Qualitative properties of parallel tempering and its infinite swapping limit

Paul Dupuis

Division of Applied Mathematics
Brown University

CERMICS workshop: **Computational statistics and molecular simulation**

with Jim Doll (Department of Chemistry, Brown University)

Yufei Liu, Pierre Nyquist, Nuria Plattner, David Lipshutz

2 February 2016

Outline

1. General perspective
2. Problem of interest
3. Large deviation theory for empirical measure
4. An accelerated algorithm–parallel tempering
5. Three uses of large deviation:
    - First large deviations analysis and the infinite swapping limit
    - Approximation of risk-sensitive functionals–a second rare event issue
    - The particle/temperature empirical measure–a diagnostic for convergence
6. Implementation issues and partial infinite swapping
7. References

General perspective

- When using Monte Carlo, there are several ways “rare events” can adversely affect performance or even viability of the method

General perspective

- When using Monte Carlo, there are several ways "rare events" can adversely affect performance or even viability of the method
  - Large relative variance of standard Monte Carlo when estimating small probabilities

General perspective

- When using Monte Carlo, there are several ways "rare events" can adversely affect performance or even viability of the method
  - Large relative variance of standard Monte Carlo when estimating small probabilities
  - Poor communication/ergodicity properties when using MCMC for stationary distributions

General perspective

- When using Monte Carlo, there are several ways "rare events" can adversely affect performance or even viability of the method
  - Large relative variance of standard Monte Carlo when estimating small probabilities
  - Poor communication/ergodicity properties when using MCMC for stationary distributions
- For purposes of design and qualitative understanding of the methods, need some way to characterize the impact of these events

General perspective

- When using Monte Carlo, there are several ways "rare events" can adversely affect performance or even viability of the method
  - Large relative variance of standard Monte Carlo when estimating small probabilities
  - Poor communication/ergodicity properties when using MCMC for stationary distributions
- For purposes of design and qualitative understanding of the methods, need some way to characterize the impact of these events
- Large deviation theory gives such information

General perspective

- When using Monte Carlo, there are several ways "rare events" can adversely affect performance or even viability of the method
  - Large relative variance of standard Monte Carlo when estimating small probabilities
  - Poor communication/ergodicity properties when using MCMC for stationary distributions
- For purposes of design and qualitative understanding of the methods, need some way to characterize the impact of these events
- Large deviation theory gives such information
  - Advantages: generally works directly with quantities of interest–not a surrogate

General perspective

- When using Monte Carlo, there are several ways "rare events" can adversely affect performance or even viability of the method
  - Large relative variance of standard Monte Carlo when estimating small probabilities
  - Poor communication/ergodicity properties when using MCMC for stationary distributions
- For purposes of design and qualitative understanding of the methods, need some way to characterize the impact of these events
- Large deviation theory gives such information
  - Advantages: generally works directly with quantities of interest–not a surrogate
  - Disadvantages: is an asymptotic theory (is it the right asymptotic for the problem at hand?); often requires solution to a variational problem

Problem of Interest

Compute the average potential energy, heat capacity, other functionals with respect to a Gibbs measure of the form

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

and V is the potential of a (relatively) complex physical system.

Problem of Interest

Compute the average potential energy, heat capacity, other functionals with respect to a Gibbs measure of the form

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

and V is the potential of a (relatively) complex physical system. Here primary interest is as the marginal (independent) distribution on spatial variables of stationary distribution of a Hamiltonian system. E.g.,

$$\bar{\pi}(dx, dp) \propto e^{-\frac{1}{\tau} V(x) - \frac{1}{\tau} \sum_{j=1}^{n} \frac{p_j^2}{2m}} dxdp.$$

Problem of Interest

Compute the average potential energy, heat capacity, other functionals with respect to a Gibbs measure of the form

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

and V is the potential of a (relatively) complex physical system. Here primary interest is as the marginal (independent) distribution on spatial variables of stationary distribution of a Hamiltonian system. E.g.,

$$\bar{\pi}(dx, dp) \propto e^{-\frac{1}{\tau}V(x) - \frac{1}{\tau}\sum_{j=1}^{n}\frac{p_j^2}{2m}}dxdp.$$

However many problems from other areas take this form, e.g., Bayesian inference, inverse problems, pattern theory, etc., where V depends on data. Here $\tau = k_B T$.

Problem of Interest

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 a variety of related discrete time models.

Problem of Interest

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 a variety of related discrete time models. The function $V(x)$ is defined on a large space, and includes, e.g., various inter-molecular potentials. In general, it may have a very complicated surface, with many deep and shallow local minima.

Problem of Interest

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 a variety of related discrete time models. The function $V(x)$ is defined on a large space, and includes, e.g., various inter-molecular potentials. In general, it may have a very complicated surface, with many deep and shallow local minima. Representative quantities of interest:

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

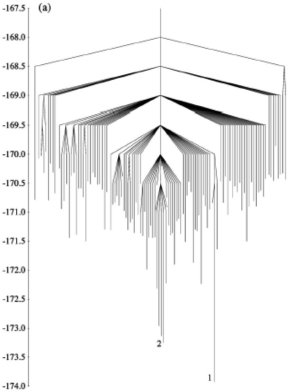
$$\text{heat capacity:} \quad \int \left[V(x) - \int V(y)\frac{e^{-V(y)/\tau}dy}{Z(\tau)}\right]^2 \frac{e^{-V(x)/\tau}dx}{Z(\tau)}.$$

Problem of Interest

An example of a potential energy surface is the Lennard-Jones cluster of 38 atoms. This potential has $\approx 10^{14}$ local minima.

