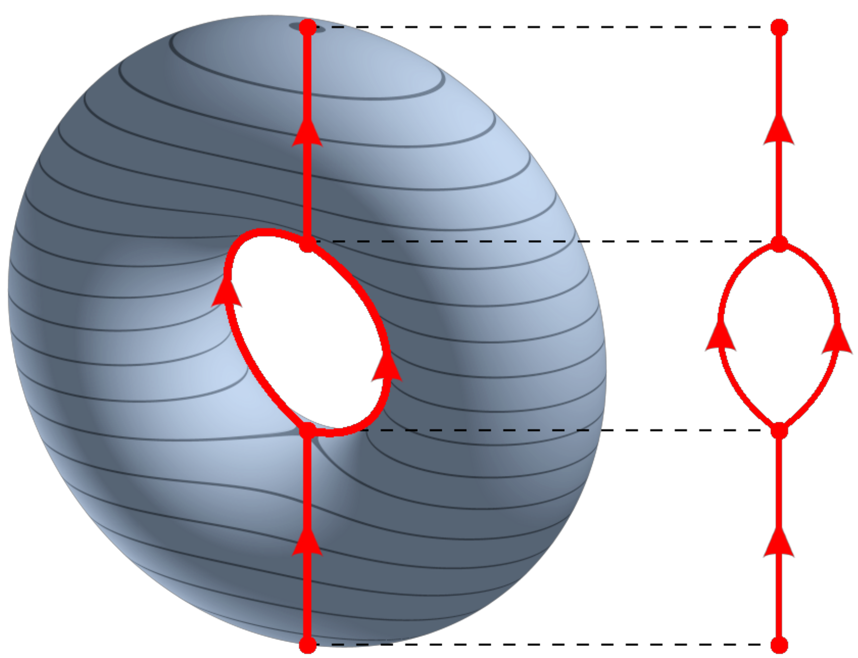
Importance sampling with non equilibrium trajectories



Eric Vanden-Eijnden
Courant Institute

Joint work with Grant Rotskoff

Refs: arXiv:1809.11132

Computational Statistics and Molecular Simulation:
A Practical Cross-Fertilization.
BIRS-CMO, Oaxaca, Nov 2018

Partition function and density of states

- Given a measure μ on $\Omega \subset \mathbb{R}^d$, the *partition function* is the normalization factor

$$Z = \int_\Omega d\mu(\boldsymbol{x})$$

- Setting $U = -\log(d\mu/d\mu_0)$ the *density of state* $D(z)$ is

$$D(u) = \frac{dV}{du} \quad \text{where} \quad V(u) = \int_{U(\boldsymbol{x})<u} d\mu_0(\boldsymbol{x}) \quad \text{so that} \quad Z = \int_{\mathbb{R}} e^{-u} D(u) du$$

  ▷ **Statistical mechanics:** If $d\mu(\boldsymbol{x}) = e^{-\beta U(\boldsymbol{x})} d\boldsymbol{x}$ for some $U : \Omega \to [0, \infty)$,

  D(E) gives Z(β) at any β

  $Z(\beta) = \int_{\mathbb{R}} e^{-\beta z} D(z) dz$

$$Z(\beta) = \int_\Omega \exp(-\beta U(\boldsymbol{x})) d\boldsymbol{x}, \qquad D(E) = \int_\Omega \delta(E - U(\boldsymbol{x})) d\boldsymbol{x}$$

  ▷ **Bayesian inference:** If $L(\boldsymbol{y}|\boldsymbol{x}, M)$ is the likelihood of the data $\boldsymbol{y}$ given the parameters $\boldsymbol{x}$ and the model M, and $d\mu_0(\boldsymbol{x})$ is the (normalized) prior, $d\mu(\boldsymbol{x}) = L(\boldsymbol{y}|\boldsymbol{x}, M) d\mu_0(\boldsymbol{x})$ is the posterior and

$$Z(\boldsymbol{y}, M) = \int_\Omega L(\boldsymbol{y}|\boldsymbol{x}, M) d\mu_0(\boldsymbol{x}) \qquad \text{is the evidence}$$

- Methods to estimate Z and $D(u)$ include thermodynamic integration, Wang-Landau, simulated / parallel tempering, nested sampling, etc. – note that $V(u) = \mathbb{P}_0(U(\boldsymbol{x}) < u)$ is an observable, but Z is not.

