# On the infinite swapping limit for parallel tempering

Yufei Liu

Division of Applied Mathematics
Brown University

with Paul Dupuis, J.D. Doll, Nuria Plattner

RESIM 2012

June, 2012

# The problem

# The problem

One is interested in computing the integral

$$\int_{\mathbb{R}^d} f(x)\pi(dx), \tag{1}$$

where $\pi$ is some probability distribution that is infeasible to sample directly from.

# The problem

One is interested in computing the integral

$$\int_{\mathbb{R}^d} f(x)\pi(dx), \tag{1}$$

where $\pi$ is some probability distribution that is infeasible to sample directly from.

A standard method to deal with this kind of problem is via MCMC in which one constructs an ergodic Markov chain with $\pi$ as its invariant distribution. If $\mu_T$ is the empirical measure, then under suitable ergodicity conditions (communicating, aperiodic)

# The problem

One is interested in computing the integral

$$\int_{\mathbb{R}^d} f(x)\pi(dx), \tag{1}$$

where $\pi$ is some probability distribution that is infeasible to sample directly from.

A standard method to deal with this kind of problem is via MCMC in which one constructs an ergodic Markov chain with $\pi$ as its invariant distribution. If $\mu_T$ is the empirical measure, then under suitable ergodicity conditions (communicating, aperiodic)

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

would converge to (1).

# The problem

One is interested in computing the integral

$$\int_{\mathbb{R}^d} f(x)\pi(dx), \tag{1}$$

where $\pi$ is some probability distribution that is infeasible to sample directly from.

A standard method to deal with this kind of problem is via MCMC in which one constructs an ergodic Markov chain with $\pi$ as its invariant distribution. If $\mu_T$ is the empirical measure, then under suitable ergodicity conditions (communicating, aperiodic)

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

would converge to (1).

If the underlying distribution $\pi$ is unimodal, the sampling is straightforward and the associated numerical results are reliable.

# Examples: 1. Statistical mechanics

# Examples: 1. Statistical mechanics

Let $V$ denote the total potential energy of a statistical mechanical system.

# Examples: 1. Statistical mechanics

Let $V$ denote the total potential energy of a statistical mechanical system. $\pi$ is the Boltzmann distribution (canonical ensemble)

$$\pi(x) \doteq \frac{1}{Z(\tau)} e^{-V(x)/\tau},$$

# Examples: 1. Statistical mechanics

Let $V$ denote the total potential energy of a statistical mechanical system. $\pi$ is the Boltzmann distribution (canonical ensemble)

$$\pi(x) \doteq \frac{1}{Z(\tau)} e^{-V(x)/\tau},$$

where $\tau$ is the scaled temperature and $Z(\tau)$ is the partition function.

# Examples: 1. Statistical mechanics

Let $V$ denote the total potential energy of a statistical mechanical system. $\pi$ is the Boltzmann distribution (canonical ensemble)

$$\pi(x) \doteq \frac{1}{Z(\tau)} e^{-V(x)/\tau},$$

where $\tau$ is the scaled temperature and $Z(\tau)$ is the partition function.

One is interested in computing quantities such as the average potential.

# Examples: 1. Statistical mechanics

Let $V$ denote the total potential energy of a statistical mechanical system. $\pi$ is the Boltzmann distribution (canonical ensemble)

$$\pi(x) \doteq \frac{1}{Z(\tau)} e^{-V(x)/\tau},$$

where $\tau$ is the scaled temperature and $Z(\tau)$ is the partition function.

One is interested in computing quantities such as the average potential.

MCMC: Metropolis-Hastings type algorithm or stochastic dynamics method (based on a physical analogy, e.g. Andersen 1980).

# Examples: 1. Statistical mechanics

Let $V$ denote the total potential energy of a statistical mechanical system. $\pi$ is the Boltzmann distribution (canonical ensemble)

$$\pi(x) \doteq \frac{1}{Z(\tau)} e^{-V(x)/\tau},$$

where $\tau$ is the scaled temperature and $Z(\tau)$ is the partition function.

One is interested in computing quantities such as the average potential.

MCMC: Metropolis-Hastings type algorithm or stochastic dynamics method (based on a physical analogy, e.g. Andersen 1980).