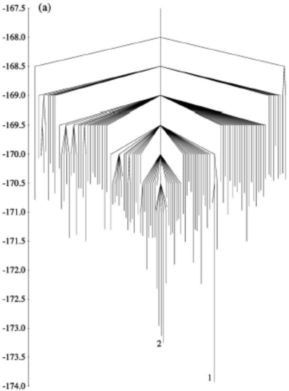
Problem of Interest

An example of a 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 the figure (taken from Doyle, Miller & Wales, *JCP*, 1999).



Problem of Interest

An example of a 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 the figure (taken from Doyle, Miller & Wales, *JCP*, 1999).



Large deviation theory for empirical measure

LD theory for empirical measure problem originates with Donsker-Varadhan and Gärtner.

Large deviation theory for empirical measure

LD theory for empirical measure problem originates with Donsker-Varadhan and Gärtner. Consider

$$dX = b(X)dt + \sigma(X)dW, \quad X(0) = x_0$$

and for large T

$$\mu^T(dx) = \frac{1}{T}\int_0^T \delta_{X(t)}(dx)dt.$$

Large deviation theory for empirical measure

LD theory for empirical measure problem originates with Donsker-Varadhan and Gärtner. Consider

$$dX = b(X)dt + \sigma(X)dW, \quad X(0) = x_0$$

and for large T

$$\mu^T(dx) = \frac{1}{T}\int_0^T \delta_{X(t)}(dx)dt.$$

Then considered as taking values in $\mathcal{P}(\mathbb{R}^d)$ and for small $\delta > 0$,

$$\mathbb{P}\left\{\mu^T \in N_\delta(\nu)\right\} \approx e^{-TJ_0(\nu)}.$$

Here $J_0(\nu) \geq 0$ measures deviations from the LLN limit (ergodic theorem) π, where π is unique invariant probability for X.

Large deviation theory for empirical measure

LD theory for empirical measure problem originates with Donsker-Varadhan and Gärtner. Consider

$$dX = b(X)dt + \sigma(X)dW, \quad X(0) = x_0$$

and for large T

$$\mu^T(dx) = \frac{1}{T}\int_0^T \delta_{X(t)}(dx)dt.$$

Then considered as taking values in $\mathcal{P}(\mathbb{R}^d)$ and for small $\delta > 0$,

$$\mathbb{P}\left\{\mu^T \in N_\delta(\nu)\right\} \approx e^{-TJ_0(\nu)}.$$

Here $J_0(\nu) \geq 0$ measures deviations from the LLN limit (ergodic theorem) π, where π is unique invariant probability for X. For diffusions satisfying a detailed balance, J_0 takes an explicit form.

Form of the rate.

Large deviation theory for empirical measure

Form of the rate. We have the variational representation*

$$-\frac{1}{T}\log \mathbb{P}\left\{\mu^T \in N_\delta(\nu)\right\} = \inf_u \mathbb{E}\left\{\frac{1}{T}\int_0^T u(t)^2 dt + \infty 1_{N_\delta(\nu)^c}(\bar{\mu}^T)\right\},$$

where u is progressively measurable with respect to W,

$$d\bar{X} = b(\bar{X})dt + \sigma(\bar{X})dW + \sigma(\bar{X})u, \quad \bar{X}(0) = x_0$$

and

$$\bar{\mu}^T(dx) = \frac{1}{T}\int_0^T \delta_{\bar{X}(t)}(dx)dt.$$

*Boué and Dupuis, *Annals of Probab.*, 1998.

Large deviation theory for empirical measure

Form of the rate.[†] Let $\mathcal{A}$ be the generator of X and

$$\|u\|_{L^2_\mu} = \left(\int_{\mathbb{R}^d} |u(x)|^2 \mu(dx) \right)^{1/2},$$

$$\mathcal{A}^u f = \mathcal{A} f + (\sigma u) \cdot \nabla f, \qquad f \in C_c^\infty,$$

$$\mathcal{S}(\mu) = \left\{ u \in L^2_\mu : \int_{\mathbb{R}^d} (\mathcal{A}^u f)(x) \mu(dx) = 0 \; \forall \; f \in C_c^\infty \right\}.$$

[†] proof from Dupuis and Lipshutz, 2016.

Large deviation theory for empirical measure

Form of the rate.[†] Let $\mathcal{A}$ be the generator of X and

$$\|u\|_{L^2_\mu} = \left(\int_{\mathbb{R}^d} |u(x)|^2 \mu(dx) \right)^{1/2},$$

$$\mathcal{A}^u f = \mathcal{A} f + (\sigma u) \cdot \nabla f, \qquad f \in C_c^\infty,$$

$$\mathcal{S}(\mu) = \left\{ u \in L^2_\mu : \int_{\mathbb{R}^d} (\mathcal{A}^u f)(x) \mu(dx) = 0 \; \forall \, f \in C_c^\infty \right\}.$$

Using the representation and a weak convergence analysis, one can show for $\theta(x) = \frac{d\nu}{d\pi}(x)$, $\theta^{1/2} \in W^{1,2}$, that

$$J_0(\mu) = \inf_{u \in \mathcal{S}(\mu)} \frac{1}{2} \|u\|^2_{L^2_\mu} \text{ if } \mathcal{S}(\mu) \neq \emptyset.$$

[†] proof from Dupuis and Lipshutz, 2016.

