

Accelerating Monte Carlo—What does the Donsker-Varadhan theory have to say?

Paul Dupuis

Division of Applied Mathematics
Brown University

with J.D. Doll, Yufei Liu, Nuria Plattner and Hui Wang

INFORMS APC

July, 2011

Large deviations and Monte Carlo

Two of the main topics in large deviations are the *Freidlin-Wentzell theory* (LDP for small random perturbations of ODEs), and the *Donsker-Varadhan theory* (LDP for empirical measure of Markov processes).

Large deviations and Monte Carlo

Two of the main topics in large deviations are the *Freidlin-Wentzell theory* (LDP for small random perturbations of ODEs), and the *Donsker-Varadhan theory* (LDP for empirical measure of Markov processes).

Example of Freidlin-Wentzell: fix $T < \infty$ and consider for small $\varepsilon > 0$

$$dX^\varepsilon = b(X^\varepsilon)dt + \varepsilon^{1/2}\sigma(X^\varepsilon)dW, \quad X^\varepsilon(0) = x_0.$$

Then considered as taking values in $\mathcal{C}([0, T] : \mathbb{R}^d)$,

$$\mathbb{P}\{X^\varepsilon \approx \phi\} \approx e^{-\frac{1}{\varepsilon}I(\phi)}.$$

Large deviations and Monte Carlo

Two of the main topics in large deviations are the *Freidlin-Wentzell theory* (LDP for small random perturbations of ODEs), and the *Donsker-Varadhan theory* (LDP for empirical measure of Markov processes).

Example of Freidlin-Wentzell: fix $T < \infty$ and consider for small $\varepsilon > 0$

$$dX^\varepsilon = b(X^\varepsilon)dt + \varepsilon^{1/2}\sigma(X^\varepsilon)dW, \quad X^\varepsilon(0) = x_0.$$

Then considered as taking values in $\mathcal{C}([0, T] : \mathbb{R}^d)$,

$$\mathbb{P}\{X^\varepsilon \approx \phi\} \approx e^{-\frac{1}{\varepsilon}I(\phi)}.$$

$I(\phi) \geq 0$ measures deviations from the LLN limit $\dot{x} = b(x), x(0) = x_0$.

Large deviations and Monte Carlo

A long history of using F-W theory to design accelerated Monte Carlo to estimate probabilities of single rare events (ruin probabilities, buffer overflow), e.g.,

$$p^\varepsilon = \mathbb{P} \{X^\varepsilon(t) \text{ hits a target set } \mathcal{A} \text{ some } t \in [0, T] | X^\varepsilon(0) = x_0\}.$$

Large deviations and Monte Carlo

A long history of using F-W theory to design accelerated Monte Carlo to estimate probabilities of single rare events (ruin probabilities, buffer overflow), e.g.,

$$p^\varepsilon = \mathbb{P} \{X^\varepsilon(t) \text{ hits a target set } \mathcal{A} \text{ some } t \in [0, T] | X^\varepsilon(0) = x_0\}.$$

Starts with

- Siegmund, D., *Importance sampling in the Monte Carlo study of sequential tests*. Ann. Statist., **4**, 673–684, 1976.

and many dozens of papers since.

Large deviations and Monte Carlo

Example of Donsker-Varadhan (see also Gärtner): consider

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

and for large T the *empirical* or *occupation* measure

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

Large deviations and Monte Carlo

Example of Donsker-Varadhan (see also Gärtner): consider

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

and for large T the *empirical* or *occupation* measure

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

Then considered as taking values in $\mathcal{P}(\mathbb{R}^d)$,

$$\mathbb{P} \left\{ \mu^T \approx \nu \right\} \approx e^{-TI(\nu)}.$$

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

Large deviations and Monte Carlo

A standard method to approximate

$$\int_{\mathbb{R}^d} f(x)\pi(dx)$$

is via some variant on the Monte Carlo scheme

$$\int_{\mathbb{R}^d} f(x)\mu^T(dx) = \frac{1}{T} \int_0^T f(X(t))dt.$$

Large deviations and Monte Carlo

A standard method to approximate

$$\int_{\mathbb{R}^d} f(x)\pi(dx)$$

is via some variant on the Monte Carlo scheme

$$\int_{\mathbb{R}^d} f(x)\mu^T(dx) = \frac{1}{T} \int_0^T f(X(t))dt.$$

When parts of \mathbb{R}^d communicate poorly under dynamics of X , the approximation can be extremely slow to converge (the rare event problem). Hence interest in accelerated Monte Carlo.

