Competing sources of variance reduction in parallel replica Monte Carlo, and optimization in the low temperature limit

Paul Dupuis

Division of Applied Mathematics
Brown University

IPAM

(J. Doll, M. Snarski, G.-J. Wu)

November 2017

Main Themes

- *Parallel tempering* and its limit *infinite swapping* are methods for accelerated Monte Carlo. They work by coupling reversible Markov chains with different "temperatures" to enhance the sampling properties of the ensemble. An important question is how to choose the temperatures.
- One source of improved sampling is straightforward from the construction–increased mobility of the lower temperature chains. A second emerges most clearly in the infinite swapping limit, and these two sources of variance reduction respond in different ways to temperature selection.
- One can explicitly identify the optimal temperature assignments in the low temperature limit, when sampling is most difficult.

Outline

- Some problems of interest
- Review of parallel tempering (aka replica exchange)
- The infinite swapping limit–symmetrized dynamics and a weighted empirical measure
- First source of variance reduction–lowered energy barriers
- Small detour–infinite swapping with independent and identically distributed samples–variance reduction originating due to weights
- Return to the Markov diffusion setting:
  - Small noise (low temperature) limit via Freidlin-Wentsell methods
  - Explicit solution to optimal temperature assignments in the small noise limit

Some problems of interest

Example 1 (physical sciences). Compute functionals with respect to a Gibbs measure of the form

$$\pi(dx) = e^{-V(x)/\tau} dx \Big/ Z(\tau),$$

where $V : \mathbb{R}^d \to \mathbb{R}$ is the potential of a (relatively) complex physical system. We use that $\pi(dx)$ is the stationary distribution of the solution to

$$dX = -\nabla V(X)dt + \sqrt{2\tau}dW,$$

as well as related discrete time models.

Some problems of interest

Example 1 (physical sciences). Compute functionals with respect to a Gibbs measure of the form

$$\pi(dx) = e^{-V(x)/\tau} dx \Big/ Z(\tau),$$

where $V : \mathbb{R}^d \to \mathbb{R}$ is the potential of a (relatively) complex physical system. We use that $\pi(dx)$ is the stationary distribution of the solution to

$$dX = -\nabla V(X)dt + \sqrt{2\tau}dW,$$

as well as related discrete time models. The function $V(x)$ is defined on a large space, and includes, e.g., various inter-molecular potentials.

Some problems of interest

Example 1 (physical sciences). Compute functionals with respect to a Gibbs measure of the form

$$\pi(dx) = e^{-V(x)/\tau} dx \Big/ Z(\tau),$$

where $V : \mathbb{R}^d \to \mathbb{R}$ is the potential of a (relatively) complex physical system. We use that $\pi(dx)$ is the stationary distribution of the solution to

$$dX = -\nabla V(X)dt + \sqrt{2\tau} dW,$$

as well as related discrete time models. The function $V(x)$ is defined on a large space, and includes, e.g., various inter-molecular potentials. Representative quantities of interest:

$$\text{average potential energy: } \int V(x) \frac{e^{-V(x)/\tau} dx}{Z(\tau)}$$

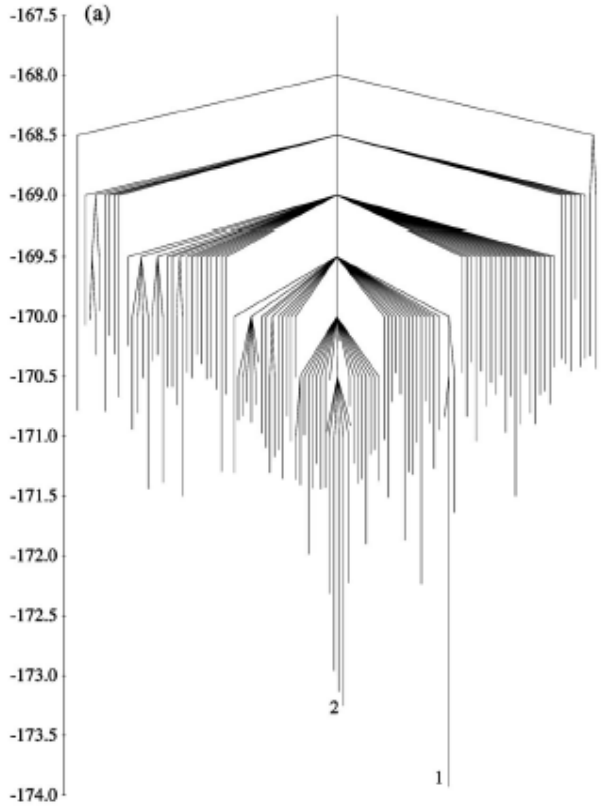
$$\text{heat capacity: } \int \left[V(x) - \int V(y) \frac{e^{-V(y)/\tau} dy}{Z(\tau)} \right]^2 \frac{e^{-V(x)/\tau} dx}{Z(\tau)}.$$

Some problems of interest

In general has a very complicated surface, with many deep and shallow local minima. An example: potential energy surface is the Lennard-Jones cluster of 38 atoms. This potential has $\approx 10^{14}$ local minima.

Some problems of interest

In general has a very complicated surface, with many deep and shallow local minima. An example: potential energy surface is the Lennard-Jones cluster of 38 atoms. This potential has $\approx 10^{14}$ local minima. The lowest 150 and their "connectivity" graph are as in figure (taken from Doyle, Miller & Wales, *JCP*, 1999).



Some problems of interest

Example 2 (path FIM). Consider reversible system with stationary distribution and transition density

$$\pi^\theta(dx) = e^{-V^\theta(x)/\tau} dx \Big/ Z^\theta(\tau), \quad p^\theta(x,y),$$

where $\theta \in \Theta$ are a collection of model parameters, and $p^\theta(x,y)$ exhibits metastability. Would like to compute the *path Fisher information matrix*

$$\int \pi^\theta(dx) \int p^\theta(x,y) \left[\nabla_\theta p^\theta(x,y)\right] \left[\nabla_\theta p^\theta(x,y)\right]^T dy$$

via Monte Carlo. Used to obtain sensitivity bounds for sensitivities in $\theta \in \Theta$.

Some problems of interest

Example 3 (combinatorics and counting). Counting problems involving subsets of very large discrete spaces, such as number of binary matrices with given row and column sums, graphs with degree sequence, etc.

Some problems of interest

Example 3 (combinatorics and counting). Counting problems involving subsets of very large discrete spaces, such as number of binary matrices with given row and column sums, graphs with degree sequence, etc. For matrices problem, let $r_i \leq n, i = 1, \ldots, m$, $c_j \leq m, j = 1, \ldots, n$, and define potential

$$V(x) = \sum_{i=1}^{m} \left| \sum_{j=1}^{n} x_{ij} - r_i \right|, \text{ where } x \in S \doteq \left\{ \{0,1\}^{n \times m} : \sum_{i=1}^{m} x_{ij} = c_j \right\}.$$

Some problems of interest

Example 3 (combinatorics and counting). Counting problems involving subsets of very large discrete spaces, such as number of binary matrices with given row and column sums, graphs with degree sequence, etc. For matrices problem, let $r_i \leq n, i = 1, \ldots, m$, $c_j \leq m, j = 1, \ldots, n$, and define potential

$$V(x) = \sum_{i=1}^{m} \left| \sum_{j=1}^{n} x_{ij} - r_i \right|, \text{ where } x \in S \doteq \left\{ \{0,1\}^{n \times m} : \sum_{i=1}^{m} x_{ij} = c_j \right\}.$$

Then can estimate $\#\{x \in S : V(x) = 0\}$ by approximating

$$\frac{1}{Z(\tau)} \sum_{x \in S_i} e^{-V(x)/\tau} = \frac{1}{Z(\tau)} |S_i| \, e^{-i/\tau}$$

for various sets $S_i \doteq \{x \in S : V(x) = i\}$ and small $\tau > 0$.

Some problems of interest

- Well-known corresponding Markov chains. We are concerned with case where structure of energy landscape $V(x)$ is complicated, with disconnected regions of importance. Very long simulation times needed for small τ.

Some problems of interest

- Well-known corresponding Markov chains. We are concerned with case where structure of energy landscape $V(x)$ is complicated, with disconnected regions of importance. Very long simulation times needed for small τ.
- Many other interesting applications have same features (e.g., Bayesian statistics), with V now depending on data.