Large deviation theory for empirical measure

One can explicitly identify the minimum. IBP on $\int_{\mathbb{R}^d}(\mathcal{A}^u f)(x)\mu(dx) = 0$ suggests we consider $u = \nabla\varphi$, φ a weak sense soln to

$$\Delta\varphi + \nabla \log \theta \cdot \nabla\varphi = \frac{1}{\theta}\mathcal{A}^*\theta.$$

Large deviation theory for empirical measure

One can explicitly identify the minimum. IBP on $\int_{\mathbb{R}^d} (\mathcal{A}^u f)(x)\mu(dx) = 0$ suggests we consider $u = \nabla\varphi$, φ a weak sense soln to

$$\Delta\varphi + \nabla \log\theta \cdot \nabla\varphi = \frac{1}{\theta}\mathcal{A}^*\theta.$$

Then completion of squares shows this u is optimal.

Large deviation theory for empirical measure

One can explicitly identify the minimum. IBP on $\int_{\mathbb{R}^d}(\mathcal{A}^u f)(x)\mu(dx) = 0$ suggests we consider $u = \nabla\varphi$, φ a weak sense soln to

$$\Delta\varphi + \nabla \log\theta \cdot \nabla\varphi = \frac{1}{\theta}\mathcal{A}^*\theta.$$

Then completion of squares shows this u is optimal. If $b = -\nabla U$, $\sigma = I$, then this reduces to

$$\Delta\varphi + \nabla \log\theta \cdot \nabla\varphi = \Delta(\log\theta + U) + \nabla\log\theta \cdot \nabla(\log\theta + U).$$

In the case solution by inspection is $\varphi = \log\theta + U$. Thus

$$J_0(\mu) = \frac{1}{2}\|\nabla \log\theta\|^2_{L^2_\mu}.$$

An accelerated algorithm: parallel tempering

How to speed up a single particle.

A key idea: "parallel tempering" (also called "replica exchange", due to Geyer, Swendsen and Wang).

Idea of parallel tempering, two temperatures.

An accelerated algorithm: parallel tempering

How to speed up a single particle.

A key idea: "parallel tempering" (also called "replica exchange", due to Geyer, Swendsen and Wang).

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

An accelerated algorithm: parallel tempering

How to speed up a single particle.

A key idea: "parallel tempering" (also called "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

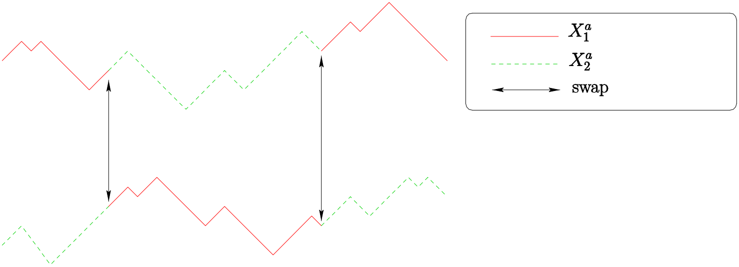
$$\pi(x_1, x_2) = e^{-\frac{V(x_1)}{\tau_1}} e^{-\frac{V(x_2)}{\tau_2}} \Big/ Z(\tau_1)Z(\tau_2).$$

An accelerated algorithm: 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."



An accelerated algorithm: parallel tempering

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

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

An accelerated algorithm: parallel tempering

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

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

$$\begin{array}{lcl} \text{Increased temperature} & \sim & \text{higher diffusivity of } X_2^a \\ & \sim & \text{easier communication for } X_2^a \\ & \sim & \text{passed to } X_1^a \text{ via swaps} \end{array}$$

This helps overcome the "rare event sampling problem." As we will see, there is a second "rare event problem" of a different sort that it also helps overcome.

Large deviations analysis

What does theory say about parallel tempering?

Large deviations analysis

What does theory say about parallel tempering? Suppose ν given such that

$$\theta(x_1, x_2) = \frac{d\nu}{d\pi}(x_1, x_2)$$

is smooth.

Large deviations analysis

What does theory say about parallel tempering? Suppose ν given such that

$$\theta(x_1, x_2) = \frac{d\nu}{d\pi}(x_1, x_2)$$

is smooth. Then we have *monotonic* form

$$I^a(\nu) = J_0(\nu) + aJ_1(\nu)$$

where J_0 is the rate for "no swap" dynamics and

Large deviations analysis

What does theory say about parallel tempering? Suppose ν given such that

$$\theta(x_1, x_2) = \frac{d\nu}{d\pi}(x_1, x_2)$$

is smooth. Then we have *monotonic* form

$$I^a(\nu) = J_0(\nu) + aJ_1(\nu)$$

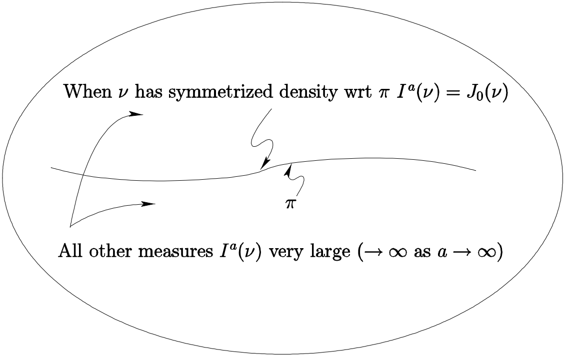
where J_0 is the rate for "no swap" dynamics and

$$J_1(\nu) = \int_{\mathbb{R}^d \times \mathbb{R}^d} g(x_1, x_2)\ell\left(\sqrt{\frac{\theta(x_2, x_1)}{\theta(x_1, x_2)}}\right)\nu(dx_1 dx_2)$$

with

$$\ell(z) = z \log z - z + 1 \begin{cases} = 0 & z = 1 \\ > 0 & z \neq 1 \end{cases} .$$

Large deviations analysis



Large deviations analysis

Rate for low temperature marginal. By contraction principle, for probability measure γ

$$I_1^a(\gamma) = \inf \left\{ I^a(\nu) : \text{ first marginal of } \nu \text{ is } \gamma \right\}.$$

Large deviations analysis

Rate for low temperature marginal. By contraction principle, for probability measure γ

$$I_1^a(\gamma) = \inf \left\{ I^a(\nu) : \text{ first marginal of } \nu \text{ is } \gamma \right\}.$$

If $\gamma(dx_1) \neq \pi_1(dx_1) = e^{-\frac{V(x_1)}{\tau_1}} dx_1 \Big/ Z(\tau_1)$, then for $a \in (0, \infty)$

$$I_1^a(\gamma) > I_1^0(\gamma)$$

and

$$I_1^a(\gamma) \uparrow \text{some finite limit.}$$

Exponentially faster decay for probability to be in any nice set that does not contain the target π_1.