Note that the higher the temperature, the "flatter" the distribution, the less likely for the Markov chain to get stuck at local minima of $V$.

# Examples: 2. Bayesian statistics

# Examples: 2. Bayesian statistics

Given a prior distribution $p(\theta)$, a likelihood model $P(D|\theta)$ and data $D$.

# Examples: 2. Bayesian statistics

Given a prior distribution $p(\theta)$, a likelihood model $P(D|\theta)$ and data $D$. $\pi$ is the posterior distribution

$$\pi(\theta|D) \propto P(D|\theta)\, p(\theta).$$

# Examples: 2. Bayesian statistics

Given a prior distribution $p(\theta)$, a likelihood model $P(D|\theta)$ and data $D$. $\pi$ is the posterior distribution

$$\pi(\theta|D) \propto P(D|\theta)\, p(\theta).$$

MCMC: Metropolis-Hastings type algorithm.

# Examples: 2. Bayesian statistics

Given a prior distribution $p(\theta)$, a likelihood model $P(D|\theta)$ and data $D$. $\pi$ is the posterior distribution

$$\pi(\theta|D) \propto P(D|\theta)\, p(\theta).$$

MCMC: Metropolis-Hastings type algorithm.
Define

$$V(\theta) \doteq -\log P(D|\theta).$$

then $\pi(\theta|D) \propto e^{-V(\theta)} p(\theta)$. For $\tau \geq 1$ define

$$\pi_\tau(\theta|D) \propto e^{-V(\theta)/\tau} p(\theta) = P(D|\theta)^{1/\tau}\, p(\theta).$$

# Examples: 2. Bayesian statistics

Given a prior distribution $p(\theta)$, a likelihood model $P(D|\theta)$ and data $D$. $\pi$ is the posterior distribution

$$\pi(\theta|D) \propto P(D|\theta)\, p(\theta).$$

MCMC: Metropolis-Hastings type algorithm.
Define

$$V(\theta) \doteq -\log P(D|\theta).$$

then $\pi(\theta|D) \propto e^{-V(\theta)} p(\theta)$. For $\tau \geq 1$ define

$$\pi_\tau(\theta|D) \propto e^{-V(\theta)/\tau} p(\theta) = P(D|\theta)^{1/\tau}\, p(\theta).$$

MC for $\pi_\tau$ $(\tau > 1)$ results in easier movement among local minima of $V(\theta)$.

# Examples: 3. A minimization problem

# Examples: 3. A minimization problem

Minimize a function $f$ over a set $\Omega$. Construct a Markov chain using a Metropolis-Hastings type algorithm with $\pi_\tau$ as the invariant distribution:

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

Here $\tau$ is chosen such that $\tau > 0$.

# Examples: 3. A minimization problem

Minimize a function $f$ over a set $\Omega$. Construct a Markov chain using a Metropolis-Hastings type algorithm with $\pi_\tau$ as the invariant distribution:

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

Here $\tau$ is chosen such that $\tau > 0$. Markov chain favors better minimization solutions of $f$.

# Examples: 3. A minimization problem

Minimize a function $f$ over a set $\Omega$. Construct a Markov chain using a Metropolis-Hastings type algorithm with $\pi_\tau$ as the invariant distribution:

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

Here $\tau$ is chosen such that $\tau > 0$. Markov chain favors better minimization solutions of $f$. As $\tau \to 0$, $\pi_\tau$ sharply peaked around global minimum; as $\tau \to \infty$, $\pi_\tau$ approximate uniform distribution on $\Omega$.

# Examples: 3. A minimization problem

Minimize a function $f$ over a set $\Omega$. Construct a Markov chain using a Metropolis-Hastings type algorithm with $\pi_\tau$ as the invariant distribution:

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

Here $\tau$ is chosen such that $\tau > 0$. Markov chain favors better minimization solutions of $f$. As $\tau \to 0$, $\pi_\tau$ sharply peaked around global minimum; as $\tau \to \infty$, $\pi_\tau$ approximate uniform distribution on $\Omega$.
Minimization algorithm: sample Markov chain under small $\tau$.