- Typically hard to compute in high dimension because of (i) multimodality of μ and (ii) entropic effects.

Importance sampling along trajectories

- **Expectations via reweighing:** Given an observable $\phi : \Omega \to \mathbb{R}$, and two measures μ_0 and μ_1 such that $\mu_0 \ll \mu_1$

$$\mu_0(\phi) = \int_\Omega \phi d\mu_0 = \int_\Omega \phi \frac{d\mu_0}{d\mu_1} d\mu_1 = \mu_1(\phi d\mu_0/d\mu_1)$$

- **Expectations along trajectories (with a flavor of PDMP):** Given $\boldsymbol{b} : \Omega \to \mathbb{R}^d$ let

$$d\boldsymbol{X}(t,\boldsymbol{x})/dt = \boldsymbol{b}(\boldsymbol{X}(t,\boldsymbol{x})), \qquad \boldsymbol{X}(0,\boldsymbol{x}) = \boldsymbol{x} \in \Omega$$

$$\tau_-(\boldsymbol{x}) = \sup\{t < 0 : \boldsymbol{X}(t,\boldsymbol{x}) \in \partial\Omega\}, \qquad \tau_+(\boldsymbol{x}) = \inf\{t > 0 : \boldsymbol{X}(t,\boldsymbol{x}) \in \partial\Omega\}$$

Given μ_0, define μ_1 via

$$\mu_1(\phi) = \bar{\tau}^{-1} \int_\Omega \left(\int_{\tau_-(\boldsymbol{x})}^{\tau_+(\boldsymbol{x})} \phi(\boldsymbol{X}(t,\boldsymbol{x}))dt \right) d\mu_0(\boldsymbol{x}), \qquad \bar{\tau} = \int_\Omega (\tau_+(\boldsymbol{x}) - \tau_-(\boldsymbol{x}))d\mu_0(\boldsymbol{x})$$

- **Combining the two:** We can write an expression for μ_1 and use it to derive

change of variable
+ invertibility of the flow map

$$\mu_0(\phi) = \int_\Omega \frac{\int_{\tau^-(\boldsymbol{x})}^{\tau^+(\boldsymbol{x})} \phi(\boldsymbol{X}(t,\boldsymbol{x}))J(t,\boldsymbol{x})\rho_0(\boldsymbol{X}(t,\boldsymbol{x}))dt}{\int_{\tau^-(\boldsymbol{x})}^{\tau^+(\boldsymbol{x})} J(t,\boldsymbol{x})\rho_0(\boldsymbol{X}(t,\boldsymbol{x}))dt} d\mu_0(\boldsymbol{x})$$

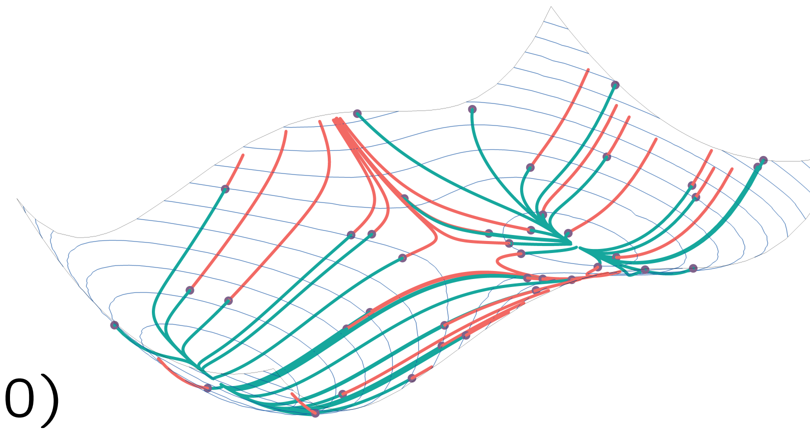
where $\rho_0 = d\mu_0/d\boldsymbol{x}$ and

transport points drawn from μ_0
towards regions that dominate $\mu_0(\phi)$?