Review of parallel tempering

Return to model from Example 1:

$$dX = -\nabla V(X)dt + \sqrt{2\tau}dW,$$

with computational approximation

$$\mu^T(dx) = \frac{1}{T}\int_0^T \delta_{X(t)}(dx)dt \in \mathcal{P}(\mathbb{R}^d).$$

Well known difficulty: the "rare event sampling problem," i.e., the infrequent moves between deep local minima of V.

Review of parallel tempering

How to speed up a single particle? Use "parallel tempering" (aka "replica exchange", due to Geyer, Swendsen and Wang).

Idea of parallel tempering, two temperatures.

Review of parallel tempering

How to speed up a single particle? Use "parallel tempering" (aka "replica exchange", due to Geyer, Swendsen and Wang).

Idea of parallel tempering, two temperatures. Besides $\tau_1 = \tau$, introduce higher temperature $\tau_2 > \tau_1$.

Review of parallel tempering

How to speed up a single particle? Use "parallel tempering" (aka "replica exchange", due to Geyer, Swendsen and Wang).

Idea of parallel tempering, two temperatures. Besides $\tau_1 = \tau$, introduce higher temperature $\tau_2 > \tau_1$. Thus

$$dX_1 = -\nabla V(X_1)dt + \sqrt{2\tau_1}dW_1$$
$$dX_2 = -\nabla V(X_2)dt + \sqrt{2\tau_2}dW_2,$$

with W_1 and W_2 independent. Then one obtains a Monte Carlo approximation to

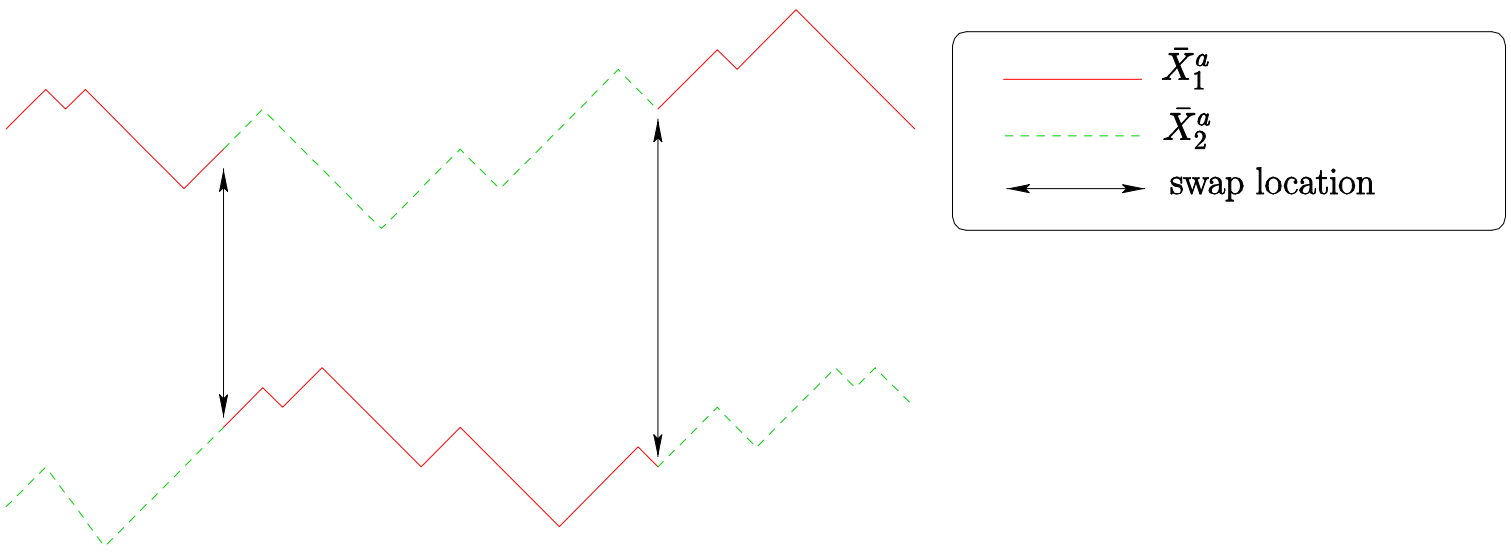
$$\pi(x_1, x_2) = e^{-\frac{V(x_1)}{\tau_1}} e^{-\frac{V(x_2)}{\tau_2}} \Big/ Z(\tau_1)Z(\tau_2).$$

Review of parallel tempering

Now introduce *swaps*, i.e., X_1 and X_2 *exchange locations* with state dependent intensity

$$ag(x_1, x_2) = a\left(1 \wedge \frac{\pi(x_2, x_1)}{\pi(x_1, x_2)}\right) = a\left(1 \wedge e^{-\left[\frac{V(x_1)}{\tau_1} + \frac{V(x_2)}{\tau_2}\right] + \left[\frac{V(x_2)}{\tau_1} + \frac{V(x_1)}{\tau_2}\right]}\right),$$

with $a > 0$ the "swap rate."



Review of parallel tempering

Now have a *Markov jump-diffusion*. Easy to check: with this swapping intensity still have detailed balance, and thus

$$\pi^a(x_1, x_2) = \pi(x_1, x_2) = e^{-\frac{V(x_1)}{\tau_1}} e^{-\frac{V(x_2)}{\tau_2}} \Big/ Z(\tau_1) Z(\tau_2).$$

Review of parallel tempering

Now have a *Markov jump-diffusion*. Easy to check: with this swapping intensity still have detailed balance, and thus

$$\pi^a(x_1, x_2) = \pi(x_1, x_2) = e^{-\frac{V(x_1)}{\tau_1}} e^{-\frac{V(x_2)}{\tau_2}} \Big/ Z(\tau_1) Z(\tau_2).$$

Higher temperature $\tau_2 > \tau_1$ $\sim$ greater diffusivity of X_2^a

$\sim$ easier communication for X_2^a

$\sim$ passed to X_1^a via swaps

This helps overcome the "rare event sampling problem."

The infinite swapping limit

Various rates of convergence (large deviation empirical measure rate, asymptotic variance) optimized by letting $a \to \infty$.

The infinite swapping limit

Various rates of convergence (large deviation empirical measure rate, asymptotic variance) optimized by letting $a \to \infty$.This suggests one consider the infinite swapping limit $a \to \infty$, except

The infinite swapping limit

Various rates of convergence (large deviation empirical measure rate, asymptotic variance) optimized by letting $a \to \infty$.This suggests one consider the infinite swapping limit $a \to \infty$, except

- if a is large but finite almost all computational effort is directed at swap attempts, rather than diffusion dynamics,

The infinite swapping limit

Various rates of convergence (large deviation empirical measure rate, asymptotic variance) optimized by letting $a \to \infty$.This suggests one consider the infinite swapping limit $a \to \infty$, except

- if a is large but finite almost all computational effort is directed at swap attempts, rather than diffusion dynamics,
- if $a \to \infty$ then limit process not well defined (no tightness).

The infinite swapping limit

Various rates of convergence (large deviation empirical measure rate, asymptotic variance) optimized by letting $a \to \infty$.This suggests one consider the infinite swapping limit $a \to \infty$, except

- if a is large but finite almost all computational effort is directed at swap attempts, rather than diffusion dynamics,
- if $a \to \infty$ then limit process not well defined (no tightness).

The infinite swapping limit

Various rates of convergence (large deviation empirical measure rate, asymptotic variance) optimized by letting $a \to \infty$.This suggests one consider the infinite swapping limit $a \to \infty$, except

- if a is large but finite almost all computational effort is directed at swap attempts, rather than diffusion dynamics,
- if $a \to \infty$ then limit process not well defined (no tightness).

An alternative perspective: rather than swap particles, swap temperatures, and use "weighted" empirical measure.

The infinite swapping limit

Various rates of convergence (large deviation empirical measure rate, asymptotic variance) optimized by letting $a \to \infty$.This suggests one consider the infinite swapping limit $a \to \infty$, except

- if a is large but finite almost all computational effort is directed at swap attempts, rather than diffusion dynamics,
- if $a \to \infty$ then limit process not well defined (no tightness).

An alternative perspective: rather than swap particles, swap temperatures, and use "weighted" empirical measure.

Particle swapping. Process:

$$(X_1^a, X_2^a)\,,$$

Approximation to $\pi(dx)$:

$$\frac{1}{T}\int_0^T \delta_{(X_1^a, X_2^a)}(dx)dt$$

The infinite swapping limit

Temperature swapping.