# Examples: 3. A minimization problem

Minimize a function $f$ over a set $\Omega$. Construct a Markov chain using a Metropolis-Hastings type algorithm with $\pi_\tau$ as the invariant distribution:

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

Here $\tau$ is chosen such that $\tau > 0$. Markov chain favors better minimization solutions of $f$. As $\tau \to 0$, $\pi_\tau$ sharply peaked around global minimum; as $\tau \to \infty$, $\pi_\tau$ approximate uniform distribution on $\Omega$.
Minimization algorithm: sample Markov chain under small $\tau$.
However, small $\tau$ results in less mobility, the chain more easily get stuck in local minima of $f$.

# The challenge

# The challenge

Use Gibbs distribution to illustrate. $\pi(x) \propto e^{-V(x)/\tau}$. When $\tau$ is small, the main contribution of

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

comes from the global minimum and "important" local minima of $V$.

# The challenge

Use Gibbs distribution to illustrate. $\pi(x) \propto e^{-V(x)/\tau}$. When $\tau$ is small, the main contribution of

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

comes from the global minimum and "important" local minima of $V$.

When $V$ has various deep local minima that are separated by steep "barriers", the underlying probability distribution $\pi$ has multiple isolated parts that communicate poorly with each other, in which case the scheme can be extremely slow to converge (the rare event problem).

# The challenge

Use Gibbs distribution to illustrate. $\pi(x) \propto e^{-V(x)/\tau}$. When $\tau$ is small, the main contribution of

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

comes from the global minimum and "important" local minima of $V$.

When $V$ has various deep local minima that are separated by steep "barriers", the underlying probability distribution $\pi$ has multiple isolated parts that communicate poorly with each other, in which case the scheme can be extremely slow to converge (the rare event problem).

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

# The challenge

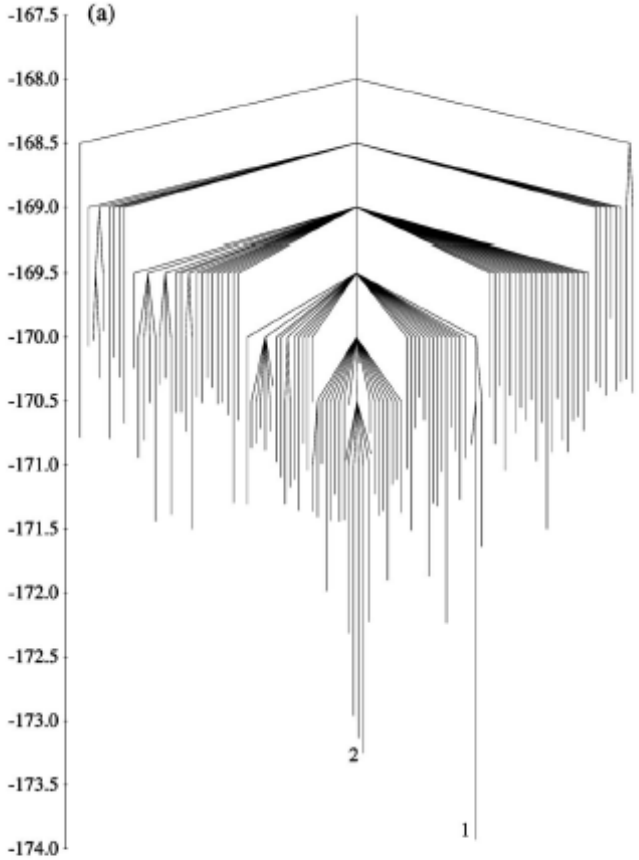
Use Gibbs distribution to illustrate. $\pi(x) \propto e^{-V(x)/\tau}$. When $\tau$ is small, the main contribution of

