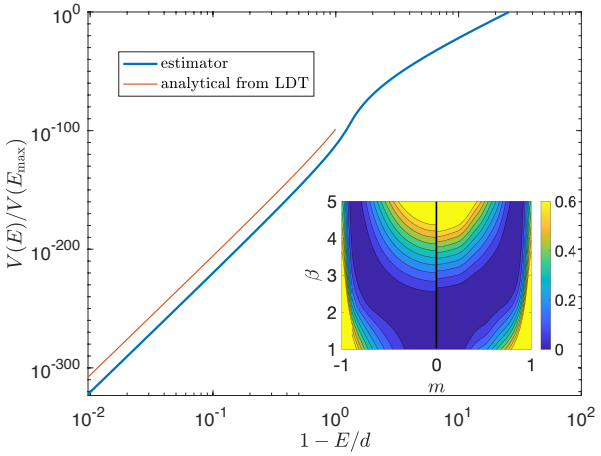

FIG. 1. Mean-field Ising model with potential (15): volume ratio obtained from (13) using a single ascent / descent trajectory. The inset shows the free energy in β and $m = d^{-1}\sum_{i=1}^{d}\cos(q_i)$ that can also be estimated from this single trajectory (right half) and analytically using large deviation theory (left half).

indirectly. In high dimensional settings, the complexity of the disconnectivity is a generic challenge, but our approach has favorable properties even in these difficult cases. In particular, the computational cost of the procedure increases only linearly in γ as we decrease this parameter to small values. We also stress that the formulae (14) for the mean and variance rely on the assumption $\gamma \ll 1$, but the estimator remains valid for any value of γ.

Phase diagram of the mean-field Ising model.— As an illustration of the statistical power contained in the nonequilibrium trajectories, we computed the phase diagram of the mean-field Ising model with potential

$$U(\boldsymbol{q}) = -\frac{1}{2d}\sum_{i,j=1}^{d}\cos(q_i)\cos(q_j). \tag{15}$$

This system is known to display a phase transition in temperature at the critical $\beta_c = 2$ [26]. The potential (15) is double-well, but because of the symmetry $\boldsymbol{q} \to -\boldsymbol{q}$ a single ascent / descent trajectory can be used to estimate the volume ratio $V(E)/V(E_{\max})$. We note that while there are only to energy minima, the energy landscape has an exponential number of critical points (the derivative with respect to θ_j vanishes 2^d when $\sin(\theta_j) = 0$), so the geometry of the landscape is nontrivial. We performed this calculation when $d = 100$ with $\gamma = 10^{-3}$ to obtain the result shown in Fig. 1: as can be seen, this result spans 300 order of magnitudes and compares very well with the analytical estimate that can be obtained in the large d limit, as described in the SM. The single ascent /descent trajectory can also be used to calculate the free energy in β and magnetization $m = d^{-1}\sum_{i=1}^{d}\cos(q_i)$ of the system (cf. SM). The result shown in the inset of Fig. 1 demonstrates that our estimate compares well with the large d estimate. The code to reproduce these experiments is available on Gitlab [27]. The numerical experiments require only several minutes of computation on a single core, but parallelization strategies could dramatically reduce the duration.

Bayes factor.—The computations for the density of states have an equivalent manifestation in Bayesian estimation. Given a model $\mathcal{M}$, one seeks to maximize the probability of a set of parameters $\theta \in \mathbb{R}^d$ conditioned on observations of data $\mathcal{D}$. Using Bayes Theorem, we can write

$$\mathbb{P}(\boldsymbol{\theta}|\mathcal{D},\mathcal{M}) = \mathcal{L}(\boldsymbol{\theta})\pi(\boldsymbol{\theta})/Z \tag{16}$$

where $\mathcal{L}(\boldsymbol{\theta}) = \mathbb{P}(\mathcal{D}|\boldsymbol{\theta},\mathcal{M})$ is the likelihood function, $\pi(\boldsymbol{\theta}) = \mathbb{P}(\boldsymbol{\theta}|\mathcal{M})$ is the prior, and $Z = \mathbb{P}(\mathcal{D}|\mathcal{M}) = \int \mathcal{L}(\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}$ is the partition function, often called the Bayesian evidence in this context; it is the canonical partition function with $\beta = 1$.

In Bayesian inference, we choose a model and then estimate its parameters without knowledge of the partition function by doing gradient descent on $-\log\mathcal{L}(\boldsymbol{\theta}) \equiv U(\boldsymbol{\theta})$, which depends on the model we have taken. However, there is no *a priori* guarantee that the chosen model is optimal, so it is often necessary to make comparisons of two distinct models $\mathcal{M}$ and $\mathcal{M}'$. Ideally, one would compare the probability of the observed data given each model, that is the Bayes factors

$$Z/Z' = \mathbb{P}(\mathcal{D}|\mathcal{M})/\mathbb{P}(\mathcal{D}|\mathcal{M}'). \tag{17}$$

Similarly, computing posterior probabilities also requires knowledge of the partition function.