$$J(t,\boldsymbol{x}) = \exp\left(\int_0^t \operatorname{div} \boldsymbol{b}(X(s,\boldsymbol{x}))ds \right)$$

$$\mu_0(\phi) = \int_\Omega \frac{\int_{\tau^-(\boldsymbol{x})}^{\tau^+(\boldsymbol{x})} \phi(\boldsymbol{X}(t,\boldsymbol{x}))J(t,\boldsymbol{x})\rho_0(\boldsymbol{X}(t,\boldsymbol{x}))dt}{\int_{\tau^-(\boldsymbol{x})}^{\tau^+(\boldsymbol{x})} J(t,\boldsymbol{x})\rho_0(\boldsymbol{X}(t,\boldsymbol{x}))dt} d\mu_0(\boldsymbol{x})$$

Back to the density of states



- **Extending the state-space:** Given $d\mu(\boldsymbol{q}) = e^{-U(\boldsymbol{q})}d\boldsymbol{q}$, let

$$d\boldsymbol{q}/dt = \boldsymbol{p}, \qquad d\boldsymbol{p}/dt = -\nabla U(\boldsymbol{q}) - \gamma \boldsymbol{p} \qquad (\gamma > 0)$$

- Then $Z = (2\pi)^{d/2} Z_q$ with

$$Z_q = \int_\Omega e^{-U(\boldsymbol{q})} d\boldsymbol{q} \quad \text{and} \quad Z = \int_{\Omega\times\mathbb{R}^d} e^{-H(\boldsymbol{q},\boldsymbol{p})} d\boldsymbol{q}d\boldsymbol{p} \quad \text{with} \quad H(\boldsymbol{q},\boldsymbol{p}) = \tfrac{1}{2}|\boldsymbol{p}|^2 + U(\boldsymbol{q})$$

- Using div $\boldsymbol{b} = d\gamma$, if in the previous formula we set

$$d\mu_0(\boldsymbol{q},\boldsymbol{p}) = V_0^{-1}\mathbf{1}(H(\boldsymbol{q},\boldsymbol{p}) < E_0)d\boldsymbol{q}d\boldsymbol{p}, \quad \text{and} \quad \phi(\boldsymbol{q},\boldsymbol{p}) = \mathbf{1}(H(\boldsymbol{q},\boldsymbol{p}) < E) \quad (E \le E_0)$$

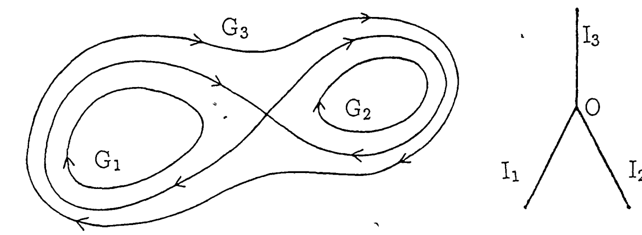
  we deduce

$$V(E) = \int_{H(\boldsymbol{q},\boldsymbol{p})<E} d\boldsymbol{q}d\boldsymbol{p} = \int_{H(\boldsymbol{q},\boldsymbol{p})<E_0} e^{-d\gamma(\tau_E(\boldsymbol{q},\boldsymbol{p}) - \tau_0(\boldsymbol{q},\boldsymbol{p}))} d\boldsymbol{q}d\boldsymbol{p}$$

  where

$$\tau_E(\boldsymbol{q},\boldsymbol{p}) = \inf\{|t| : H(\boldsymbol{q}(t),\boldsymbol{p}(t)) = E\}, \qquad \tau_0(\boldsymbol{q},\boldsymbol{p}) = \inf\{t < 0 : H(\boldsymbol{q}(t),\boldsymbol{p}(t)) = E_0\}$$