Large deviations and Monte Carlo

A standard method to approximate

$$\int_{\mathbb{R}^d} f(x)\pi(dx)$$

is via some variant on the Monte Carlo scheme

$$\int_{\mathbb{R}^d} f(x)\mu^T(dx) = \frac{1}{T} \int_0^T f(X(t))dt.$$

When parts of \mathbb{R}^d communicate poorly under dynamics of X , the approximation can be extremely slow to converge (the rare event problem). Hence interest in accelerated Monte Carlo.

Question: Can the D-V theory be used as a tool for design and understanding of these algorithms?

Outline

1 Examples

Outline

- 1 Examples
- 2 Standard measures of performance and their shortcomings

Outline

- 1 Examples
- 2 Standard measures of performance and their shortcomings
- 3 An accelerated algorithm: parallel tempering (aka replica exchange)

Outline

- 1 Examples
- 2 Standard measures of performance and their shortcomings
- 3 An accelerated algorithm: parallel tempering (aka replica exchange)
- 4 Large deviation properties and the infinite swapping limit

Outline

- 1 Examples
- 2 Standard measures of performance and their shortcomings
- 3 An accelerated algorithm: parallel tempering (aka replica exchange)
- 4 Large deviation properties and the infinite swapping limit
- 5 Implementation issues and partial infinite swapping

Outline

- 1 Examples
- 2 Standard measures of performance and their shortcomings
- 3 An accelerated algorithm: parallel tempering (aka replica exchange)
- 4 Large deviation properties and the infinite swapping limit
- 5 Implementation issues and partial infinite swapping
- 6 Numerical example

Outline

- 1 Examples
- 2 Standard measures of performance and their shortcomings
- 3 An accelerated algorithm: parallel tempering (aka replica exchange)
- 4 Large deviation properties and the infinite swapping limit
- 5 Implementation issues and partial infinite swapping
- 6 Numerical example
- 7 Concluding remarks

Examples

A representative example. Compute the average potential energy and other functionals with respect to a Gibbs measure of the form

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

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

Examples

A representative example. Compute the average potential energy and other functionals with respect to a Gibbs measure of the form

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

and V is the potential of a (relatively) complex physical system. A corresponding continuous time model is

$$dX = -\nabla V(X)dt + \sqrt{2\tau}dW, \quad X(0) = x_0,$$

where τ is a fixed temperature.

Examples

A representative example. Compute the average potential energy and other functionals with respect to a Gibbs measure of the form

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

and V is the potential of a (relatively) complex physical system. A corresponding continuous time model is

$$dX = -\nabla V(X)dt + \sqrt{2\tau}dW, \quad X(0) = x_0,$$

where τ is a fixed temperature.

- Simulations are done using a discrete time model.

Examples

At low temperatures the main contribution comes from the global minimum and “important” local minima, which often do not communicate well.

Examples

At low temperatures the main contribution comes from the global minimum and “important” local minima, which often do not communicate well.

To extract useful information requires very large scale computing, e.g., 10^{10} times steps for a more complex model than X itself.

Examples

At low temperatures the main contribution comes from the global minimum and “important” local minima, which often do not communicate well.

To extract useful information requires very large scale computing, e.g., 10^{10} times steps for a more complex model than X itself.

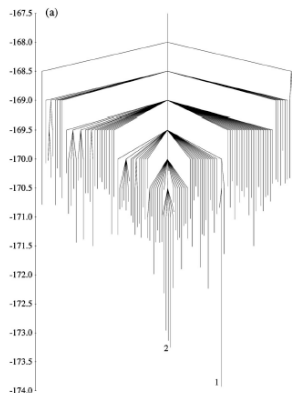
Dimensions on the order of 10,000, but can be much larger.