Infinite swapping limit

This suggests one consider the infinite swapping limit $a \to \infty$, except

Infinite swapping limit

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,

Infinite swapping limit

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).

Infinite swapping limit

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).

Infinite swapping limit

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.

Infinite swapping limit

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:

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

Approximation to $\pi(dx)$:

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

Infinite swapping limit

Temperature swapping.

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 $ag(Y_1^a(t), Y_2^a(t))$.

Infinite swapping limit

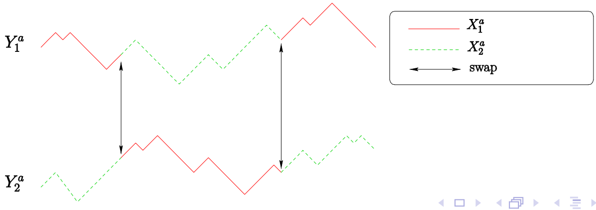
Temperature swapping. Process:

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

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 $\pi(dx)$:

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



Infinite swapping limit

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

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

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

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

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

Theorem: $\{\eta^T\}$ satisfies the large deviation principle with rate I^∞.

Infinite swapping limit

Remarks

- Prove a uniform result (can let $a \to a^* \in [0, \infty]$, $T \to \infty$ in any order).
- 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"

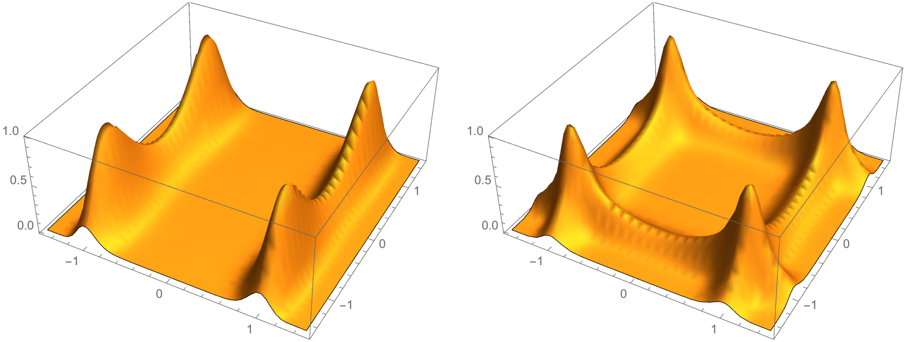
$$-\log\left[e^{-\frac{V(x_1)}{\tau_1}} e^{-\frac{V(x_2)}{\tau_2}} + e^{-\frac{V(x_2)}{\tau_1}} e^{-\frac{V(x_1)}{\tau_2}}\right]$$

has lower energy barriers than the original

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

Infinite swapping limit

Densities when $V(x)$ is a double well, orginal product density and density of implied potential:



Infinite swapping limit

Remarks

- To get the INS approximation $\eta^T(dx)$, we simulate (Y_1, Y_2), form its empirical measure, and push this through a deterministic "re-weighting" map M to get η^T:

$$M[\alpha](A) = \int_A [\rho_1(y_1, y_2)\alpha(dy_1 dy_2) + \rho_2(y_1, y_2)\alpha(dy_2 dy_1)] \,.$$

Infinite swapping limit

Remarks

- To get the INS approximation $\eta^T(dx)$, we simulate (Y_1, Y_2), form its empirical measure, and push this through a deterministic "re-weighting" map M to get η^T:

$$M[\alpha](A) = \int_A [\rho_1(y_1, y_2)\alpha(dy_1 dy_2) + \rho_2(y_1, y_2)\alpha(dy_2 dy_1)] .$$

- Two interpretations of the rate function in terms of rates J_0 of (X_1, X_2) and K of (Y_1, Y_2):

$$I^\infty(\nu) = \begin{cases} J_0(\nu) & \theta(x_1, x_2) = \theta(x_2, x_1) \\ \infty & \text{else} \end{cases} = \inf \{K(\alpha) : \nu = M[\alpha]\} .$$

The minimizing α is always symmetric, and enforces the "weighted" symmetry on ν due to the form of M.

A second rare event issue

- A problem that has attracted a lot of attention: how to select the temperatures.

A second rare event issue

- A problem that has attracted a lot of attention: how to select the temperatures.
- Suppose τ_1 fixed and assume interest is in this temperature. How to choose τ_2 (and $\tau_3, \ldots$)?

A second rare event issue

- A problem that has attracted a lot of attention: how to select the temperatures.
- Suppose τ_1 fixed and assume interest is in this temperature. How to choose τ_2 (and $\tau_3, \ldots$)?
- If only rare event issue sampling problem (long time correlation) tempted to choose τ_2 *large*.