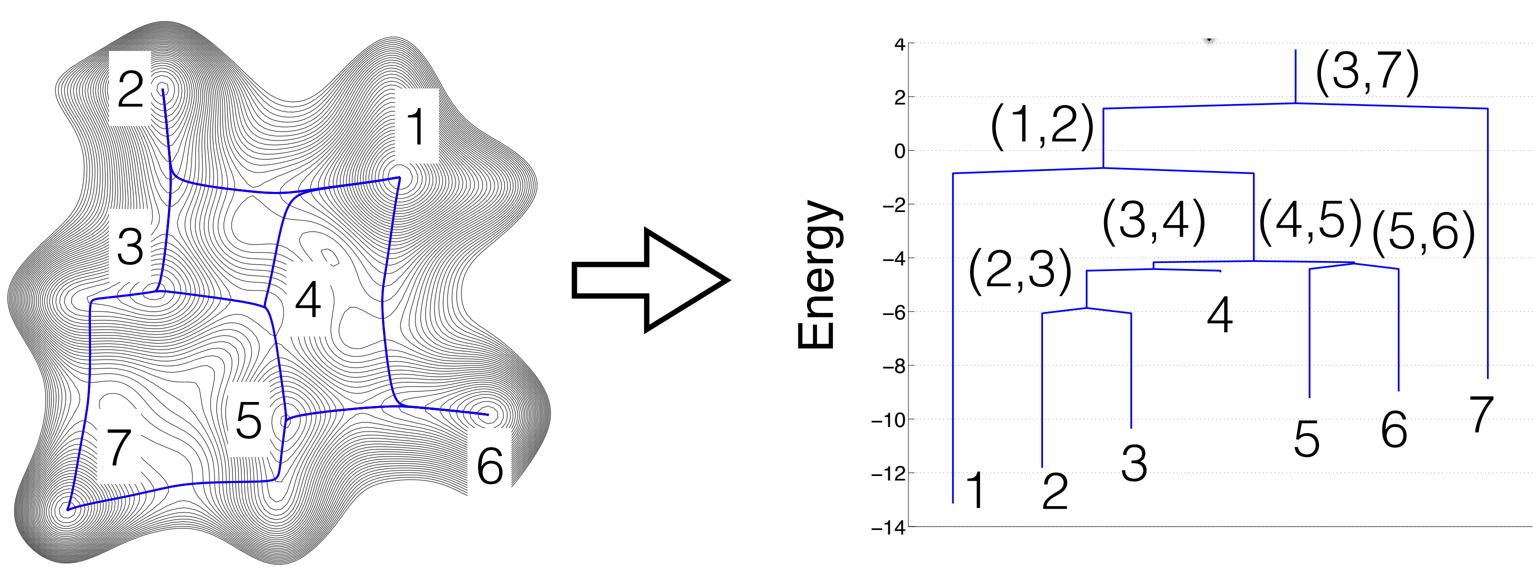
- That is, $V(E)/V_0$ is the expectation of $e^{-d\gamma(\tau_E(\boldsymbol{q},\boldsymbol{p}) - \tau_0(\boldsymbol{q},\boldsymbol{p}))}$ over initial data uniform in $H(\boldsymbol{q},\boldsymbol{p}) < E_0$.

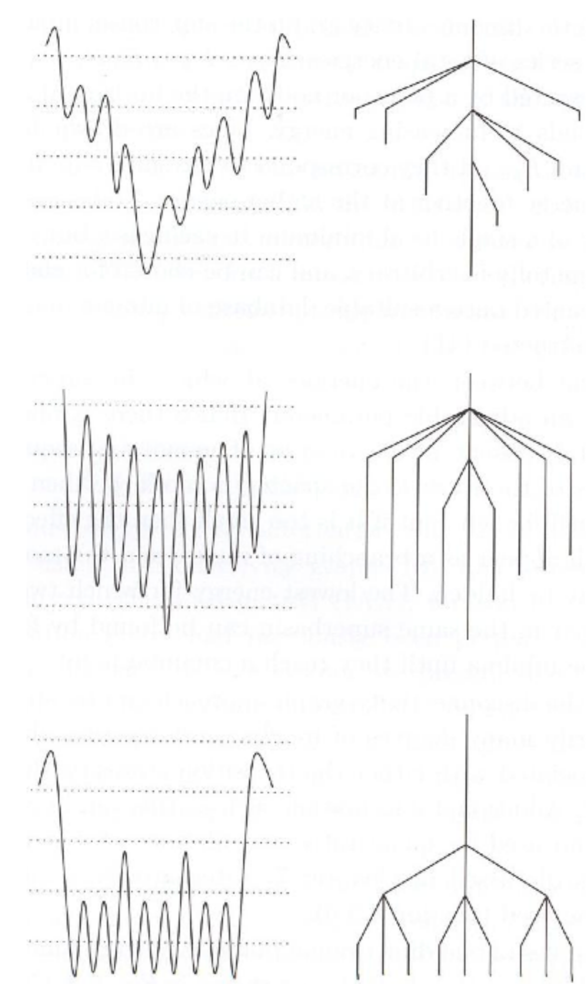


From Freidlin & Wentzell,
Annals of Prob, **21**, 2015 (1993)