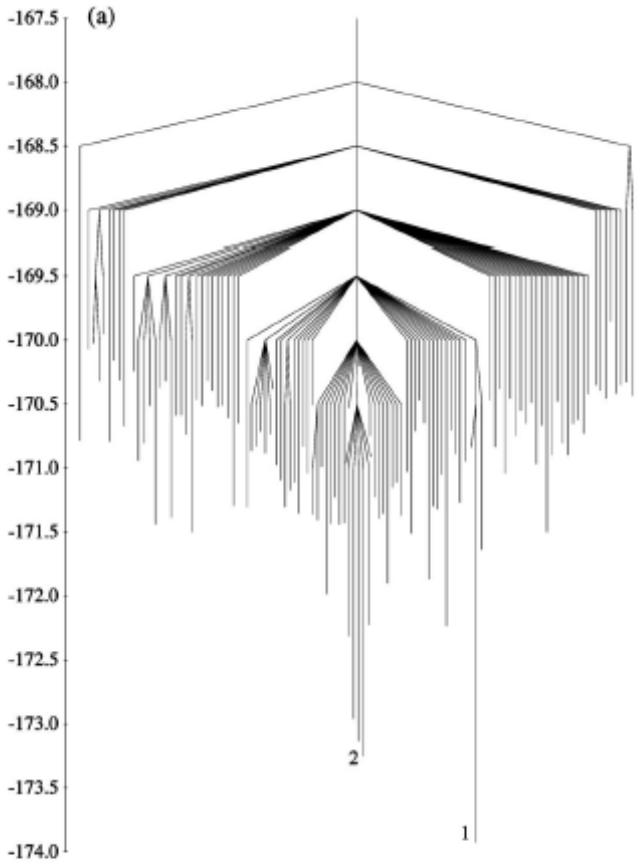
$$\int_{\mathbb{R}^d} f(x)\pi(dx)$$

comes from the global minimum and "important" local minima of $V$.

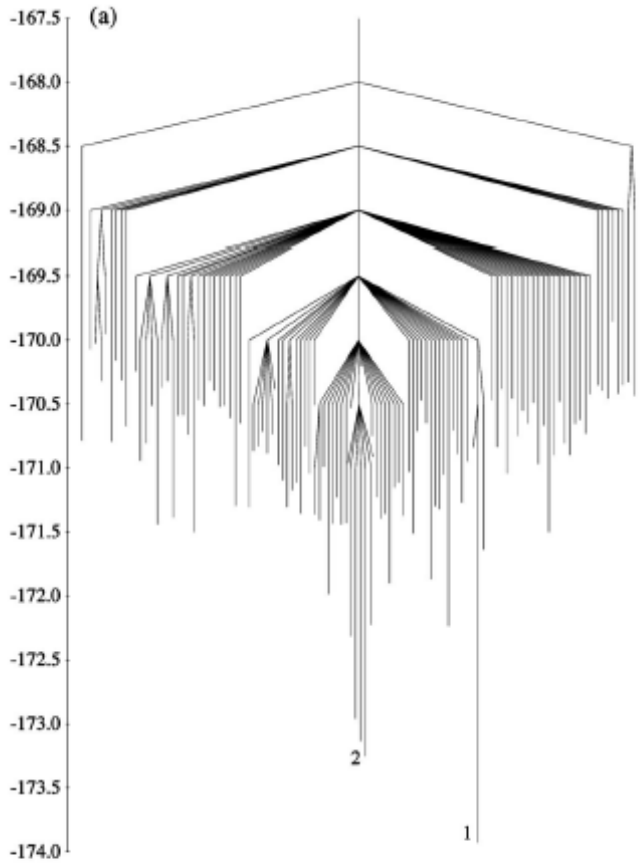
When $V$ has various deep local minima that are separated by steep "barriers", the underlying probability distribution $\pi$ has multiple isolated parts that communicate poorly with each other, in which case the scheme can be extremely slow to converge (the rare event problem).

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).






Global minimum only discovered 10+ years ago.




Global minimum only discovered 10+ years ago. Focus on overcoming rare-event sampling issues.

# Outline

# Outline

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

# Outline

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

# Outline

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

# Outline

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

# Standard measures of performance

# Standard measures of performance

*How should one describe the rate of convergence*

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

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*

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

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

**$2^{nd}$ 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*

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

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

**$2^{nd}$ 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

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

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 and their shortcomings

**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 and their shortcomings

**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 and their shortcomings

**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.

# A representative example

# A representative example

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

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

# A representative example

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

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

A corresponding continuous time model is

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

where $\tau$ is a fixed temperature (properly scaled).