A second rare event issue

- A problem that has attracted a lot of attention: how to select the temperatures.
- Suppose τ_1 fixed and assume interest is in this temperature. How to choose τ_2 (and $\tau_3, \ldots$)?
- If only rare event issue sampling problem (long time correlation) tempted to choose τ_2 *large*.
- One cannot do this for parallel tempering since too few acceptances if gap too large.

A second rare event issue

- A problem that has attracted a lot of attention: how to select the temperatures.
- Suppose τ_1 fixed and assume interest is in this temperature. How to choose τ_2 (and $\tau_3, \ldots$)?
- If only rare event issue sampling problem (long time correlation) tempted to choose τ_2 *large*.
- One cannot do this for parallel tempering since too few acceptances if gap too large.
- But infinite swapping "hard codes" the swaps, so why not?

A second rare event issue

- A problem that has attracted a lot of attention: how to select the temperatures.
- Suppose τ_1 fixed and assume interest is in this temperature. How to choose τ_2 (and $\tau_3, \ldots$)?
- If only rare event issue sampling problem (long time correlation) tempted to choose τ_2 *large*.
- One cannot do this for parallel tempering since too few acceptances if gap too large.
- But infinite swapping "hard codes" the swaps, so why not?
- One reason: for many interesting functionals (*risk-sensitive functionals*), there is *another rare event problem*.

A second rare event issue

What is an ordinary integral? One of the form

$$\frac{1}{Z(\tau)}\int F(x)e^{-V(x)/\tau}dx, \text{ e.g., } \frac{1}{Z(\tau)}\int V(x)e^{-V(x)/\tau}dx.$$

A second rare event issue

What is an ordinary integral? One of the form

$$\frac{1}{Z(\tau)}\int F(x)e^{-V(x)/\tau}dx, \text{ e.g., } \frac{1}{Z(\tau)}\int V(x)e^{-V(x)/\tau}dx.$$

What is a risk-sensitive functional? Ones of the form

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

and integrals heavily influenced by the tail of the distribution.

A second rare event issue

What is an ordinary integral? One of the form

$$\frac{1}{Z(\tau)}\int F(x)e^{-V(x)/\tau}dx, \text{ e.g., } \frac{1}{Z(\tau)}\int V(x)e^{-V(x)/\tau}dx.$$

What is a risk-sensitive functional? Ones of the form

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

and integrals heavily influenced by the tail of the distribution.

Examples: heat capacity, functionals arising in "free energy calculations"‡.

‡Lelièvre, Stoltz, Rousset, *Free Energy Computations: A Mathematical Perspective*, 2010.

A second rare event issue

What is an ordinary integral? One of the form

$$\frac{1}{Z(\tau)}\int F(x)e^{-V(x)/\tau}dx, \text{ e.g., } \frac{1}{Z(\tau)}\int V(x)e^{-V(x)/\tau}dx.$$

What is a risk-sensitive functional? Ones of the form

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

and integrals heavily influenced by the tail of the distribution.

Examples: heat capacity, functionals arising in "free energy calculations"[‡].

For ordinary functionals with INS, bigger τ_2 is true at least for some circumstances. *Not so* for risk-sensitive functionals.

[‡] Lelièvre, Stoltz, Rousset, *Free Energy Computations: A Mathematical Perspective*, 2010.

A second rare event issue

To illustrate, we eliminate the time correlation aspect, and assume iid samples (Y_1, Y_2) drawn from the target symmetrized distribution

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

A second rare event issue

To illustrate, we eliminate the time correlation aspect, and assume iid samples (Y_1, Y_2) drawn from the target symmetrized distribution

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

To simplify consider the risk-sensitive quantity

$$\frac{1}{Z(\tau_1)}\int_A e^{-V(x)/\tau_1}dx=\frac{1}{Z(\tau_1)}\int e^{-\infty 1_{A^c}(x)/\tau_1}e^{-V(x)/\tau_1}dx,$$

where A does not include the global min of V. Probability of interest decay rate:

$$-\tau\log\left(\frac{1}{Z(\tau)}\int_A e^{-V(x)/\tau}dx\right)\rightarrow\inf_{x\in A}V(x).$$

A second rare event issue

The estimate given by INS is

$$
\begin{aligned}
\theta^{\tau_1,\tau_2} &= 1_A(Y_1)\frac{e^{-\frac{V(Y_1)}{\tau_1}-\frac{V(Y_2)}{\tau_2}}}{e^{-\frac{V(Y_1)}{\tau_1}-\frac{V(Y_2)}{\tau_2}}+e^{-\frac{V(Y_2)}{\tau_1}-\frac{V(Y_1)}{\tau_2}}} \\
&\quad +1_A(Y_2)\frac{e^{-\frac{V(Y_2)}{\tau_1}-\frac{V(Y_1)}{\tau_2}}}{e^{-\frac{V(Y_1)}{\tau_1}-\frac{V(Y_2)}{\tau_2}}+e^{-\frac{V(Y_2)}{\tau_1}-\frac{V(Y_1)}{\tau_2}}}.
\end{aligned}
$$

Now let $\tau_2 = a\tau_1 = a\tau$, $a \in [1,\infty)$ and consider the limit $\tau \downarrow 0$.