Examples

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

Examples

An example of such 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).



Examples

Problems with the same “rare event” issue:

- Bayesian statistics
- Pattern theory (image analysis and related estimation)
- Counting problems in computer science
- Computational physics/chemistry (condensed matter, quantum systems)
- Computational biology (motif sampling)
- ...

Standard measures of performance

How should one describe the rate of convergence

$$\frac{1}{T} \int_0^T f(X(t)) dt \rightarrow \int_{\mathbb{R}^d} f(x) \pi(dx)?$$

None of the standard descriptions work directly with convergence of the empirical measure.

Standard measures of performance

How should one describe the rate of convergence

$$\frac{1}{T} \int_0^T f(X(t)) dt \rightarrow \int_{\mathbb{R}^d} f(x) \pi(dx)?$$

None of the standard descriptions work directly with convergence of the empirical measure.

2nd eigenvalue. Consider the transition kernel

$$p(dx, T, x_0) = \mathbb{P} \{X(T) \in dx | X(0) = x_0\}.$$

Standard measures of performance

How should one describe the rate of convergence

$$\frac{1}{T} \int_0^T f(X(t)) dt \rightarrow \int_{\mathbb{R}^d} f(x) \pi(dx)?$$

None of the standard descriptions work directly with convergence of the empirical measure.

2nd eigenvalue. Consider the transition kernel

$$p(dx, T, x_0) = \mathbb{P}\{X(T) \in dx | X(0) = x_0\}.$$

Under mild conditions the exponential rate of convergence

$$p(dx, T, x_0) \rightarrow \pi(dx)$$

is determined by the sub-dominant eigenvalue of the operator corresponding to X . Used to characterize “efficiency” of the corresponding Monte Carlo.

Standard measures of performance

Problem: Only indirectly related to problem of interest. Information on density, but not on empirical measure which depends on sample path; neglects potentially significant effect of *time averaging* in empirical measure (Rosenthal, Gubernatis).

Standard measures of performance

Problem: Only indirectly related to problem of interest. Information on density, but not on empirical measure which depends on sample path; neglects potentially significant effect of *time averaging* in empirical measure (Rosenthal, Gubernatis).

Asymptotic variance. Also a popular quantity for comparing efficiency of algorithms, but is a property of the algorithm once one is already at equilibrium. Also does not properly reflect the time averaging.

Standard measures of performance

Problem: Only indirectly related to problem of interest. Information on density, but not on empirical measure which depends on sample path; neglects potentially significant effect of *time averaging* in empirical measure (Rosenthal, Gubernatis).

Asymptotic variance. Also a popular quantity for comparing efficiency of algorithms, but is a property of the algorithm once one is already at equilibrium. Also does not properly reflect the time averaging.

Large deviation rate. We will use the LD rate I , where a larger rate implies faster convergence.

An accelerated algorithm: parallel tempering

Idea of parallel tempering, two temperatures.

An accelerated algorithm: parallel tempering

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

An accelerated algorithm: parallel tempering

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

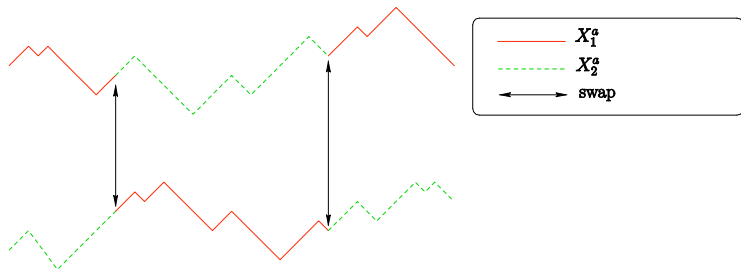
$$\pi(x_1, x_2) = e^{-\frac{V(x_1)}{\tau_1}} e^{-\frac{V(x_2)}{\tau_2}} / Z(\tau_1, \tau_2).$$