# A representative example

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

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

A corresponding continuous time model is

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

where $\tau$ is a fixed temperature (properly scaled).

- Simulations are done using a discrete time model.

# An accelerated algorithm: parallel tempering

# An accelerated algorithm: parallel tempering

Besides $\tau_1 = \tau$, introduce higher temperature $\tau_2 > \tau_1$.

# An accelerated algorithm: parallel tempering

Besides $\tau_1 = \tau$, introduce higher temperature $\tau_2 > \tau_1$. Thus

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

with $W_1$ and $W_2$ independent.

# An accelerated algorithm: parallel tempering

Besides $\tau_1 = \tau$, introduce higher temperature $\tau_2 > \tau_1$. Thus

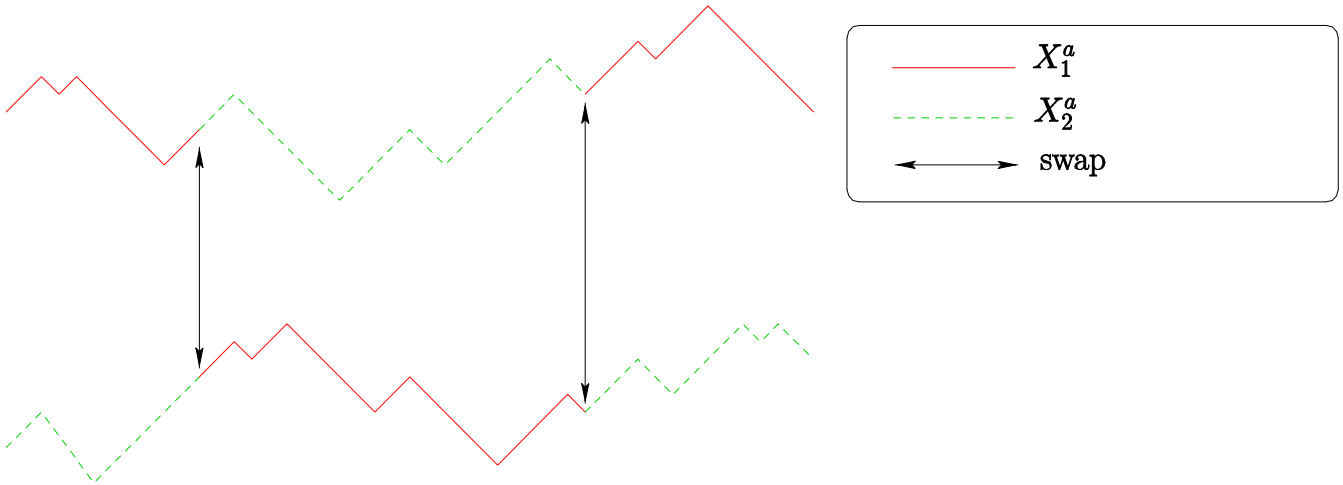
$$dX_1^a = -\nabla V(X_1^a)dt + \sqrt{2\tau_1}dW_1$$
$$dX_2^a = -\nabla V(X_2^a)dt + \sqrt{2\tau_2}dW_2,$$

with $W_1$ and $W_2$ independent. Now introduce *swaps* (Swendsen, Geyer), i.e., $X_1^a$ and $X_2^a$ *exchange locations* with state dependent intensity