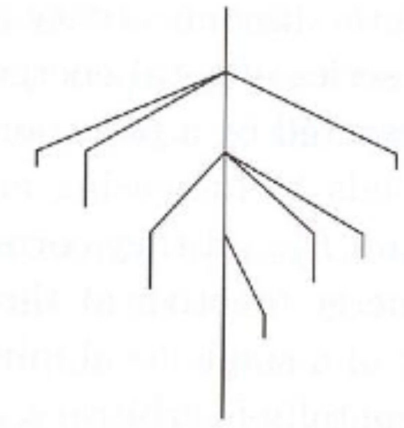
Variance of the estimator

- If we rescale time as $\gamma t \to t$ and let $\gamma \to 0$, the damped Hamiltonian dynamics reduces to descent on the Reeb (aka disconnectivity) graph of $H(\boldsymbol{q},\boldsymbol{p})$ (which is that of $U(\boldsymbol{q})$), that is:
  - ▷ On each branch of the graph $E(t) = H(\boldsymbol{q}(t),\boldsymbol{p}(t))$ satisfies a closed equation depending on the geometry of the underlying basin;
  - ▷ At every branching point, the trajectory picks a branch at random with a probability that also depends only on the geometry of the basins.





Variance of the estimator



- If we rescale time as $\gamma t \to t$ and let $\gamma \to 0$, the damped Hamiltonian dynamics reduces to descent on the Reeb (aka disconnectivity) graph of $H(\boldsymbol{q},\boldsymbol{p})$ (which is that of $U(\boldsymbol{q})$), that is:
  - ▷ On each branch of the graph $E(t) = H(\boldsymbol{q}(t),\boldsymbol{p}(t))$ satisfies a closed equation depending on the geometry of the underlying basin;
  - ▷ At every branching point, the trajectory picks a branch at random with a probability that also depends only on the geometry of the basins.

- Indexing for $j = 1,\ldots,M$ all the branches of the graph, let $\tau_j(E) > 0$ (possibly infinite) be the (deterministic) time it takes the trajectory to go from $H(\boldsymbol{q}(0),\boldsymbol{p}(0)) = E_0$ to $H(\boldsymbol{q}(t),\boldsymbol{p}(t)) = E$.

- Denote by $p_j > 0$ with $\sum_{j=1}^{M} p_j = 1$ the probability (computed over initial data uniformly drawn over $H(\boldsymbol{q},\boldsymbol{p}) < E_0$) that the trajectory takes branch j.

- Then $\tau_E(\boldsymbol{q},\boldsymbol{p}) - \tau_0(\boldsymbol{q},\boldsymbol{p}) = \tau_j(E)$ with probability p_j (i.e. depending only on whether the trajectory initiated at $(\boldsymbol{q},\boldsymbol{p})$ travels on branch j).

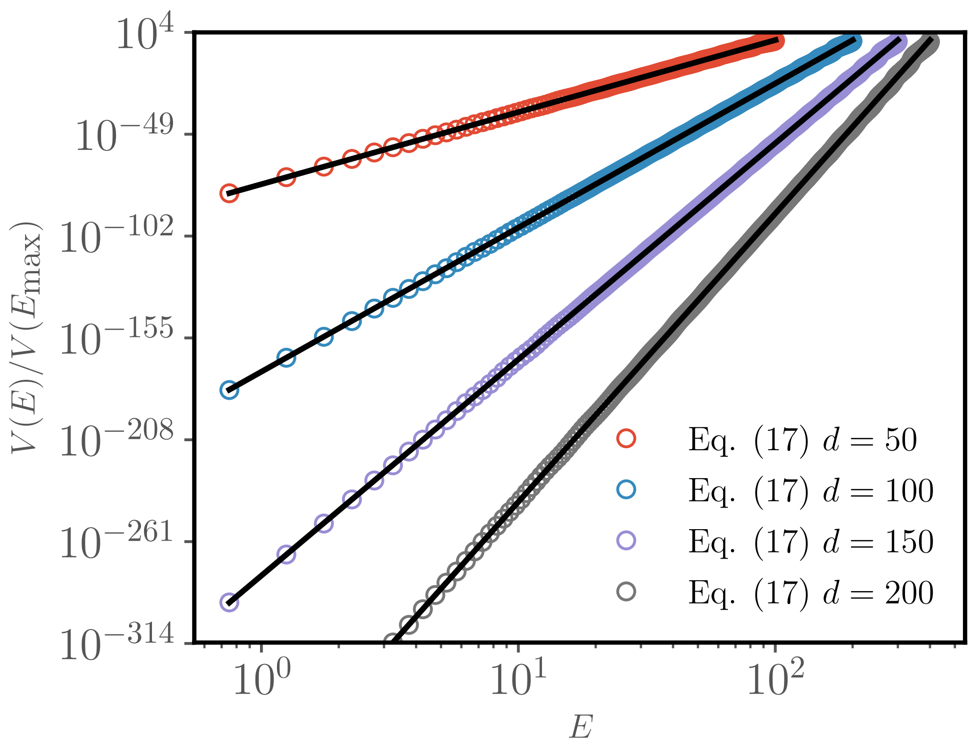
$$\text{mean} = V(E)/V_0 = \sum_{j=1}^{M} p_j e^{-\gamma d \tau_j(E)}, \qquad \text{var} = \sum_{j=1}^{M} p_j e^{-2\gamma d \tau_j(E)} - \text{mean}^2.$$