An accelerated algorithm: parallel tempering

Now introduce *swaps* (Swendsen, Geyer), 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

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

An accelerated algorithm: parallel tempering

Now have a *Markov jump-diffusion*. Easy to check: with this swapping intensity still have

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

- Increased temperature \sim higher diffusivity of X_2^a
- \sim easier communication for X_2^a
- \sim passed to X_1^a via swaps

An accelerated algorithm: parallel tempering

Now have a *Markov jump-diffusion*. Easy to check: with this swapping intensity still have

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

- Increased temperature \sim higher diffusivity of X_2^a
- \sim easier communication for X_2^a
- \sim passed to X_1^a via swaps

Natural question: how does convergence depend on a and τ_2 ? We focus on a (may consider τ_2 in future).

An accelerated algorithm: parallel tempering

Now have a *Markov jump-diffusion*. Easy to check: with this swapping intensity still have

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

- Increased temperature \sim higher diffusivity of X_2^a
- \sim easier communication for X_2^a
- \sim passed to X_1^a via swaps

Natural question: how does convergence depend on a and τ_2 ? We focus on a (may consider τ_2 in future).

Implementation uses discrete time version, and conventional wisdom is that swap rate should be chosen so that

one swap attempt \longleftrightarrow six steps discrete dynamics

Large deviation properties and the infinite swapping limit

What does the Donsker-Varadhan theory say?

Large deviation properties and the infinite swapping limit

What does the Donsker-Varadhan theory say? Suppose ν given such that

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

is smooth.

Large deviation properties and the infinite swapping limit

What does the Donsker-Varadhan theory say? 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 deviation properties and the infinite swapping limit

What does the Donsker-Varadhan theory say? 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

$$\ell(z) = z \log z - z + 1 = 0 \text{ iff } z = 1.$$

Large deviation properties and the infinite swapping limit

Thus for large a $I^a(\nu)$ is large (ν is very unlikely) unless

$$\int_{\mathbb{R}^d \times \mathbb{R}^d} g(x_1, x_2) \ell \left(\sqrt{\frac{\theta(x_2, x_1)}{\theta(x_1, x_2)}} \right) \nu(dx_1 dx_2) = 0.$$

Large deviation properties and the infinite swapping limit

Thus for large a $I^a(\nu)$ is large (ν is very unlikely) unless

$$\int_{\mathbb{R}^d \times \mathbb{R}^d} g(x_1, x_2) \ell \left(\sqrt{\frac{\theta(x_2, x_1)}{\theta(x_1, x_2)}} \right) \nu(dx_1 dx_2) = 0.$$

That means

$$\theta(x_2, x_1) = \theta(x_1, x_2) \text{ } \nu\text{-a.s.}$$

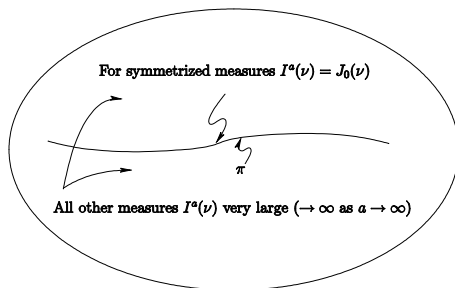
which is equivalent to

ν places precisely same relative weight on permutations

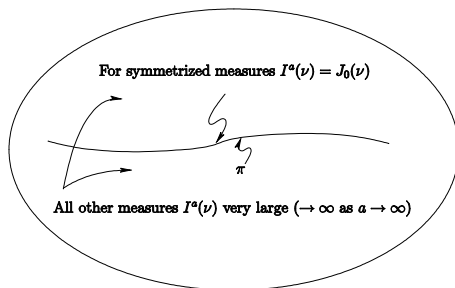
$$(x_2, x_1) \text{ and } (x_1, x_2) \text{ as } \pi, \text{ all } (x_1, x_2) \in \mathbb{R}^d \times \mathbb{R}^d.$$

Call these *symmetrized* measures.