The infinite swapping limit

Temperature swapping. Process:

$$dY_1^a = -\nabla V(Y_1^a)dt + \sqrt{2r_1(Z^a)}dW_1$$
$$dY_2^a = -\nabla V(Y_2^a)dt + \sqrt{2r_2(Z^a)}dW_2,$$

where $r(Z^a(t))$ jumps between τ_1 and τ_2 with intensity $ag(Y_1^a(t), Y_2^a(t))$.

The infinite swapping limit

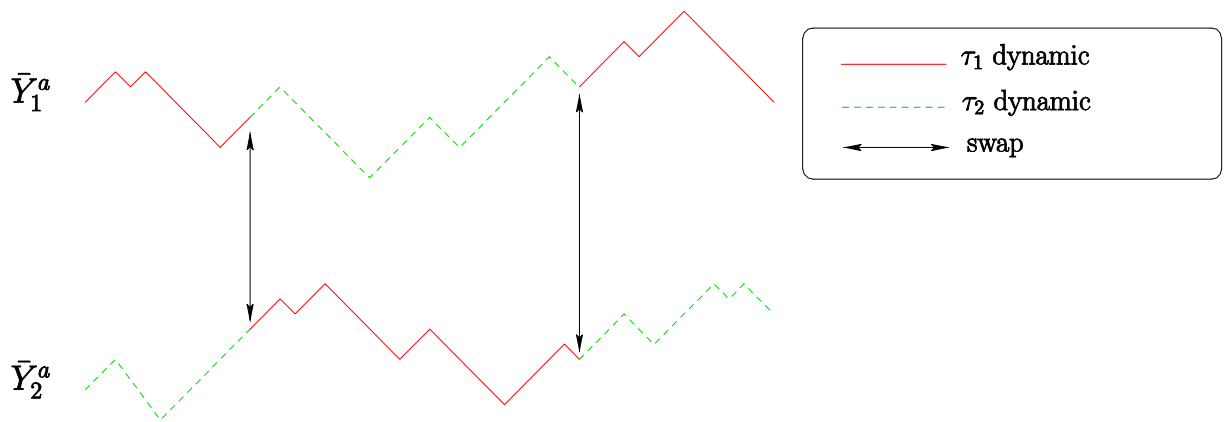
Temperature swapping. Process:

$$dY_1^a = -\nabla V(Y_1^a)dt + \sqrt{2r_1(Z^a)}dW_1$$
$$dY_2^a = -\nabla V(Y_2^a)dt + \sqrt{2r_2(Z^a)}dW_2,$$

where $r(Z^a(t))$ jumps between τ_1 and τ_2 with intensity $ag(Y_1^a(t), Y_2^a(t))$.

Approximation to $\pi(dx)$:

$$\frac{1}{T}\int_0^T \left[1_{\{0\}}(Z^a)\delta_{(Y_1^a,Y_2^a)}(dx) + 1_{\{1\}}(Z^a)\delta_{(Y_2^a,Y_1^a)}(dx)\right] dt.$$



The infinite swapping limit

The advantage is a well defined weak limit as $a \to \infty$:

The infinite swapping limit

The advantage is a well defined weak limit as $a \to \infty$:

$$dY_1 = -\nabla V(Y_1)dt + \sqrt{2\tau_1\rho_1(Y_1, Y_2) + 2\tau_2\rho_2(Y_1, Y_2)}dW_1$$
$$dY_2 = -\nabla V(Y_2)dt + \sqrt{2\tau_2\rho_1(Y_1, Y_2) + 2\tau_1\rho_2(Y_1, Y_2)}dW_2,$$

$$\eta^T(dx) = \frac{1}{T}\int_0^T \left[\rho_1(Y_1, Y_2)\delta_{(Y_1,Y_2)} + \rho_2(Y_1, Y_2)\delta_{(Y_2,Y_1)}\right] ds,$$

and

$$\rho_1(x_1, x_2) = \frac{e^{-\left[\frac{V(x_1)}{\tau_1} + \frac{V(x_2)}{\tau_2}\right]}}{Z_\rho(x_1, x_2)}, \quad \rho_2(x_1, x_2) = \frac{e^{-\left[\frac{V(x_2)}{\tau_1} + \frac{V(x_1)}{\tau_2}\right]}}{Z_\rho(x_1, x_2)}.$$

The infinite swapping limit

The advantage is a well defined weak limit as $a \to \infty$:

$$dY_1 = -\nabla V(Y_1)dt + \sqrt{2\tau_1\rho_1(Y_1, Y_2) + 2\tau_2\rho_2(Y_1, Y_2)}dW_1$$
$$dY_2 = -\nabla V(Y_2)dt + \sqrt{2\tau_2\rho_1(Y_1, Y_2) + 2\tau_1\rho_2(Y_1, Y_2)}dW_2,$$

$$\eta^T(dx) = \frac{1}{T}\int_0^T \left[\rho_1(Y_1, Y_2)\delta_{(Y_1,Y_2)} + \rho_2(Y_1, Y_2)\delta_{(Y_2,Y_1)}\right] ds,$$

and

$$\rho_1(x_1, x_2) = \frac{e^{-\left[\frac{V(x_1)}{\tau_1} + \frac{V(x_2)}{\tau_2}\right]}}{Z_\rho(x_1, x_2)}, \quad \rho_2(x_1, x_2) = \frac{e^{-\left[\frac{V(x_2)}{\tau_1} + \frac{V(x_1)}{\tau_2}\right]}}{Z_\rho(x_1, x_2)}.$$

For generalization to K temperatures $\tau_K \geq \cdots \geq \tau_1$ one must compute ρ weights of all permutations of $(x_1, \ldots, x_K)$, practical implementation requires PINS (partial INS) for $K \geq 7$.

Variance reduction–lowered energy barriers

How do PT and INS improve sampling?

- The invariant distribution of (Y_1, Y_2) is the symmetrized measure

$$\frac{1}{2}\left[\pi(x_1, x_2) + \pi(x_2, x_1)\right]$$
$$= \frac{1}{2Z(\tau_1)Z(\tau_2)}\left[e^{-\frac{V(x_1)}{\tau_1}} e^{-\frac{V(x_2)}{\tau_2}} + e^{-\frac{V(x_2)}{\tau_1}} e^{-\frac{V(x_1)}{\tau_2}}\right].$$

  The "implied potential"

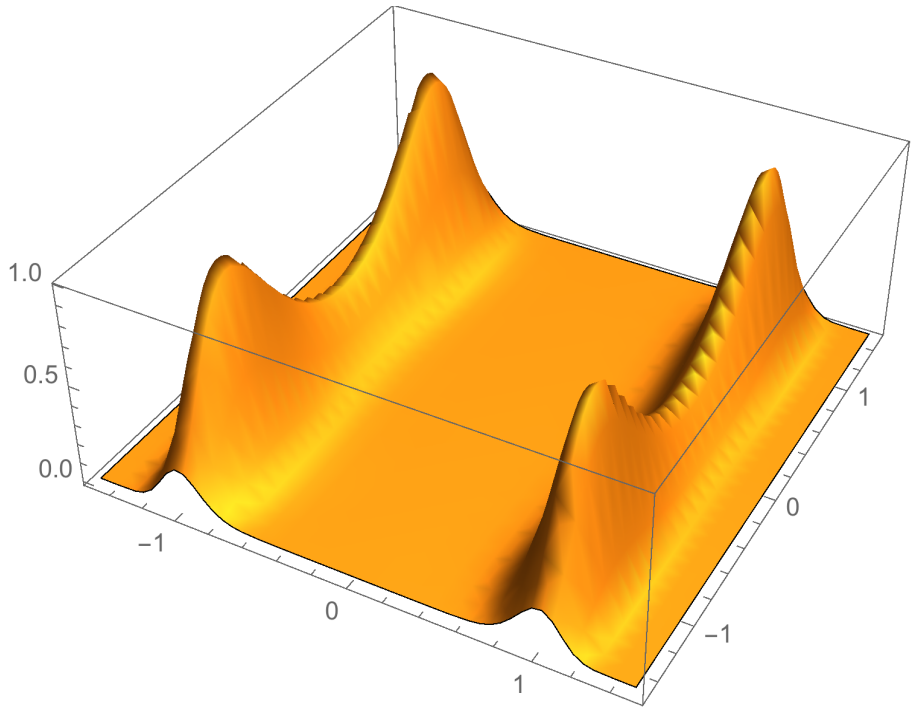
$$-\log\left[e^{-\frac{V(x_1)}{\tau_1}} e^{-\frac{V(x_2)}{\tau_2}} + e^{-\frac{V(x_2)}{\tau_1}} e^{-\frac{V(x_1)}{\tau_2}}\right]$$