Note that the estimator is consistent and unbiased at every γ

$$\text{var} = \sum_{j=1}^{M} p_j e^{-2\gamma d \tau_j(E)} - \text{mean}^2$$

Quartic well example

- If there is only one well (μ_0 is monomodal), the variance is zero! A single trajectory does the job if γ is small enough



Results with a single trajectory for

$$U(\boldsymbol{q}) = \sum_{j=1}^{d} (\boldsymbol{b}_j \cdot \boldsymbol{q})^4$$

with some random $\boldsymbol{b}_j \in \mathbb{R}^d$. Here

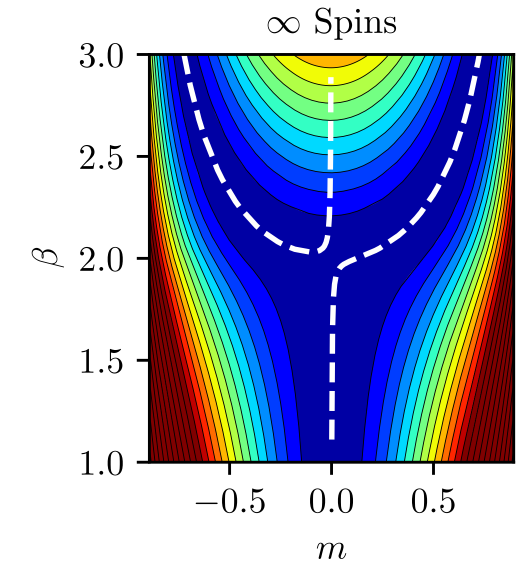
$$V(E)/V_0 = (E/E_0)^{3d/4}$$

and we took $\gamma = .1 \min_j |b_j|$.

Similar results for $U(\boldsymbol{q}) = \sum_{j=1}^{d} (\boldsymbol{b}_j \cdot \boldsymbol{q})^2$.

- Note that this implies a $O(\gamma^{-1})$ cost to integrate the equations to the relevant time scale, and how small γ needs to be depends on the dimension in general.