Large deviation properties and the infinite swapping limit



Large deviation properties and the infinite swapping limit



Benefit even greater with more temperatures—all permutations of (x_1, x_2, \dots, x_K) perfectly balanced according to target distribution $\pi(dx_1, \dots, dx_K)$.

Large deviation properties and the infinite swapping limit

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

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

Large deviation properties and the infinite swapping limit

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

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

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

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

and

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

Large deviation properties and the infinite swapping limit

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

Large deviation properties and the infinite swapping limit

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

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

Large deviation properties and the infinite swapping limit

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

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

Large deviation properties and the infinite swapping limit

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

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

Large deviation properties and the infinite swapping limit

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

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

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

Large deviation properties and the infinite swapping limit

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

- if a is large but finite almost all computational effort is directed at swap attempts, rather than diffusion dynamics,
- if $a \rightarrow \infty$ 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 $\pi(dx)$:

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

Large deviation properties and the infinite swapping limit

Temperature swapping.

Large deviation properties and 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 $ag(Y_1^a(t), Y_2^a(t))$.

Large deviation properties and the infinite swapping limit

Temperature swapping. Process:

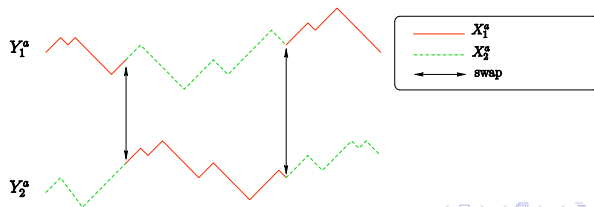
$$dY_1^a = -\nabla V(Y_1^a)dt + \sqrt{2r_1(Z^a)}dW_1$$

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

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 $\pi(dx)$:

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



Large deviation properties and the infinite swapping limit

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

Large deviation properties and the infinite swapping limit

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

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

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

and

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

Large deviation properties and the infinite swapping limit

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

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

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

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^∞ .

Implementation issues and partial infinite swapping

- 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

- Applications of parallel tempering use many temperatures (e.g., $K = 30$ to 50) when V is complicated to overcome barriers of all different heights.
- Swaps typically between near-neighbor only, with randomized or deterministic schedule of which pair to try (discrete time).

Implementation issues and partial infinite swapping

- Applications of parallel tempering use many temperatures (e.g., $K = 30$ to 50) when V is complicated to overcome barriers of all different heights.
- Swaps typically between near-neighbor only, with randomized or deterministic schedule of which pair to try (discrete time).
- Straightforward extension of infinite swapping to K temperatures $\tau_1 < \tau_2 < \dots < \tau_K$. Benefits of symmetrization/equilibration even greater, e.g. $I^\infty(\nu) < \infty$ means very fast equilibration of relative contributions from all permutations of (x_1, \dots, x_K) to joint distribution, even larger rate for lowest marginal.

Implementation issues and partial infinite swapping

- Applications of parallel tempering use many temperatures (e.g., $K = 30$ to 50) when V is complicated to overcome barriers of all different heights.
- Swaps typically between near-neighbor only, with randomized or deterministic schedule of which pair to try (discrete time).
- Straightforward extension of infinite swapping to K temperatures $\tau_1 < \tau_2 < \dots < \tau_K$. Benefits of symmetrization/equilibration even greater, e.g. $I^\infty(\nu) < \infty$ means very fast equilibration of relative contributions from all permutations of (x_1, \dots, x_K) to joint distribution, even 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

- Applications of parallel tempering use many temperatures (e.g., $K = 30$ to 50) when V is complicated to overcome barriers of all different heights.
- Swaps typically between near-neighbor only, with randomized or deterministic schedule of which pair to try (discrete time).
- Straightforward extension of infinite swapping to K temperatures $\tau_1 < \tau_2 < \dots < \tau_K$. Benefits of symmetrization/equilibration even greater, e.g. $I^\infty(\nu) < \infty$ means very fast equilibration of relative contributions from all permutations of (x_1, \dots, x_K) to joint distribution, even 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

Full infinite swapping has the property that if

$$(Y_1(t), \dots, Y_K(t)) = (y_1, \dots, y_K),$$

then contributions from all permutations of (y_1, \dots, y_K) are added to $\eta^T(dx)$ according to their relative weights under $\pi(dx)$.