As we have already emphasized, computing Z is intractable analytically in all but the simplest cases. Skilling [16] demonstrated that it is possible to numerically evaluate the "prior volume",

$$V(L) = \int_{\mathcal{L}(\boldsymbol{\theta})\geq L}\pi(\boldsymbol{\theta})d\boldsymbol{\theta} \tag{18}$$

to produce an estimate of Z via

$$Z = \int_0^{L_0} V(L)\, dL, \tag{19}$$

where L_0 is the maximum value of the likelihood. Just as in the density of states calculation, we can evaluate the Bayesian evidence by using trajectorial estimators.

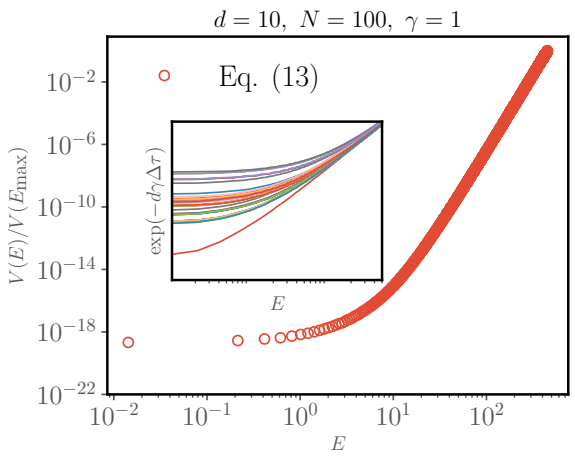

FIG. 2. Mixture of Gaussians inference problem with $d = 10$ and 50 wells in the mixture. The volume of states below $E = -\log \mathcal{L}$ is shown as red circles. In the small gamma limit, the time to reach energy E, $\Delta\tau = \tau^E - \tau^-$, should be independent of the initial condition. The inset shows that the dissipative dynamics converges to many different local minima, corresponding to the individually discernible lines in the inset. A total of $N = 100$ trajectories are plotted, but there are fewer than one hundred visible lines, meaning that some trajectories not only end in the same basin, but also have decay in energy at the same rate, as the small γ limit predicts. This demonstrates that the low variance regime can be achieved even with modest values of γ.

To do so, we sample parameters of the model $\mathcal{M}$ uniformly and define a flow of parameters via dissipative Langevin dynamics with $U(\boldsymbol{\theta}) = -\log \mathcal{L}(\boldsymbol{\theta})$ (which also gives L_0 from the terminal point of the descent trajectories). We construct an estimate of Z by computing $V(L)$ using Eq. (13) and numerically integrating Eq. (19) using quadrature. Note that the contribution from the momenta can be factored out and the resulting Gaussian integral can be computed exactly.

We tested our approach using a mixture of Gaussians model, a benchmark which has been used to characterize nested sampling for inference problems [28]. The model is defined as a mixture of n distributions in dimension d with amplitudes A_i,

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{i=1}^{n} A_i e^{-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\theta}-\boldsymbol{\mu}_i)}. \qquad (20)$$

Though we do not have access to the exact expression for $V(E)$ at all energy levels in this model, we can evaluate the partition function Z exactly.

We used $n = 50$ wells with depths exponentially distributed in dimension $d = 10$, an example much more complex than previous benchmarks. While this landscape is not rugged, in mixture of Gaussian problems entropic effects can lead to extremely difficult optimization problems because we are required to sample exponentially small volumes. In this regime, brute force Monte Carlo approaches fail dramatically. Fig. 2 illustrates the statistical power of the trajectory reweighting approach. With only 100 trajectories, we recover the volume of states for the likelihood function extremely accurately, especially at low energies, where the standard error is vanishingly small. An accurate estimate at low energies leads to robust estimates of Z because the contribution to Z decays exponentially with E. In particular, we know the low energy volume estimates are accurate because we compute $Z = 17.41$ versus the exact result $Z = 17.10$. For this calculation we set $E_{\max} = 450$, meaning that the states we neglected have likelihood lower than e^{-450}.

Conclusions.—Any estimate of the microcanonical partition function requires a thorough exploration of the states of the system. Both naive Monte Carlo sampling and equilibrium dynamics often fail to visit states, which, though rare, dramatically impact the thermodynamic properties of the system. A nonequilibrium dynamics suffers from precisely the opposite problem: it explores the states rapidly, but not in proportion to their equilibrium probabilities. Our estimator, via Eq. (13) establishes a simple link between a nonequilibrium dynamical observable and a static property, the volume of phase space.

With a properly formulated algorithm, we can fully account for the statistical bias of a nonequilibrium dynamics. The resulting estimators can access states that are extremely atypical in equilibrium sampling schemes, but nevertheless physically consequential. While we demonstrated the potential of these estimators by computing the density of states and the computationally analogous Bayes factor, the expression in (8) is extremely general. Attractive applications within reach include adapting this approach to basin volume calculations [29, 30], computing the partition function of restricted Boltzmann machines [6, 31], and importance sampling to compute properties of systems in nonequilibrium stationary states, like active matter [32, 33].