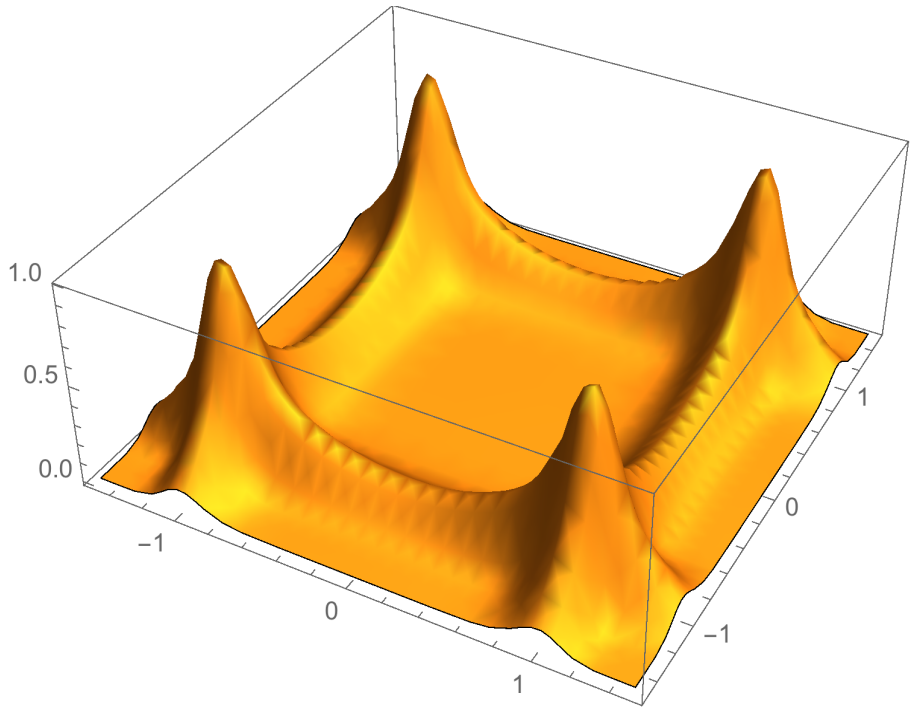
  has lower energy barriers than the original

$$\frac{V(x_1)}{\tau_1} + \frac{V(x_2)}{\tau_2}.$$

Variance reduction–lowered energy barriers

E.g., densities when $V(x)$ is a double well, orginal product density and density of implied potential:





However, to simply minimize the maximum barrier height one should let $\tau_2 \longrightarrow \infty$.

Infinite swapping with IID samples

To remove dynamics and identify other sources of variance reduction, *temporarily* assume can generate iid samples from stationary distribution

$$\pi_\tau(dx) = \frac{1}{Z(\tau)} e^{-V(x)/\tau} dx,$$

and hence iid samples (Y_1, Y_2) from symmetrized distribution

$$\frac{1}{2}\left[\pi_{\tau_1}(dy_1)\pi_{\tau_2}(dy_2) + \pi_{\tau_2}(dy_1)\pi_{\tau_1}(dy_2)\right].$$

Infinite swapping with IID samples

To remove dynamics and identify other sources of variance reduction, *temporarily* assume can generate iid samples from stationary distribution

$$\pi_\tau(dx) = \frac{1}{Z(\tau)} e^{-V(x)/\tau} dx,$$

and hence iid samples (Y_1, Y_2) from symmetrized distribution

$$\frac{1}{2}\left[\pi_{\tau_1}(dy_1)\pi_{\tau_2}(dy_2) + \pi_{\tau_2}(dy_1)\pi_{\tau_1}(dy_2)\right].$$

To obtain an unbiased estimator for integrals wrt $\pi_{\tau_1}(dx_1)\pi_{\tau_2}(dx_2)$ must again use weighted samples (weighted empirical measure):

$$\rho_1(Y_1, Y_2)\delta_{(Y_1,Y_2)} + \rho_2(Y_1, Y_2)\delta_{(Y_2,Y_1)}$$

with as before

$$\rho_1(x_1, x_2) = \frac{e^{-\left[\frac{V(x_1)}{\tau_1} + \frac{V(x_2)}{\tau_2}\right]}}{Z_\rho(x_1, x_2)}, \quad \rho_2(x_1, x_2) = \frac{e^{-\left[\frac{V(x_2)}{\tau_1} + \frac{V(x_1)}{\tau_2}\right]}}{Z_\rho(x_1, x_2)}.$$

Infinite swapping with IID samples

- Is this useful? Specifically, if we can compute the ρ's, is it better than standard MC?

Infinite swapping with IID samples

- Is this useful? Specifically, if we can compute the ρ's, is it better than standard MC?
- The answer is yes, and reason is analogous to why (well designed) importance sampling improves MC.

Infinite swapping with IID samples

- Is this useful? Specifically, if we can compute the ρ's, is it better than standard MC?
- The answer is yes, and reason is analogous to why (well designed) importance sampling improves MC.
- Note that probabilities

$$\pi_\tau(dx) = \frac{1}{Z(\tau)} e^{-V(x)/\tau} dx$$

have an obvious large deviation property when $\tau \to 0$. If $A \subset \mathbb{R}^d$ does not contain global minimum of V and ∂A "nice", then $\pi_\tau(A)$ decays exponentially in τ: $\pi_\tau(A) \approx \exp - \left[\inf_{x \in A} V(x)\right] / \tau$.

Infinite swapping with IID samples

- Is this useful? Specifically, if we can compute the ρ's, is it better than standard MC?
- The answer is yes, and reason is analogous to why (well designed) importance sampling improves MC.
- Note that probabilities

$$\pi_\tau(dx) = \frac{1}{Z(\tau)} e^{-V(x)/\tau} dx$$

have an obvious large deviation property when $\tau \to 0$. If $A \subset \mathbb{R}^d$ does not contain global minimum of V and ∂A "nice", then $\pi_\tau(A)$ decays exponentially in τ: $\pi_\tau(A) \approx \exp - \left[\inf_{x \in A} V(x)\right] / \tau$.

- How to assess performance for approximating probabilities $\pi_\tau(A)$ and expected values

$$\int e^{-\frac{1}{\tau} F(x)} \pi_\tau(dx)$$

via MC?

Infinite swapping with IID samples

- For standard Monte Carlo we average iid copies of $1_{\{X \in A\}}$. One needs $K \approx e^{V^*/\tau}$, $V^* = [\inf_{x \in A} V(x)]$ samples for bounded relative error.

Infinite swapping with IID samples

- For standard Monte Carlo we average iid copies of $1_{\{X \in A\}}$. One needs $K \approx e^{V^*/\tau}$, $V^* = [\inf_{x \in A} V(x)]$ samples for bounded relative error.
- Alternative approach: construct iid random variables $s_1^\tau, \ldots, s_K^\tau$ with $E s_1^\tau = \pi_\tau(A)$ and use the unbiased estimator

$$\hat{q}_{\tau,K} \doteq \frac{s_1^\tau + \cdots + s_K^\tau}{K}.$$

Infinite swapping with IID samples

- For standard Monte Carlo we average iid copies of $1_{\{X \in A\}}$. One needs $K \approx e^{V^*/\tau}$, $V^* = [\inf_{x \in A} V(x)]$ samples for bounded relative error.
- Alternative approach: construct iid random variables $s_1^\tau, \ldots, s_K^\tau$ with $E s_1^\tau = \pi_\tau(A)$ and use the unbiased estimator

$$\hat{q}_{\tau,K} \doteq \frac{s_1^\tau + \cdots + s_K^\tau}{K}.$$

- Performance determined by $Var(s_1^\tau)$, and since unbiased by $E\,(s_1^\tau)^2$.