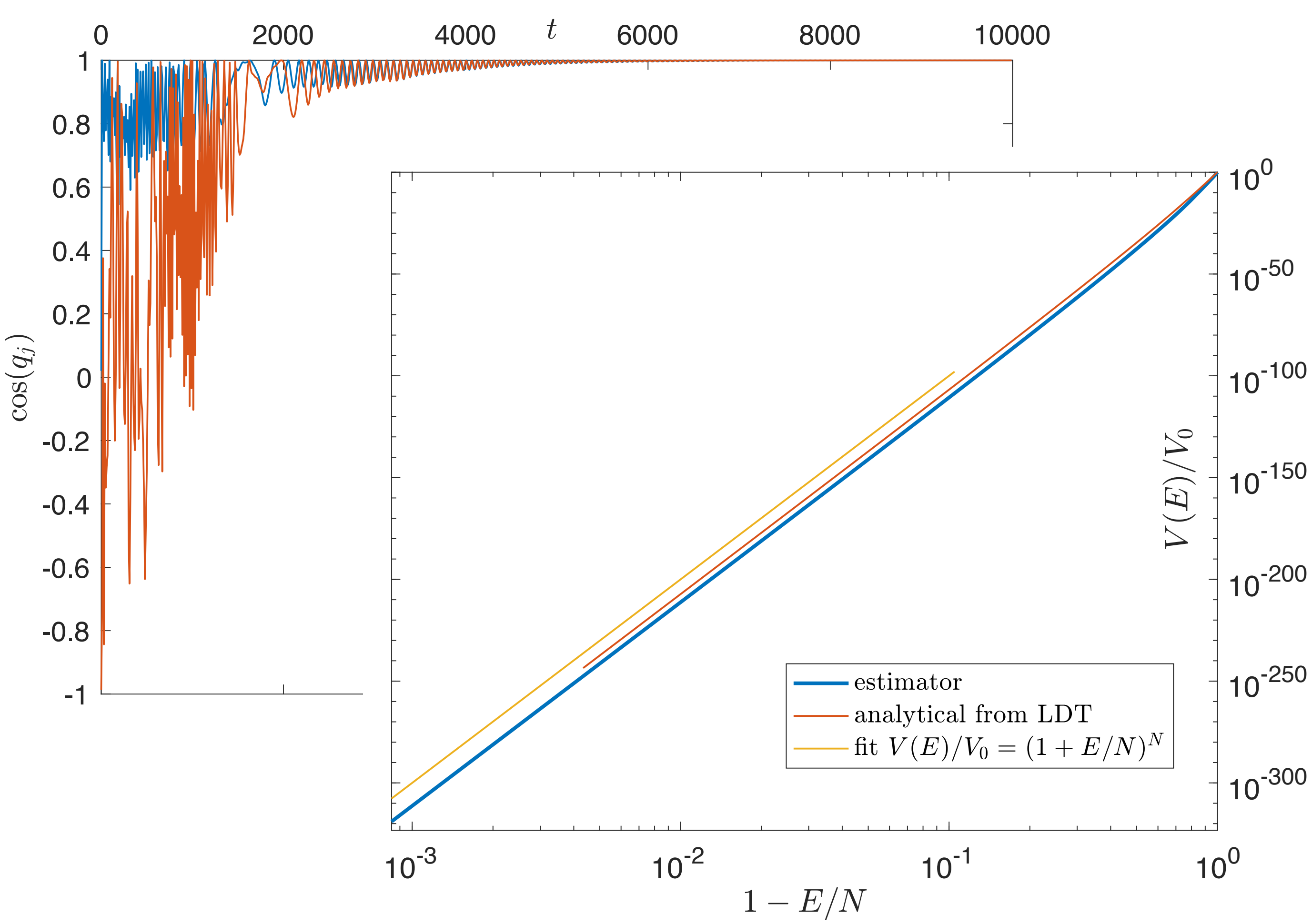
Curie-Weiss model



- Curie-Weiss model for N continuous spins $\sigma_i = \cos(q_i)$ with potential

$$U(\boldsymbol{q}) = -N^{-1} \sum_{i,j=1}^{N} \cos(q_i)\cos(q_j)$$

In the limit as $N \to \infty$, the model exhibits a second order phase transition at $\beta = 2$, because entropic effects that favor disorganized spin configurations dominate at high temperatures, whereas energetic effects that favor $\cos(q_j) = \pm 1$ dominate at low temperatures.

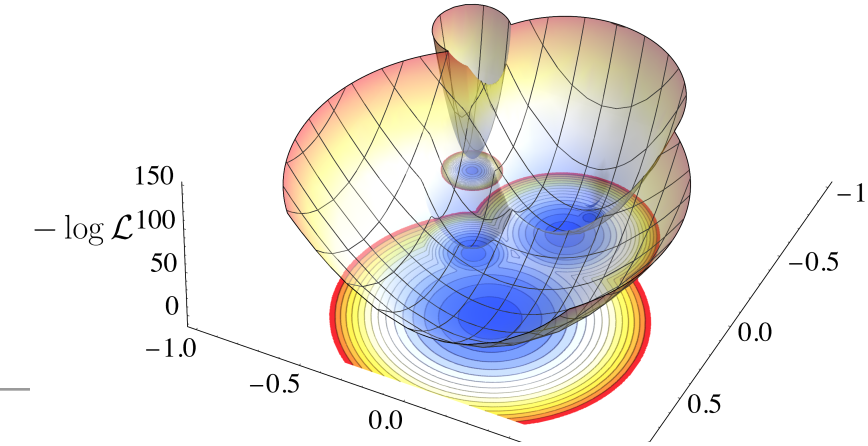


Result for $N = 100$ spins with a single trajectory run at $\gamma = 10^{-3}$.