Acknowledgements.—The authors thank Daan Frenkel, Stefano Martiniani, K. Julian Schrenk, and Shang-Wei Ye for discussions that helped motivate this work. G.M.R. acknowledges support from the James S. McDonnell Foundation. E.V.E. was supported by National Science Foundation (NSF) Materials Research Science and Engineering Center Program Award DMR-1420073; and by NSF Award

DMS-1522767.

* rotskoff@cims.nyu.edu
† eve2@cims.nyu.edu

[1] C. Jarzynski, Phys. Rev. Lett. **78**, 2690 (1997).
[2] G. Hummer and A. Szabo, Proc. Natl. Acad. Sci. USA **98**, 3658 (2001).
[3] S. X. Sun, J. Chem. Phys. **118**, 5769 (2003).
[4] M. Athénes, Eur. Phys. J. B **38**, 651 (2004).
[5] C. Dellago and G. Hummer, Entropy **16**, 41 (2013).
[6] R. M. Neal, Stat. Comput. **11**, 125 (2001).
[7] J. P. Nilmeier, G. E. Crooks, D. D. L. Minh, and J. D. Chodera, Proc. Nat. Acad. Sci. USA **108**, E1009 (2011).
[8] R. J. Allen, C. Valeriani, and P. Rein ten Wolde, J. Phys. Condens. Matter **21**, 463102 (2009).
[9] A. Bouchard-Côté, S. J. Vollmer, and A. Doucet, J. Am. Stat. Assoc. **113**, 855 (2018).
[10] E. A. J. F. Peters and G. de With, Phys. Rev. E **85**, 026703 (2012).
[11] M. Michel, S. C. Kapfer, and W. Krauth, J. Chem. Phys. **140**, 054116 (2014).
[12] As made clear by the calculation in the supplementary material, the main requirement to derive the estimator (8) is that $\tau^{\pm}(\boldsymbol{x})$ satisfy $\tau^{\pm}(\boldsymbol{X}(t,\boldsymbol{x})) = \tau^{\pm}(\boldsymbol{x}) - t$ for all $t \in [\tau^{-}(\boldsymbol{x}), \tau^{+}(\boldsymbol{x})]$. This requirement is satisfied with $\tau^{\pm}(\boldsymbol{x})$ defined as the times when the trajectory $\boldsymbol{X}(t,\boldsymbol{x})$ hits fixed target sets.
[13] Note that these formula assume that $\langle\tau\rangle$ is finite. However, the final expression for (2) makes sense even in situations where this time diverges.
[14] Note that this comparison neglects the cost of generating the ascent / descent trajectories from the points $\boldsymbol{x}_i$.
[15] This indicates that the density $\rho_{\mathrm{ne}}(\boldsymbol{x})$ is singular here because the density becomes atomic on the local minima of $\mathcal{H}(\boldsymbol{x})$. However the estimator (7) remains valid and explicitly given by (13).
[16] J. Skilling, Bayesian Anal. **1**, 833 (2006).
[17] B. J. Brewer, L. B. Pártay, and G. Csányi, Stat. Comput. **21**, 649 (2011).
[18] L. B. Pártay, A. P. Bartók, and G. Csányi, Phys. Rev. E **89**, 022302 (2014).
[19] S. Martiniani, J. D. Stevenson, D. J. Wales, and D. Frenkel, Phys. Rev. X **4**, 031034 (2014).
[20] P. G. Bolhuis and G. Csányi, Phys. Rev. Lett. **120**, 250601 (2018).
[21] D. Frenkel and B. Smit, *Understanding Molecular Simulation: From Algorithms to Applications* (Elsevier, 2001).
[22] A. Gelman and X. L. Meng, Stat. Sci. **13**, 163 (1998).
[23] E. Marinari and G. Parisi, EPL **19**, 451 (1992).
[24] D. J. Earl and M. W. Deem, Phys. Chem. Chem. Phys. **7**, 3910 (2005).
[25] D. J. Wales, M. A. Miller, and T. R. Walsh, Nature **394**, 758 (1998).
[26] A. Martinsson, J. Lu, B. Leimkuhler, and E. Vanden-Eijnden, arXiv (2018), arXiv:1809.05066 [math, stat].
[27] Source code available at: `https://gitlab.com/rotskoff/trajectory_estimators`.
[28] F. Feroz and M. P. Hobson, Mon. Notices Royal Astron. Soc. **384**, 449 (2008).
[29] D. Asenjo, F. Paillusson, and D. Frenkel, Phys. Rev. Lett. **112**, 098002 (2014).
[30] S. Martiniani, K. J. Schrenk, J. D. Stevenson, D. J. Wales, and D. Frenkel, Phys. Rev. E **93**, 012906 (2016).
[31] R. B. Grosse, C. J. Maddison, and R. R. Salakhutdinov, in *Advances in Neural Information Processing Systems 26*, edited by C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger (Curran Associates, Inc., 2013) pp. 2769–2777.
[32] C. Bechinger, R. Di Leonardo, H. Löwen, C. Reichhardt, G. Volpe, and G. Volpe, Rev. Mod. Phys. **88**, 045006 (2016).
[33] M. E. Cates and J. Tailleur, Annu. Rev. Condens. Matter Phys. **6**, 219 (2015).