Infinite swapping with IID samples

- For standard Monte Carlo we average iid copies of $1_{\{X \in A\}}$. One needs $K \approx e^{V^*/\tau}$, $V^* = [\inf_{x \in A} V(x)]$ samples for bounded relative error.
- Alternative approach: construct iid random variables $s_1^\tau, \ldots, s_K^\tau$ with $E s_1^\tau = \pi_\tau(A)$ and use the unbiased estimator

$$\hat{q}_{\tau,K} \doteq \frac{s_1^\tau + \cdots + s_K^\tau}{K}.$$

- Performance determined by $Var(s_1^\tau)$, and since unbiased by $E\,(s_1^\tau)^2$.
- By Jensen's inequality

$$-\tau \log E\,(s_1^\tau)^2 \leq -2\tau \log E s_1^\tau = -2\tau \log \pi_\tau(A) \to 2V^*.$$

Infinite swapping with IID samples

- For standard Monte Carlo we average iid copies of $1_{\{X \in A\}}$. One needs $K \approx e^{V^*/\tau}$, $V^* = [\inf_{x \in A} V(x)]$ samples for bounded relative error.
- Alternative approach: construct iid random variables $s_1^\tau, \ldots, s_K^\tau$ with $E s_1^\tau = \pi_\tau(A)$ and use the unbiased estimator

$$\hat{q}_{\tau,K} \doteq \frac{s_1^\tau + \cdots + s_K^\tau}{K}.$$

- Performance determined by $Var(s_1^\tau)$, and since unbiased by $E\,(s_1^\tau)^2$.
- By Jensen's inequality

$$-\tau \log E\,(s_1^\tau)^2 \leq -2\tau \log E s_1^\tau = -2\tau \log \pi_\tau(A) \to 2V^*.$$

- An estimator is called *asymptotically efficient* if

$$\liminf_{\tau \to 0} -\tau \log E\,(s_1^\tau)^2 \geq \ 2V^*.$$

Infinite swapping with IID samples

- For standard Monte Carlo we average iid copies of $1_{\{X \in A\}}$. One needs $K \approx e^{V^*/\tau}$, $V^* = [\inf_{x \in A} V(x)]$ samples for bounded relative error.
- Alternative approach: construct iid random variables $s_1^\tau, \ldots, s_K^\tau$ with $E s_1^\tau = \pi_\tau(A)$ and use the unbiased estimator

$$\hat{q}_{\tau,K} \doteq \frac{s_1^\tau + \cdots + s_K^\tau}{K}.$$

- Performance determined by $Var(s_1^\tau)$, and since unbiased by $E\,(s_1^\tau)^2$.
- By Jensen's inequality

$$-\tau \log E\,(s_1^\tau)^2 \leq -2\tau \log E s_1^\tau = -2\tau \log \pi_\tau(A) \to 2V^*.$$

- An estimator is called *asymptotically efficient* if

$$\liminf_{\tau \to 0} -\tau \log E\,(s_1^\tau)^2 \geq \; 2V^*.$$

- For standard MC

$$\lim_{\tau \to 0} -\tau \log E\left(1_{\{X \in A\}}\right)^2 = \; V^*.$$

Infinite swapping with IID samples

To approximate $\pi_\tau(A)$, we propose the INS estimator based on IID samples:

$$\begin{aligned} s^\tau &= \rho_1(Y_1^\tau, Y_2^\tau)\delta_{(Y_1^\tau, Y_2^\tau)}(A, \mathbb{R}^d) + \rho_2(Y_1^\tau, Y_2^\tau)\delta_{(Y_2^\tau, Y_1^\tau)}(A, \mathbb{R}^d) \\ &= \rho_1(Y_1^\tau, Y_2^\tau)1_{\{Y_1^\tau \in A\}} + \rho_2(Y_1^\tau, Y_2^\tau)1_{\{Y_2^\tau \in A\}}, \end{aligned}$$

where (Y_1^τ, Y_2^τ) sampled from

$$\frac{1}{2}\left[\pi_\tau(dy_1)\pi_{\tau r}(dy_2) + \pi_{\tau r}(dy_1)\pi_\tau(dy_2)\right],$$

with $r \geq 1$, so that $\tau_2 = \tau r \geq \tau = \tau_1$.

Infinite swapping with IID samples

To approximate $\pi_\tau(A)$, we propose the INS estimator based on IID samples:

$$\begin{aligned} s^\tau &= \rho_1(Y_1^\tau, Y_2^\tau)\delta_{(Y_1^\tau, Y_2^\tau)}(A, \mathbb{R}^d) + \rho_2(Y_1^\tau, Y_2^\tau)\delta_{(Y_2^\tau, Y_1^\tau)}(A, \mathbb{R}^d) \\ &= \rho_1(Y_1^\tau, Y_2^\tau)1_{\{Y_1^\tau \in A\}} + \rho_2(Y_1^\tau, Y_2^\tau)1_{\{Y_2^\tau \in A\}}, \end{aligned}$$

where (Y_1^τ, Y_2^τ) sampled from

$$\frac{1}{2}\left[\pi_\tau(dy_1)\pi_{\tau r}(dy_2) + \pi_{\tau r}(dy_1)\pi_\tau(dy_2)\right],$$

with $r \geq 1$, so that $\tau_2 = \tau r \geq \tau = \tau_1$. Can also consider analogous estimator based on K temperatures

$$\tau_K = \tau r_K \text{ with } r_K \geq r_{K-1} \geq \cdots \geq r_2 \geq r_1 = 1.$$

Infinite swapping with IID samples

Theorem

For the INS estimator based on K temperatures

$$\lim_{\tau \to 0} -\tau \log E\left(s^{\tau}\right)^2 = M(r_1, \ldots, r_K) \left[\inf_{x \in A} V(x)\right],$$

where $M(r_1, \ldots, r_K)$ solves the LP

$$M(r_1, \ldots, r_K) = \inf_{\{I : I_1 = 1, I_k \in [0,1] \text{ for } k=2,\ldots,K\}} \left[2 \sum_{j=1}^{K} \frac{1}{r_j} I_j - \min_{\sigma \in \Sigma_K} \left\{ \sum_{j=1}^{K} \frac{1}{r_j} I_{\sigma(j)} \right\}\right].$$

Moreover the supremum over $r_K \geq r_{K-1} \geq \cdots \geq r_2 \geq r_1$ is $M(r_1, \ldots, r_K) = 2 - (1/2)^K$ and is uniquely achieved at $r_k = 2^{k-1}$.

Infinite swapping with IID samples

Theorem

For the INS estimator based on K temperatures

$$\lim_{\tau \to 0} -\tau \log E\left(s^{\tau}\right)^2 = M(r_1, \ldots, r_K) \left[\inf_{x \in A} V(x)\right],$$

where $M(r_1, \ldots, r_K)$ solves the LP

$$M(r_1, \ldots, r_K) = \inf_{\{I : I_1 = 1, I_k \in [0,1] \text{ for } k = 2, \ldots, K\}} \left[2 \sum_{j=1}^{K} \frac{1}{r_j} I_j - \min_{\sigma \in \Sigma_K} \left\{ \sum_{j=1}^{K} \frac{1}{r_j} I_{\sigma(j)} \right\}\right].$$

Moreover the supremum over $r_K \geq r_{K-1} \geq \cdots \geq r_2 \geq r_1$ is $M(r_1, \ldots, r_K) = 2 - (1/2)^K$ and is uniquely achieved at $r_k = 2^{k-1}$.

Thus $K = 5$ temperatures gives 1.96875 or 98.4% of optimal value.
Analogous result for functionals $\int e^{-\frac{1}{\tau} F(x)} \pi_{\tau}(dx)$.

Infinite swapping with IID samples

Why does it work? The weights ρ act much like likelihood ratio in importance sampling.

Infinite swapping with IID samples

Why does it work? The weights ρ act much like likelihood ratio in importance sampling. For $K = 2$ and small $\tau > 0$, there are three types of outcomes, [recall $V^* = \inf_{x \in A} V(x)$]:

Infinite swapping with IID samples

Why does it work? The weights ρ act much like likelihood ratio in importance sampling. For $K = 2$ and small $\tau > 0$, there are three types of outcomes, [recall $V^* = \inf_{x \in A} V(x)$]:

- $s^\tau = 1$ when $(Y_1^\tau, Y_2^\tau) \in A \times A$, which occurs with approximate probability $P\left((Y_1^\tau, Y_2^\tau) \in A \times A\right) \approx e^{-\frac{1}{\tau} V^*} \cdot e^{-\frac{1}{\tau r} V^*} = e^{-\frac{1}{\tau}\left(1+\frac{1}{r}\right) V^*}$.