- Correspondingly, the density of states decreases rapidly with the energy, since lowering $U(\boldsymbol{q})$ to its minimum value $E = -N$ requires to align the spins, and the number of aligned configurations is much less that the number of disorganized ones.
- This effect can be estimated analytically via LDT by estimating the entropy of the magnetization

$$m = N^{-1} \sum_{i=1}^{N} \cos(q_i).$$

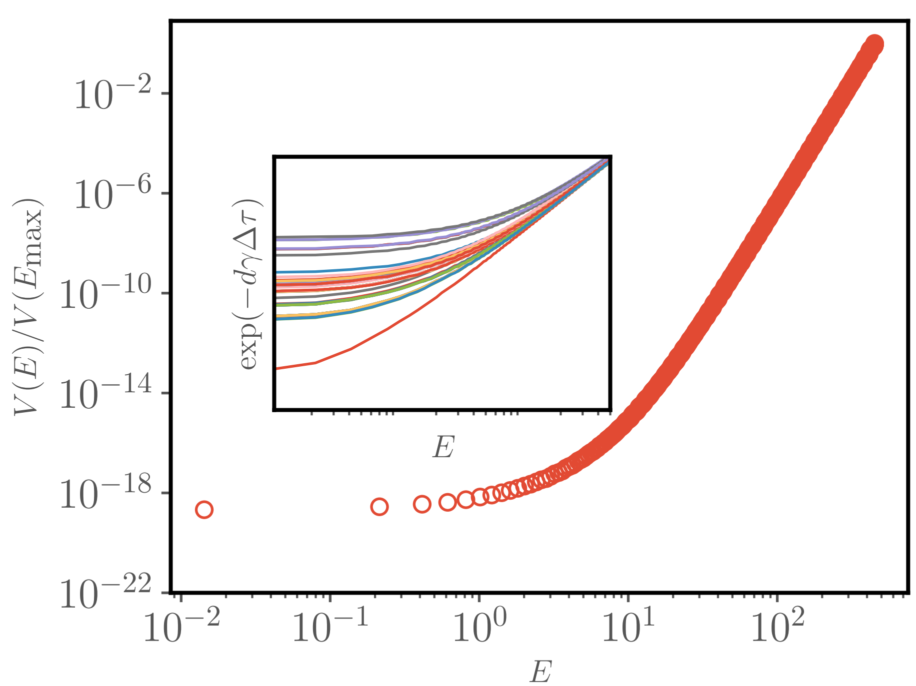
Bayesian inference test-case



- Mixture of Gaussians model as benchmark for inference problems. The model is defined as a mixture of n distributions in dimension d with amplitudes A_i, means $\boldsymbol{\mu}_i$ and covariances $\boldsymbol{\Sigma}_i$

$$L(\boldsymbol{x}) = \sum_{i=1}^{n} A_i \exp\left(-\tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)\right).$$

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.



Result with $n = 50$ wells with depths exponentially distributed in dimension $d = 10$, an example much more complex than previous benchmarks. In this regime, brute force Monte Carlo approaches fail dramatically. The volume estimator, with only 100 trajectories, reaches the deepest minima in a nontrivial estimation problem. Furthermore, the low energy volume estimates are reasonably accurate: we compute $Z = 17.41$ versus the exact result $Z = 17.10$.

Conclusions

- Estimator using trajectories that are guaranteed to visit regions of low energy / high likelihood around local minima of that would otherwise be difficult to select by direct sampling of the prior.

- Approach similar in spirit to Skilling's nested sampling method, but with the advantage that it does not require uniform sampling below / above every energy / likelihood level, which is required in nested sampling and is hard to implement in practice.

- Every trajectory contributes independently to the estimator, meaning that the implementation is trivially parallelizable.

- Variance can be estimated in the small friction limit, and depends on the complexity of the Reeb graph of the energy / likelihood.