Implementation issues and partial infinite swapping

Full infinite swapping has the property that if

$$(Y_1(t), \dots, Y_K(t)) = (y_1, \dots, y_K),$$

then contributions from all permutations of (y_1, \dots, y_K) are added to $\eta^T(dx)$ according to their relative weights under $\pi(dx)$. We say that the full set of permutations are *instantaneously equilibrated*.

Implementation issues and partial infinite swapping

Full infinite swapping has the property that if

$$(Y_1(t), \dots, Y_K(t)) = (y_1, \dots, y_K),$$

then contributions from all permutations of (y_1, \dots, y_K) are added to $\eta^T(dx)$ according to their relative weights under $\pi(dx)$. We say that the full set of permutations are *instantaneously equilibrated*.

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

Implementation issues and partial infinite swapping

Full infinite swapping has the property that if

$$(Y_1(t), \dots, Y_K(t)) = (y_1, \dots, y_K),$$

then contributions from all permutations of (y_1, \dots, y_K) are added to $\eta^T(dx)$ according to their relative weights under $\pi(dx)$. We say that the full set of permutations are *instantaneously equilibrated*.

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

Subgroup property means the corresponding dynamic can be realized as weak limit of a parallel tempering algorithm.

Implementation issues and partial infinite swapping

Full infinite swapping has the property that if

$$(Y_1(t), \dots, Y_K(t)) = (y_1, \dots, y_K),$$

then contributions from all permutations of (y_1, \dots, y_K) are added to $\eta^T(dx)$ according to their relative weights under $\pi(dx)$. We say that the full set of permutations are *instantaneously equilibrated*.

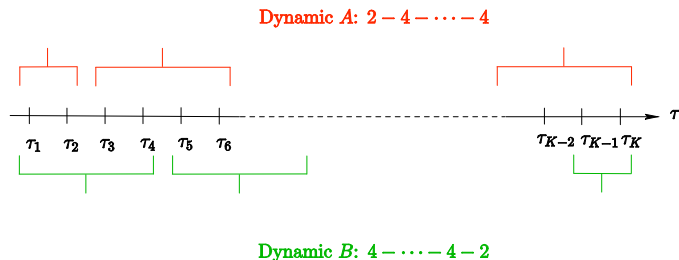
Partial infinite swapping. Given any subgroup of set of permutations one can construct a corresponding *partial infinite swapping* dynamic.

Subgroup property means the corresponding dynamic can be realized as weak limit of a parallel tempering algorithm.

Depending on the starting point, a particular coset of the subgroup is instantaneously equilibrated.

Implementation issues and partial infinite swapping

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 physical 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 physical 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.

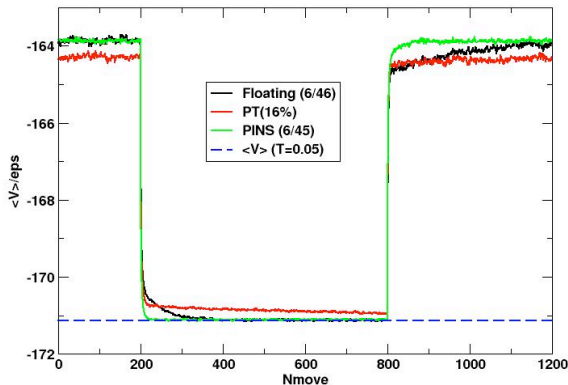
Numerical example

Relaxation study of convergence to equilibrium for LJ-38.