Infinite swapping with IID samples

Why does it work? The weights ρ act much like likelihood ratio in importance sampling. For $K = 2$ and small $\tau > 0$, there are three types of outcomes, [recall $V^* = \inf_{x \in A} V(x)$]:

- $s^\tau = 1$ when $(Y_1^\tau, Y_2^\tau) \in A \times A$, which occurs with approximate probability $P\left((Y_1^\tau, Y_2^\tau) \in A \times A\right) \approx e^{-\frac{1}{\tau} V^*} \cdot e^{-\frac{1}{\tau r} V^*} = e^{-\frac{1}{\tau}\left(1+\frac{1}{r}\right) V^*}$.
- $s^\tau = 0$ when $(Y_1^\tau, Y_2^\tau) \in A^c \times A^c$, with approximate probability $P\left((Y_1^\tau, Y_2^\tau) \in A^c \times A^c\right) \approx 1$.

Infinite swapping with IID samples

Why does it work? The weights ρ act much like likelihood ratio in importance sampling. For $K = 2$ and small $\tau > 0$, there are three types of outcomes, [recall $V^* = \inf_{x \in A} V(x)$]:

- $s^\tau = 1$ when $(Y_1^\tau, Y_2^\tau) \in A \times A$, which occurs with approximate probability $P((Y_1^\tau, Y_2^\tau) \in A \times A) \approx e^{-\frac{1}{\tau}V^*} \cdot e^{-\frac{1}{\tau r}V^*} = e^{-\frac{1}{\tau}\left(1+\frac{1}{r}\right)V^*}$.
- $s^\tau = 0$ when $(Y_1^\tau, Y_2^\tau) \in A^c \times A^c$, with approximate probability $P((Y_1^\tau, Y_2^\tau) \in A^c \times A^c) \approx 1$.
- $s^\tau \approx \rho_1(Y_1^\tau, Y_2^\tau)$ when $(Y_1^\tau, Y_2^\tau) \in A \times A^c$ or $A^c \times A$, with approximate probability

$$\begin{aligned} & P((Y_1^\tau, Y_2^\tau) \in A \times A^c) + P((Y_1^\tau, Y_2^\tau) \in A^c \times A) \\ & \quad \approx e^{-\frac{1}{\tau}V^*} + e^{-\frac{1}{\tau r}V^*} \\ & \quad \approx e^{-\frac{1}{\tau r}V^*}. \end{aligned}$$

Infinite swapping with IID samples

The definition of ρ_1 gives

$$\rho_1\left(Y_1^\tau, Y_2^\tau\right) \approx \frac{e^{-\frac{1}{\tau}V^*}}{e^{-\frac{1}{\tau}V^*} + e^{-\frac{1}{\tau r}V^*}} \approx e^{-\frac{1}{\tau}\left(1-\frac{1}{r}\right)V^*},$$

and combining the three possibilities gives

$$\begin{aligned} E(s^\tau)^2 &\approx 1^2 \cdot e^{-\frac{1}{\tau}\left(1+\frac{1}{r}\right)V^*} + \left(\rho_1\left(Y_1^\tau, Y_2^\tau\right)\right)^2 \cdot e^{-\frac{1}{\tau r}V^*} \\ &\approx e^{-\frac{1}{\tau}\left(1+\frac{1}{r}\right)V^*} + e^{-\frac{1}{\tau}\left(2-\frac{1}{r}\right)V^*} \\ &\approx e^{-\frac{1}{\tau}\left[\left(1+\frac{1}{r}\right)\wedge\left(2-\frac{1}{r}\right)\right]V^*}. \end{aligned}$$

Infinite swapping with IID samples

The definition of ρ_1 gives

$$\rho_1\left(Y_1^\tau, Y_2^\tau\right) \approx \frac{e^{-\frac{1}{\tau}V^*}}{e^{-\frac{1}{\tau}V^*} + e^{-\frac{1}{\tau r}V^*}} \approx e^{-\frac{1}{\tau}\left(1-\frac{1}{r}\right)V^*},$$

and combining the three possibilities gives

$$\begin{aligned} E(s^\tau)^2 &\approx 1^2 \cdot e^{-\frac{1}{\tau}\left(1+\frac{1}{r}\right)V^*} + \left(\rho_1\left(Y_1^\tau, Y_2^\tau\right)\right)^2 \cdot e^{-\frac{1}{\tau r}V^*} \\ &\approx e^{-\frac{1}{\tau}\left(1+\frac{1}{r}\right)V^*} + e^{-\frac{1}{\tau}\left(2-\frac{1}{r}\right)V^*} \\ &\approx e^{-\frac{1}{\tau}\left[\left(1+\frac{1}{r}\right)\wedge\left(2-\frac{1}{r}\right)\right]V^*}. \end{aligned}$$

Maximum decay rate at $r = 2$.

Small noise limit via Freidlin-Wentsell methods

Return to the diffusion model and two temperature INS:

$$dY_1 = -\nabla V(Y_1)dt + \sqrt{2\tau_1\rho_1(Y_1, Y_2) + 2\tau_2\rho_2(Y_1, Y_2)}dW_1$$
$$dY_2 = -\nabla V(Y_2)dt + \sqrt{2\tau_2\rho_1(Y_1, Y_2) + 2\tau_1\rho_2(Y_1, Y_2)}dW_2,$$

with

$$\rho_1(x_1, x_2) = \frac{e^{-\left[\frac{V(x_1)}{\tau_1} + \frac{V(x_2)}{\tau_2}\right]}}{Z_\rho(x_1, x_2)}, \quad \rho_2(x_1, x_2) = \frac{e^{-\left[\frac{V(x_2)}{\tau_1} + \frac{V(x_1)}{\tau_2}\right]}}{Z_\rho(x_1, x_2)},$$

stationary with respect to symmetrized distribution

$$\frac{1}{2Z(\tau_1)Z(\tau_2)}\left[e^{-\frac{V(x_1)}{\tau_1}}e^{-\frac{V(x_2)}{\tau_2}} + e^{-\frac{V(x_2)}{\tau_1}}e^{-\frac{V(x_1)}{\tau_2}}\right].$$

Small noise limit via Freidlin-Wentsell methods

Return to the diffusion model and two temperature INS:

$$dY_1 = -\nabla V(Y_1)dt + \sqrt{2\tau_1\rho_1(Y_1, Y_2) + 2\tau_2\rho_2(Y_1, Y_2)}dW_1$$
$$dY_2 = -\nabla V(Y_2)dt + \sqrt{2\tau_2\rho_1(Y_1, Y_2) + 2\tau_1\rho_2(Y_1, Y_2)}dW_2,$$

with

$$\rho_1(x_1, x_2) = \frac{e^{-\left[\frac{V(x_1)}{\tau_1} + \frac{V(x_2)}{\tau_2}\right]}}{Z_\rho(x_1, x_2)}, \quad \rho_2(x_1, x_2) = \frac{e^{-\left[\frac{V(x_2)}{\tau_1} + \frac{V(x_1)}{\tau_2}\right]}}{Z_\rho(x_1, x_2)},$$

stationary with respect to symmetrized distribution

$$\frac{1}{2Z(\tau_1)Z(\tau_2)}\left[e^{-\frac{V(x_1)}{\tau_1}}e^{-\frac{V(x_2)}{\tau_2}} + e^{-\frac{V(x_2)}{\tau_1}}e^{-\frac{V(x_1)}{\tau_2}}\right].$$

We will let $\tau_2 = \tau r \geq \tau = \tau_1$, identify the analogue of the decay rate of second moment, and optimize in the limit $\tau \to 0$. Will also present corresponding results for K temperatures.

Small noise limit via Freidlin-Wentsell methods

Problem of interest is again to estimate $\pi_\tau(A)$, but now using

$$s_T^\tau = \frac{1}{T}\int_0^T \left[\rho_1(Y_1^\tau(t), Y_2^\tau(t))1_{\{Y_1^\tau(t)\in A\}} + \rho_2(Y_1^\tau(t), Y_2^\tau(t))1_{\{Y_2^\tau(t)\in A\}}\right] dt.$$

Small noise limit via Freidlin-Wentsell methods

Problem of interest is again to estimate $\pi_\tau(A)$, but now using

$$s_T^\tau = \frac{1}{T}\int_0^T \left[\rho_1(Y_1^\tau(t), Y_2^\tau(t))1_{\{Y_1^\tau(t)\in A\}} + \rho_2(Y_1^\tau(t), Y_2^\tau(t))1_{\{Y_2^\tau(t)\in A\}}\right] dt.$$

Performance criteria is the *rate of growth of variance per unit time*:

$$T(\tau)\mathrm{Var}(s^\tau_{T(\tau)})^2 = \frac{1}{T(\tau)}\mathrm{Var}(T(\tau)s^\tau_{T(\tau)})^2,$$

where $T(\tau) = \exp M/\tau$.