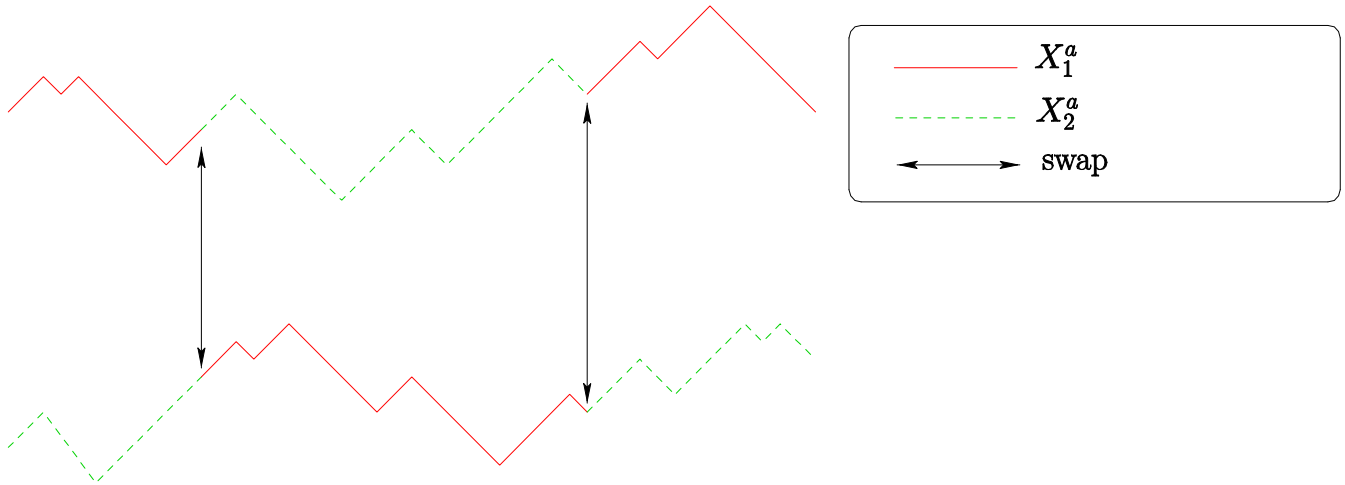
$$ag(x_1, x_2) \doteq a\left(1 \wedge \frac{\pi_{\tau_1}(x_2)\,\pi_{\tau_2}(x_1)}{\pi_{\tau_1}(x_1)\,\pi_{\tau_2}(x_2)}\right),$$

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

# An accelerated algorithm: parallel tempering



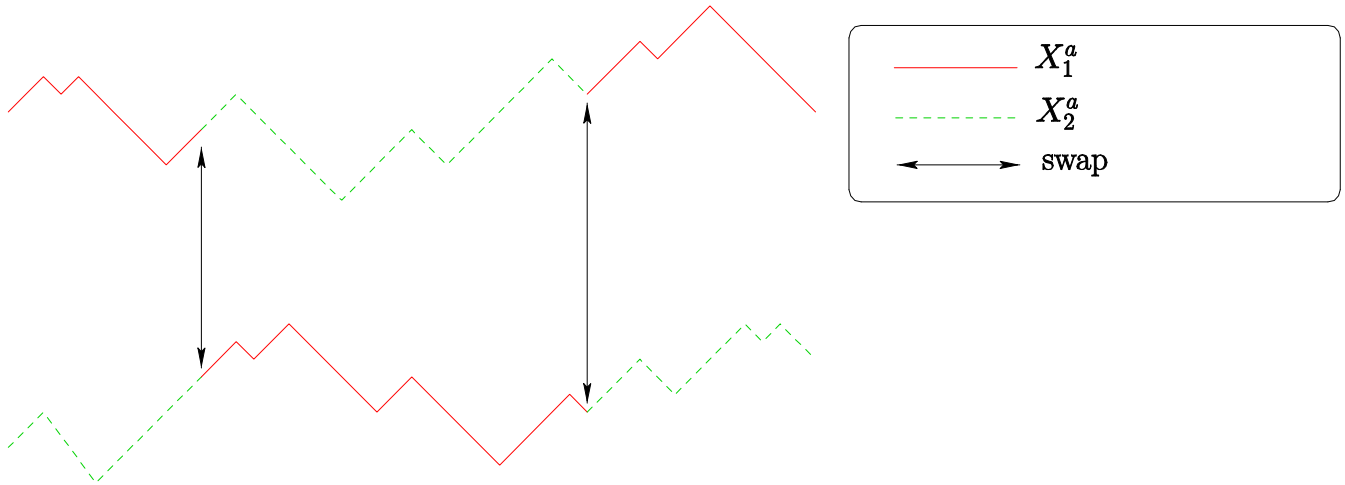
# An accelerated algorithm: parallel tempering




One can check (detailed balance condition): with this swapping intensity, invariant distribution of the joint process

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

# An accelerated algorithm: parallel tempering




One can check (detailed balance condition): with this swapping intensity, invariant distribution of the joint process

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

Use the first marginal of the empirical measure.