A second rare event issue

The estimate given by INS is

$$\theta^{\tau_1,\tau_2} = 1_A(Y_1)\frac{e^{-\frac{V(Y_1)}{\tau_1}-\frac{V(Y_2)}{\tau_2}}}{e^{-\frac{V(Y_1)}{\tau_1}-\frac{V(Y_2)}{\tau_2}}+e^{-\frac{V(Y_2)}{\tau_1}-\frac{V(Y_1)}{\tau_2}}} + 1_A(Y_2)\frac{e^{-\frac{V(Y_2)}{\tau_1}-\frac{V(Y_1)}{\tau_2}}}{e^{-\frac{V(Y_1)}{\tau_1}-\frac{V(Y_2)}{\tau_2}}+e^{-\frac{V(Y_2)}{\tau_1}-\frac{V(Y_1)}{\tau_2}}}.$$

Now let $\tau_2 = a\tau_1 = a\tau$, $a \in [1,\infty)$ and consider the limit $\tau \downarrow 0$. Appropriate measure of performance the decay rate of variance/second moment. Using LD calculations,

$$-\tau \log\left(\mathbb{E}[\theta^{\tau_1,\tau_2}]^2\right) \to \min\left\{1+\frac{1}{a}, 2-\frac{1}{a}\right\}\left(\inf_{x\in A} V(x)\right).$$

Optimal decay rate is $\frac{3}{2}\inf_{x\in A} V(x)$, achieved when $a=2$.

A second rare event issue

Remarks

- Can extend to more temperatures
- Current project: put dynamics back in and consider double limit $T \to \infty$ and $\tau_1 \downarrow 0$

A diagnostic for convergence

For temperature swapped parallel tempering, let

$$\sigma_i(t) = \text{process component assigned dynamic with temperature } \tau_i,$$

so $\{\sigma_i(t), i = 1, \ldots, K\}$ is a permutation on $\{1, \ldots, K\}$, with $\sigma(0) = \iota$ (the identity permutation).

A diagnostic for convergence

For temperature swapped parallel tempering, let

$$\sigma_i(t) = \text{process component assigned dynamic with temperature } \tau_i,$$

so $\{\sigma_i(t), i = 1, \ldots, K\}$ is a permutation on $\{1, \ldots, K\}$, with $\sigma(0) = \iota$ (the identity permutation). Then for parallel tempering,

$$\gamma^T(\sigma) = \frac{1}{T}\int_0^T 1_{\{\sigma_i(t)=\sigma\}} dt$$

is the empirical measure on the particle/temperature associations, and is a (random) probability measure on $\Sigma_K = \{\text{permutations of } \{1, \ldots, K\}\}$. There is an analogue for INS.

A diagnostic for convergence

Lemma

For either PT or INS,

$$\gamma^T(\sigma) \to \frac{1}{K!} \text{ for } \sigma \in \Sigma_K, \text{ a.s.}$$

As a consequence, for any given particle

$$\text{fraction of time particle assigned dynamic } k \ \to \frac{1}{K}.$$

Easy to compute during simulation.

A diagnostic for convergence

For INS (notation of $K = 2$) the analogue of $\gamma^T(\sigma)$ is

$$\gamma^T(\{1,2\}) = \frac{1}{T}\int_0^T \rho_1(Y_1, Y_2)dt, \quad \gamma^T(\{2,1\}) = \frac{1}{T}\int_0^T \rho_2(Y_1, Y_2)dt,$$

defined in terms of the weights

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

A diagnostic for convergence

For INS (notation of $K = 2$) the analogue of $\gamma^T(\sigma)$ is

$$\gamma^T(\{1,2\}) = \frac{1}{T}\int_0^T \rho_1(Y_1, Y_2)dt, \quad \gamma^T(\{2,1\}) = \frac{1}{T}\int_0^T \rho_2(Y_1, Y_2)dt,$$

defined in terms of the weights

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

Rate function for the pair (η^T, γ^T):

$$I(\nu, (w_1, w_2)) = \inf\left\{K(\alpha) : \ \nu = M\alpha, \ \int \rho_1(x_1, x_2)\alpha(dx_1, dx_2) = w_1\right\},$$

where K is rate for empirical measure of (Y_1, Y_2),

$$M[\alpha](A) = \int_A \left[\rho_1(y_1, y_2)\alpha(dy_1 dy_2) + \rho_2(y_1, y_2)\alpha(dy_2 dy_1)\right].$$

A diagnostic for convergence

Theorem

If $(w_1, w_2) \neq (1/2, 1/2)$, *then the minimizer in*

$$\nu \to I(\nu, (w_1, w_2))$$

is **not** $\nu = \pi$.

A diagnostic for convergence

Theorem

If $(w_1, w_2) \neq (1/2, 1/2)$*, then the minimizer in*

$$\nu \to I(\nu, (w_1, w_2))$$

is **not** $\nu = \pi$*.*

Interpretation. By using the LDP (in the form of *Gibbs conditioning principle*), the minimizing ν is the overwhelmingly most likely location of η^T given $\gamma^T = (w_1, w_2)$. Thus if the particle/temperature association is not close to uniform, then η^T will not have converged.

A diagnostic for convergence

Theorem

If $(w_1, w_2) \neq (1/2, 1/2)$*, then the minimizer in*

$$\nu \to I(\nu, (w_1, w_2))$$

is **not** $\nu = \pi$.

Interpretation. By using the LDP (in the form of *Gibbs conditioning principle*), the minimizing ν is the overwhelmingly most likely location of η^T given $\gamma^T = (w_1, w_2)$. Thus if the particle/temperature association is not close to uniform, then η^T will not have converged.

Remark. Generalizes to K temperatures. The converse (unfortunately) does not hold. However, the analysis of the variationals problem $w \to I(\nu, (w_1, w_2))$ gives indicates other interesting aspects of INS.

Implementation issues and partial infinite swapping

- As noted applications of parallel tempering use many temperatures (e.g., $K = 30$ to 50) when V is complicated to overcome barriers of all different heights.