Small noise limit via Freidlin-Wentsell methods

Problem of interest is again to estimate $\pi_\tau(A)$, but now using

$$s_T^\tau = \frac{1}{T}\int_0^T \left[\rho_1(Y_1^\tau(t), Y_2^\tau(t))1_{\{Y_1^\tau(t)\in A\}} + \rho_2(Y_1^\tau(t), Y_2^\tau(t))1_{\{Y_2^\tau(t)\in A\}}\right] dt.$$

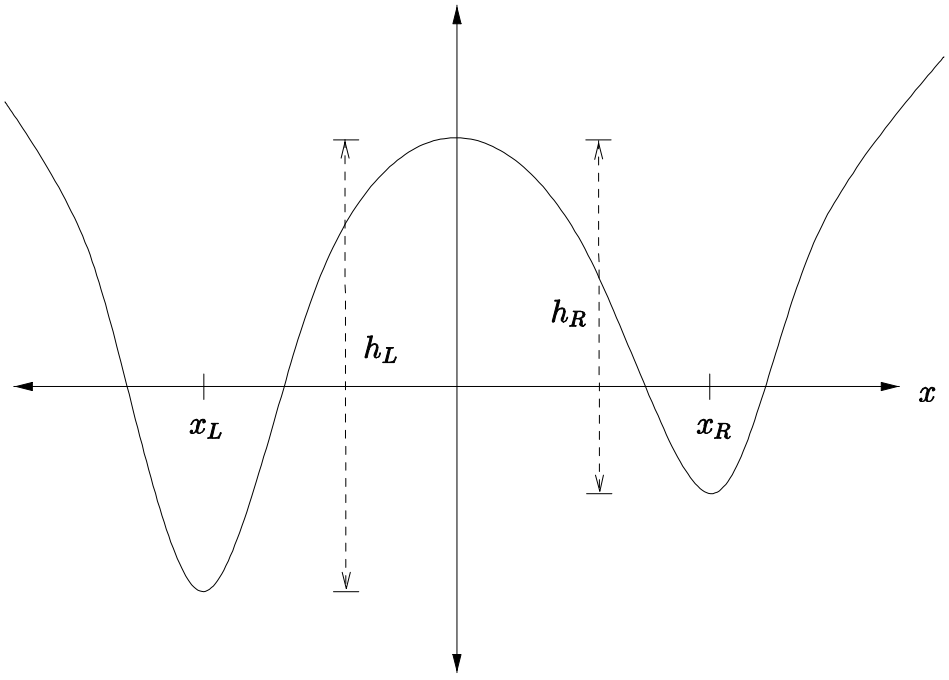
Performance criteria is the *rate of growth of variance per unit time*:

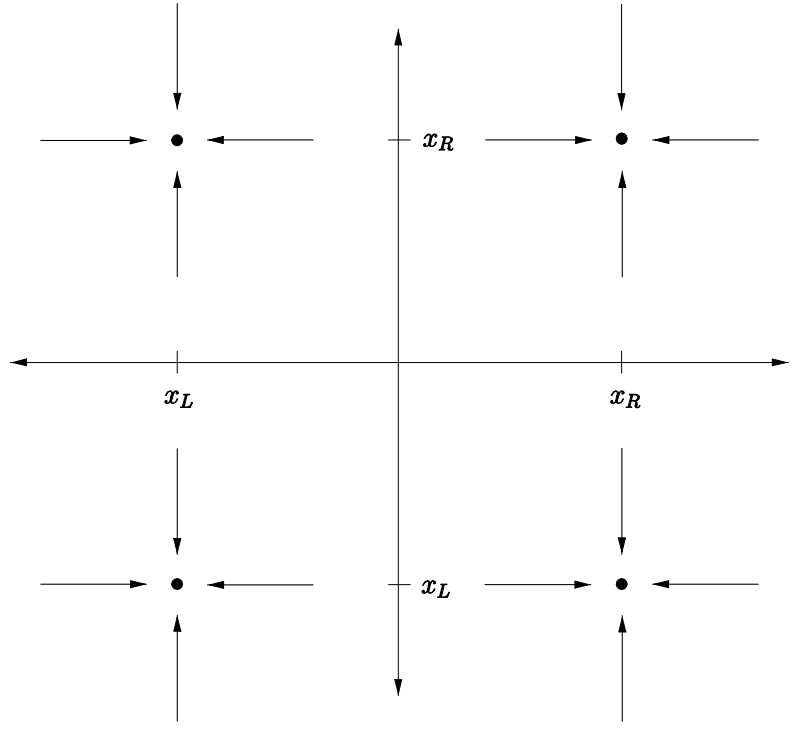
$$T(\tau)\mathrm{Var}(s_{T(\tau)}^\tau)^2 = \frac{1}{T(\tau)}\mathrm{Var}(T(\tau)s_{T(\tau)}^\tau)^2,$$

where $T(\tau) = \exp M/\tau$. The optimizer for temperature placement will be *independent* of M, but one should imagine M large enough that a regenerative structure can be used. This regenerative structure is the key to evaluating the limit $\tau \to 0$.

Small noise limit via Freidlin-Wentsell methods

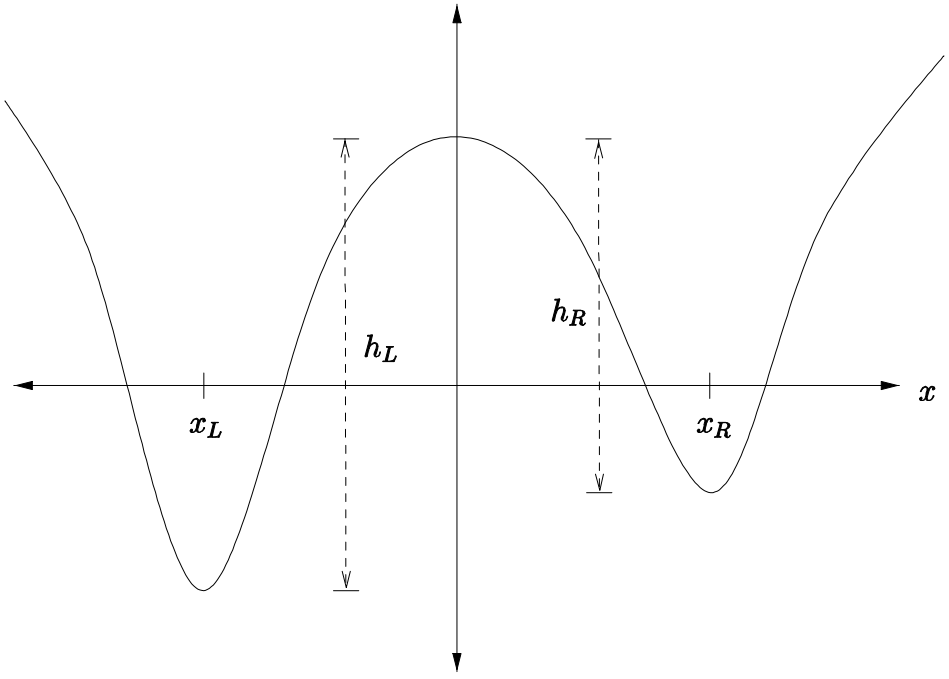
Consider for example a two well model for V and corresponding noiseless dynamics for (Y_1^τ, Y_2^τ):