# Parallel tempering analysis

In practice, much more temperatures $(30 - 50)$ are used.

# Parallel tempering analysis

In practice, much more temperatures $(30 - 50)$ are used. Bring in higher temperatures

1. Higher temperature simulations correspond to higher volatility.

# Parallel tempering analysis

In practice, much more temperatures $(30 - 50)$ are used. Bring in higher temperatures

1. Higher temperature simulations correspond to higher volatility.
2. High-energy barriers are more easily crossed for simulations carried out in higher temperatures.

# Parallel tempering analysis

In practice, much more temperatures $(30 - 50)$ are used. Bring in higher temperatures

1. Higher temperature simulations correspond to higher volatility.
2. High-energy barriers are more easily crossed for simulations carried out in higher temperatures.
3. Swapping enables information flow from high temperatures to low temperatures.

# Donsker-Varadhan rate of decay

# Donsker-Varadhan rate of decay

How does convergence depend on swap rate $a$?

# Donsker-Varadhan rate of decay

How does convergence depend on swap rate $a$?
Donsker-Varadhan theory for empirical measure. Let $I$ denote the large deviations rate function.

# Donsker-Varadhan rate of decay

How does convergence depend on swap rate $a$?
Donsker-Varadhan theory for empirical measure. Let $I$ denote the large deviations rate function.
Let $S$ denote the state space, for any $\mu \in \mathcal{P}(S)$, if $N_\delta(\mu)$ is a $\delta$-neighborhood of $\mu$ under weak topology

# Donsker-Varadhan rate of decay

How does convergence depend on swap rate $a$?
Donsker-Varadhan theory for empirical measure. Let $I$ denote the large deviations rate function.
Let $S$ denote the state space, for any $\mu \in \mathcal{P}(S)$, if $N_\delta(\mu)$ is a $\delta$-neighborhood of $\mu$ under weak topology

$$P(\mu_T \in N_\delta(\mu)) \approx e^{-T(I(\mu)+\varepsilon(\delta))}$$

# Donsker-Varadhan rate of decay

How does convergence depend on swap rate $a$?
Donsker-Varadhan theory for empirical measure. Let $I$ denote the large deviations rate function.
Let $S$ denote the state space, for any $\mu \in \mathcal{P}(S)$, if $N_\delta(\mu)$ is a $\delta$-neighborhood of $\mu$ under weak topology

$$P(\mu_T \in N_\delta(\mu)) \approx e^{-T(I(\mu)+\varepsilon(\delta))}$$

To achieve maximum rate of convergence, we choose $a$ such that $I^a$ is the largest possible.

# Large deviation rate function

Under mild conditions on $V$, one can calculate $I$ explicitly (Donser-Varadhan).

# Large deviation rate function

Under mild conditions on $V$, one can calculate $I$ explicitly (Donser-Varadhan). Suppose $\nu \in \mathcal{P}(S)$ is given by

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

# Large deviation rate function

Under mild conditions on $V$, one can calculate $I$ explicitly (Donser-Varadhan). Suppose $\nu \in \mathcal{P}(S)$ is given by

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

Then we have *monotonic* form

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

where $J_0$ is the rate for "no swap" dynamics; $J_1$ is nonnegative and

# Large deviation rate function

Under mild conditions on $V$, one can calculate $I$ explicitly (Donser-Varadhan). Suppose $\nu \in \mathcal{P}(S)$ is given by

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

Then we have *monotonic* form

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

where $J_0$ is the rate for "no swap" dynamics; $J_1$ is nonnegative and

$$J_1(\nu) = 0 \text{ iff } \theta(x_2, x_1) = \theta(x_1, x_2) \ \nu\text{-a.s.}$$

# Limit of rate function

Thus for $I^a(\nu) \uparrow \infty$ as $a \uparrow \infty$ ($\nu$ is very unlikely) unless

# Limit of rate function

Thus for $I^a(\nu) \uparrow \infty$ as $a \uparrow \infty$ ($\nu$ is very unlikely) unless

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

# Limit of rate function

Thus for $I^a(\nu) \uparrow \infty$ as $a \uparrow \infty$ ($\nu$ is very unlikely) unless