Implementation issues and partial infinite swapping

- As noted applications of parallel tempering use many temperatures (e.g., $K = 30$ to 50) when V is complicated to overcome barriers of all different heights.
- Straightforward extension of infinite swapping to K temperatures $\tau_1 < \tau_2 < \cdots < \tau_K$. Benefits of symmetrization/equilibration even greater, larger rate for lowest marginal.

Implementation issues and partial infinite swapping

- As noted applications of parallel tempering use many temperatures (e.g., $K = 30$ to 50) when V is complicated to overcome barriers of all different heights.
- Straightforward extension of infinite swapping to K temperatures $\tau_1 < \tau_2 < \cdots < \tau_K$. Benefits of symmetrization/equilibration even greater, larger rate for lowest marginal.
- But, coefficients become complex, e.g., $K!$ weights, and each involves many calculations. Not practical if $K \geq 7$.

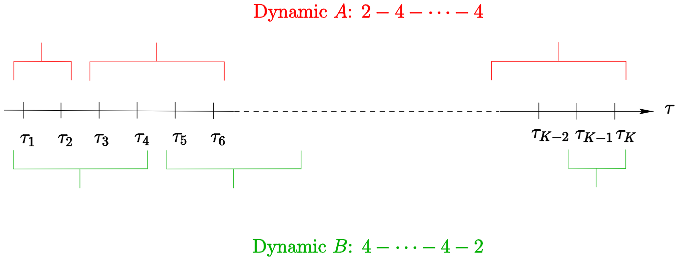
Implementation issues and partial infinite swapping

- As noted applications of parallel tempering use many temperatures (e.g., $K = 30$ to 50) when V is complicated to overcome barriers of all different heights.
- Straightforward extension of infinite swapping to K temperatures $\tau_1 < \tau_2 < \cdots < \tau_K$. Benefits of symmetrization/equilibration even greater, larger rate for lowest marginal.
- But, coefficients become complex, e.g., $K!$ weights, and each involves many calculations. Not practical if $K \geq 7$.
- Need for computational feasibility leads to *partial infinite swapping*.

Implementation issues and partial infinite swapping

Partial infinite swapping. Given any subgroup of set of permutations one can construct a corresponding *partial infinite swapping* dynamic.

Partial infinite swapping. Given any subgroup of set of permutations one can construct a corresponding *partial infinite swapping* dynamic. Two examples are Dynamics A and B in figure:



Implementation issues and partial infinite swapping

- Using partial infinite swapping one can control the complexity of the coefficients and algorithm.

Implementation issues and partial infinite swapping

- Using partial infinite swapping one can control the complexity of the coefficients and algorithm.
- If one alternates between subgroups that generate full group of permutations, one approximates full infinite swapping (convergence theorem in continuous time).

Implementation issues and partial infinite swapping

- Using partial infinite swapping one can control the complexity of the coefficients and algorithm.
- If one alternates between subgroups that generate full group of permutations, one approximates full infinite swapping (convergence theorem in continuous time).
- However, particles lose their temperature identity in infinite swapping limit (partial or otherwise). Cannot simply alternate–need a proper "handoff" rule.

Implementation issues and partial infinite swapping

- Using partial infinite swapping one can control the complexity of the coefficients and algorithm.
- If one alternates between subgroups that generate full group of permutations, one approximates full infinite swapping (convergence theorem in continuous time).
- However, particles lose their temperature identity in infinite swapping limit (partial or otherwise). Cannot simply alternate–need a proper "handoff" rule.
- Can identify the "distributionally correct" handoff rule, using that partial swappings are limits of "physically meaningful" processes. E.g., in a block of 4 locations x_i associated with 4 temperatures τ_i, select a permutation σ according to

$$e^{-\left[\frac{V(x_{\sigma(1)})}{\tau_1}+\frac{V(x_{\sigma(2)})}{\tau_2}+\frac{V(x_{\sigma(3)})}{\tau_3}+\frac{V(x_{\sigma(4)})}{\tau_4}\right]} \Bigg/ \sum_{\bar{\sigma}} e^{-\left[\frac{V(x_{\bar{\sigma}(1)})}{\tau_1}+\frac{V(x_{\bar{\sigma}(2)})}{\tau_2}+\frac{V(x_{\bar{\sigma}(3)})}{\tau_3}+\frac{V(x_{\bar{\sigma}(4)})}{\tau_4}\right]},$$

and assign τ_i to $x_{\sigma(i)}$.

Implementation issues and partial infinite swapping

- Using partial infinite swapping one can control the complexity of the coefficients and algorithm.
- If one alternates between subgroups that generate full group of permutations, one approximates full infinite swapping (convergence theorem in continuous time).
- However, particles lose their temperature identity in infinite swapping limit (partial or otherwise). Cannot simply alternate–need a proper "handoff" rule.
- Can identify the "distributionally correct" handoff rule, using that partial swappings are limits of "physically meaningful" processes. E.g., in a block of 4 locations x_i associated with 4 temperatures τ_i, select a permutation σ according to

$$e^{-\left[\frac{V(x_{\sigma(1)})}{\tau_1}+\frac{V(x_{\sigma(2)})}{\tau_2}+\frac{V(x_{\sigma(3)})}{\tau_3}+\frac{V(x_{\sigma(4)})}{\tau_4}\right]} \Bigg/ \sum_{\bar{\sigma}} e^{-\left[\frac{V(x_{\bar{\sigma}(1)})}{\tau_1}+\frac{V(x_{\bar{\sigma}(2)})}{\tau_2}+\frac{V(x_{\bar{\sigma}(3)})}{\tau_3}+\frac{V(x_{\bar{\sigma}(4)})}{\tau_4}\right]},$$

  and assign τ_i to $x_{\sigma(i)}$.
- CHARMM codes (http://www.charmm.org) due to Plattner, Meuwly

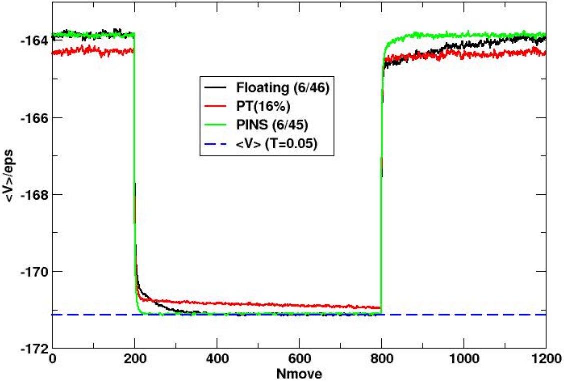
Numerical examples

Relaxation study of convergence to equilibrium for LJ-38.