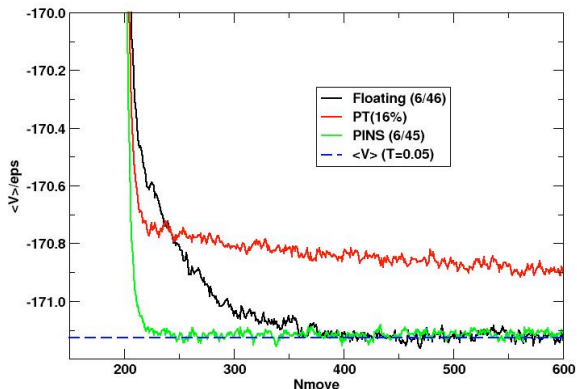
- quantity of interest: average potential energy at various temperatures
- used 45 temperatures, alternate 3-6-6-...-6 and 6-6-...-6-3 type dynamics for partial infinite swapping
- lowest 1/3 of temperatures raised to push process away from equilibrium (low temperature components pushed away from deep 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 example

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



Numerical example



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

Mathematical paper:

- “On the infinite swapping limit for parallel tempering”, D, Wang, Liu, Plattner and Doll, to be submitted to *SIAM J. on MMS*

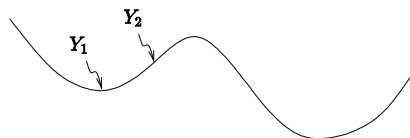
Applications paper (lots of numerical data):

- “An infinite swapping approach to the rare-event sampling problem”, Plattner, Doll, D, Wang, Liu and Gubernatis, submitted to *J. of Chem. Phy.* ([//arxiv.org/abs/1106.6305](http://arxiv.org/abs/1106.6305))

Concluding remarks

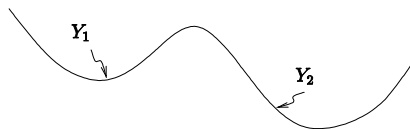
Function minimization attributes:

The infinite swapping process (Y_1, Y_2) has symmetric dynamics, interesting qualitative properties w.r.t. function minimization.



diffusion coefficient of Y_1 is $\approx \tau_1$

diffusion coefficient of Y_2 is $\approx \tau_2$

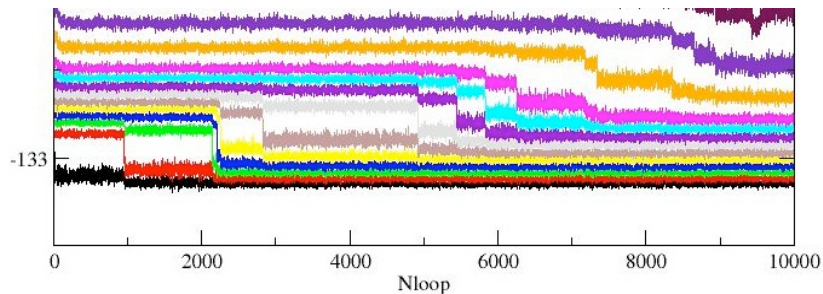


diffusion coefficient of Y_1 is $\approx \tau_2$

diffusion coefficient of Y_2 is $\approx \tau_1$

Concluding remarks

Convergence to equilibrium, single sample, 12 lowest temperatures:



Concluding remarks

Many open questions.

- Selection of “best” partial infinite swapping approximations.

Concluding remarks

Many open questions.

- Selection of “best” partial infinite swapping approximations.
- Use of other parameters besides temperature.

Concluding remarks

Many open questions.

- Selection of “best” partial infinite swapping approximations.
- Use of other parameters besides temperature.
- Better quantitative understanding of rate of marginals such as $I_1^\infty(\gamma)$

Concluding remarks

Many open questions.

- Selection of “best” partial infinite swapping approximations.
- Use of other parameters besides temperature.
- Better quantitative understanding of rate of marginals such as $I_1^\infty(\gamma)$
- Extension to related algorithms (simulated tempering).

Concluding remarks

Many open questions.

- Selection of “best” partial infinite swapping approximations.
- Use of other parameters besides temperature.
- Better quantitative understanding of rate of marginals such as $I_1^\infty(\gamma)$
- Extension to related algorithms (simulated tempering).
- Application to other difficult problems.