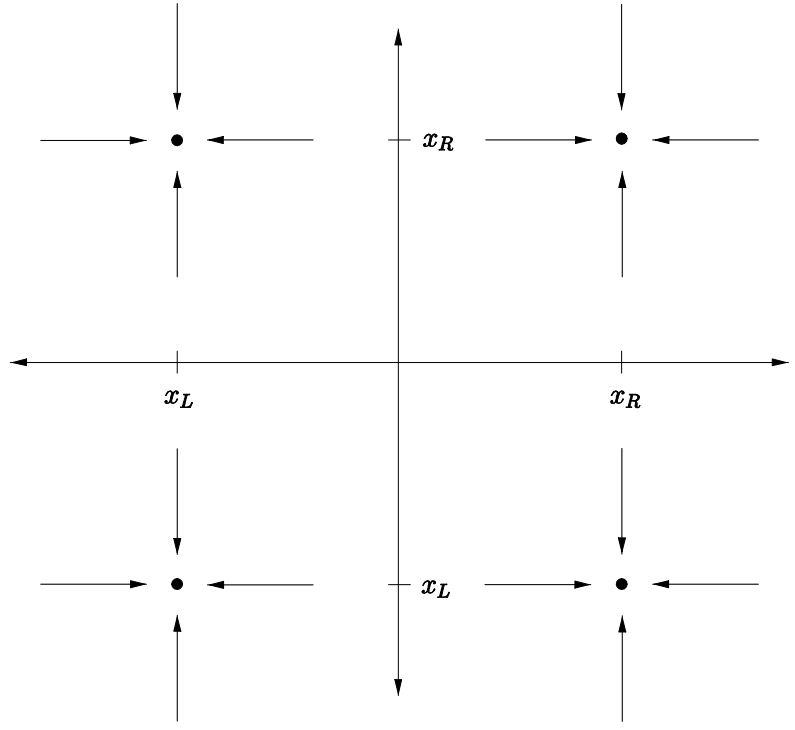




Small noise limit via Freidlin-Wentsell methods

Consider for example a two well model for V and corresponding noiseless dynamics for (Y_1^τ, Y_2^τ):





deepest well at (x_L, x_L), next deepest at (x_L, x_R) and (x_R, x_L), shallowest at (x_R, x_R).

Small noise limit via Freidlin-Wentsell methods

An extension of Freidlin-Wentsell theory used to justify the approximation of

$$\{(Y_1^\tau(t), Y_2^\tau(t)), t \in [0, T]\}$$

when $\tau > 0$ small by *finite state continuous time Markov chain*, with states

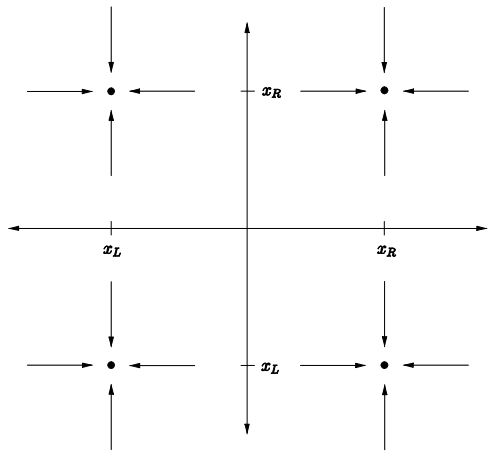
$$\{(x_L, x_L), (x_L, x_R), (x_R, x_L), (x_R, x_R)\},$$

transition rates determined by the *quasipotential*

$$Q(y_1, y_2) = \min\left\{V(y_1) + \frac{1}{r}V(y_2), V(y_2) + \frac{1}{r}V(y_1)\right\} - \left(1 + \frac{1}{r}\right)V(x_L).$$

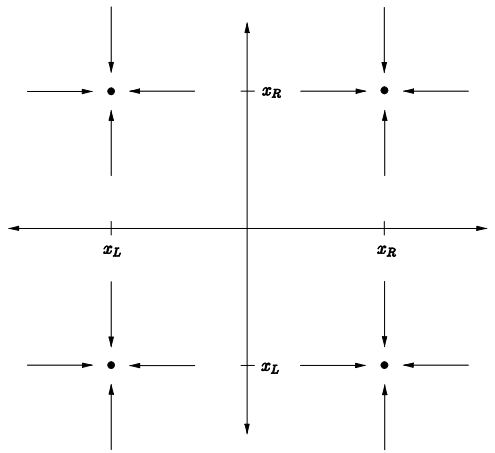
Small noise limit via Freidlin-Wentsell methods

$$Q(y_1, y_2) = \min\left\{ V(y_1) + \frac{1}{r} V(y_2), V(y_2) + \frac{1}{r} V(y_1) \right\} - \left(1 + \frac{1}{r}\right) V(x_L),$$



Small noise limit via Freidlin-Wentsell methods

$$Q(y_1, y_2) = \min\left\{ V(y_1) + \frac{1}{r} V(y_2), V(y_2) + \frac{1}{r} V(y_1) \right\} - \left(1 + \frac{1}{r}\right) V(x_L),$$



Not standard since non-smooth potential. Then compute

$$\lim_{\tau \to 0} -\tau \log \left[T(\tau) \mathrm{Var}(s^{\tau}_{T(\tau)})^2 \right]$$

using regenerative structure.

Optimal temperatures in the small noise limit

- $\pi_\tau(A)$ with A a subset of shallower well that includes mimimum.
- $\lim_{\tau \to 0} -\tau \log \pi_\tau(A) = (h_L - h_R)$.

Optimal temperatures in the small noise limit

- $\pi_\tau(A)$ with A a subset of shallower well that includes mimimum.
- $\lim_{\tau\to 0} -\tau \log \pi_\tau(A) = (h_L - h_R)$.

Theorem

Consider the two well, K temperature problem for estimating $\pi_\tau(A)$. Let $h_R = \beta h_L$ with $\beta \in (0,1]$. Then

$$\inf_{1\le r_2\le\cdots\le r_K\le\infty} \lim_{\tau\to 0} -\tau \log\left[T(\tau)\,Var(s^\tau_{T(\tau)})^2\right]$$
$$= \begin{cases} \left(2-\left(\frac{1}{2}\right)^{K-1}\right) h_L - 2h_R & \text{if } \beta \le 1/2 \\ \left(2-\left(\frac{1}{2}\right)^{K-2}\right)(h_L - h_R) & \text{if } \beta \ge 1/2 \end{cases}$$

with optimal r's

$$\left(1,2,\ldots,2^{K-2},2^{K-1}\right) \text{ and } \left(1,2,\ldots,2^{K-2},\infty\right).$$

Optimal temperatures in the small noise limit

Generalizations/comments:

- All cases of three well model have optimal temperatures in small noise limit as one of these two forms
- Conjecture that same is true for arbitrary finite number of wells
- Analogous results for functionals, discrete time models
- Geometric spacing has been suggested based on other arguments for PT/INS (see references)

Summary

- "Rare events" issues of various sorts are one of the banes of efficient Monte Carlo
- As such, it is natural to use various asymptotic theories to understand issues of algorithm design
- There is a relatively long history of the use of large deviation ideas in the design of algorithms for estimating probabilities of single rare events, but less on how to overcome impact of rare events on MCMC
- Have presented one such use in the context of parallel replica algorithms to understand and optimize the mechanisms that produce variance reduction

References

Parallel tempering:

- "Replica Monte Carlo simulation of spin glasses", Swendsen and Wang, *Phys. Rev. Lett.*, **57**, 2607–2609, 1986.
- "Markov chain Monte Carlo maximum likelihood", Geyer in *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*, ASA, 156–163, 1991.

A paper suggesting it is good to swap a lot:

- "Exchange frequency in replica exchange molecular dynamics", Sindhikara, Meng and Roitberg, *J. of Chem. Phy.*, **128**, 024104, 2008.

Use of pFIM:

- "Path-space information bounds for uncertainty quantification and sensitivity analysis of stochastic dynamics", (D, M.A. Katsoulakis, Y. Pantazis and P. Plecháč), *SIAM/ASA J. Uncertainty Quantification*, **4**, 80-111, 2016.

References

Infinite swapping as a limit of PT:

- "On the infinite swapping limit for parallel tempering", D, Liu, Plattner and Doll, *SIAM J. on MMS*, **10**, 986–1022, 2012.

Infinite swapping for IID samples:

- "Infinite swapping using IID Samples", D, Wu, Snarski, submitted to *TOMACS*.

Properties of INS and formulation for discrete state:

- "A large deviations analysis of certain qualitative properties of parallel tempering and infinite swapping algorithms", Doll, D and Nyquist, *Appl Math Optim,* doi:10.1007/s00245-017-9401-9, 2017.

References

Applications to biology:

- "Overcoming the rare-event sampling problem in biological systems with infinite swapping", Plattner, Doll and Meuwly, *J. of Chem. Th. and Comp.* **9**, 4215–4224, 2013.
- "Partial infinite swapping: Implementation and application to alanine-decapeptide and myoglobin in the gas phase and in solution", Hédin, Plattner, Doll and Meuwly, *J. Chem. Theory,* to appear.

Paper suggesting that geometric ratios of temperatures is good:

- "Infinite swapping replica exchange molecular dynamics leads to a simple simulation patch using mixture potentials", Lu and Vanden-Eijnden, *J Chem Phys.*, **138**, 084105, 2013.