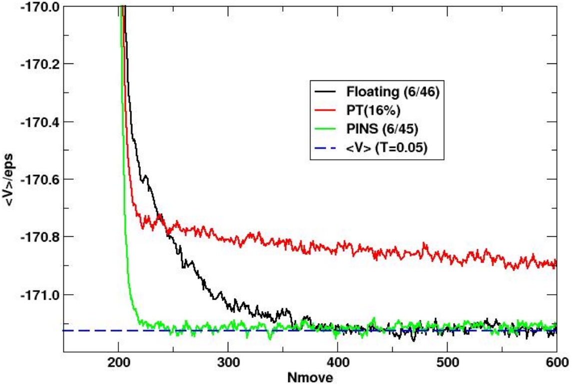
- quantity of interest: average potential energy at various temperatures
- used 45 temperatures, 3–6–6–$\cdots$–6 type dynamic for partial infinite swapping
- used Smart Monte Carlo for particle dynamics
- lowest $1/3$ of temperatures raised to push process away from equilibrium (low temperature components pushed away from deep minima)
- then reduced to correct temperatures for 600 discrete time steps to study return to equilibria
- repeated 2000 times, we plot averages for lowest (and hardest) temperature

Numerical examples

Relaxation study of convergence to equilibrium for LJ-38: parallel tempering versus partial infinite swapping, only lowest temperature illustrated.



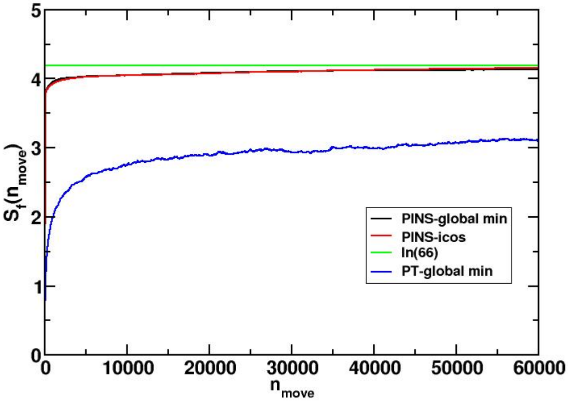
Numerical examples



For this system, reduction relative to parallel tempering: 10^{10} reduced to 10^6 steps with additional overhead of approximately 10%.

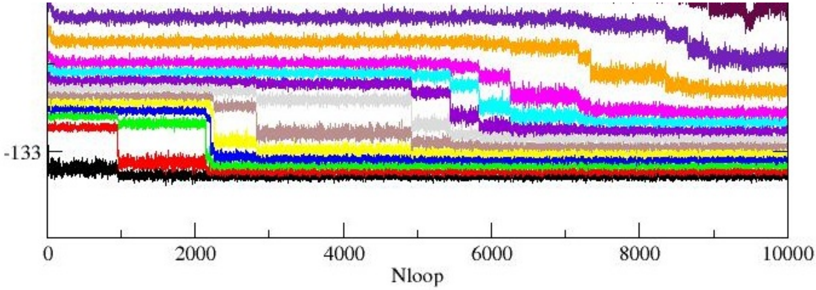
Numerical examples

Convergence of the empirical measure on temperatures to uniform distribution.



Numerical examples

Convergence to equilibrium, single sample, 12 lowest temperatures:



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.

LD for empirical measure and control form of rate:

- "On the large deviation rate for the empirical measure of a reversible pure jump Markov processes", Dupuis and Liu, *Annals of Probability*, **43**, (2015), 1121–1156.
- "Large deviations for the empirical measure of a diffusion via weak convergence methods", Dupuis and Lipshutz, preprint.

References

Infinite swapping:

- "On the infinite swapping limit for parallel tempering", Dupuis, Liu, Plattner and Doll, *SIAM J. on MMS*, **10**, 986–1022, 2012.
- "An infinite swapping approach to the rare-event sampling problem", Plattner, Doll, Dupuis, Wang, Liu and Gubernatis, *J. of Chem. Phy.*, **135**, 134111, 2011.
- "On performance measures for infinite swapping Monte Carlo methods", Doll and Dupuis, *J. of Chemical Physics*, **142**, (2015), 024111.
- "A large deviations analysis of certain qualitative properties of parallel tempering and infinite swapping algorithms", Doll, Dupuis and Nyquist, preprint.

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.
- "The structure and dynamics of alanine dipeptide and decapeptide at multiple thermodynamic states using the PINS method", Hédin, Plattner, Doll and Meuwly, preprint.

Paper using LD rate to study non-reversible single particle scheme:

- "Irreversible Langevin samplers and variance reduction: a large deviation approach", Rey-Bellet and Spiliopoulos, *Nonlinearity*, **28**, 2081–2103, 2015.

Paper considering alternative implementation of INS and PINS:

- "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.