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

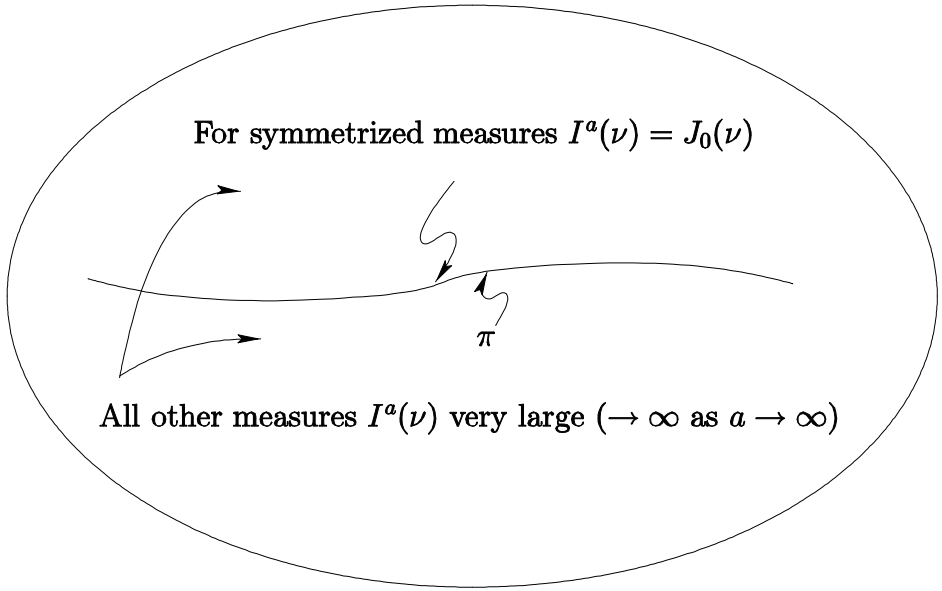
If we call measures that place precisely same relative weight on permutations $(x_1, x_2)$ and $(x_2, x_1)$ as $\pi$ *symmetrized measures*, then

# Limit of rate function

Thus for $I^a(\nu) \uparrow \infty$ as $a \uparrow \infty$ ($\nu$ is very unlikely) unless

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

If we call measures that place precisely same relative weight on permutations $(x_1, x_2)$ and $(x_2, x_1)$ as $\pi$ *symmetrized measures*, then



# Limit of rate function (cont'd)

By contraction principle, for probability measure $\gamma$

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

# Limit of rate function (cont'd)

By contraction principle, for probability measure $\gamma$

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

$$I_1^a(\gamma) \uparrow \text{ as } a \uparrow .$$

# Limit of rate function (cont'd)

By contraction principle, for probability measure $\gamma$

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

$$I_1^a(\gamma) \uparrow \text{ as } a \uparrow .$$

This suggests one consider the infinite swapping limit $a \uparrow \infty$.

# Limit of rate function (cont'd)

By contraction principle, for probability measure $\gamma$

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

$$I_1^a(\gamma) \uparrow \text{ as } a \uparrow .$$

This suggests one consider the infinite swapping limit $a \uparrow \infty$.
Unfortunately, limit process is not well defined (no tightness).

# Limit of rate function (cont'd)

By contraction principle, for probability measure $\gamma$

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

$$I_1^a(\gamma) \uparrow \text{ as } a \uparrow .$$

This suggests one consider the infinite swapping limit $a \uparrow \infty$. Unfortunately, limit process is not well defined (no tightness).
**An alternative perspective:** rather than swap particles, swap temperatures, and use "weighted" empirical measure.

# "Temperature swapping "process

Temperature swapping process:

# "Temperature swapping "process

Temperature swapping process:

$$dY_1^a = -\nabla V(Y_1^a)dt + \sqrt{2\tau_1 \mathbf{1}(Z^a = 1) + 2\tau_2 \mathbf{1}(Z^a = 2)}dW_1$$
$$dY_2^a = -\nabla V(Y_2^a)dt + \sqrt{2\tau_2 \mathbf{1}(Z^a = 1) + 2\tau_1 \mathbf{1}(Z^a = 2)}dW_2,$$

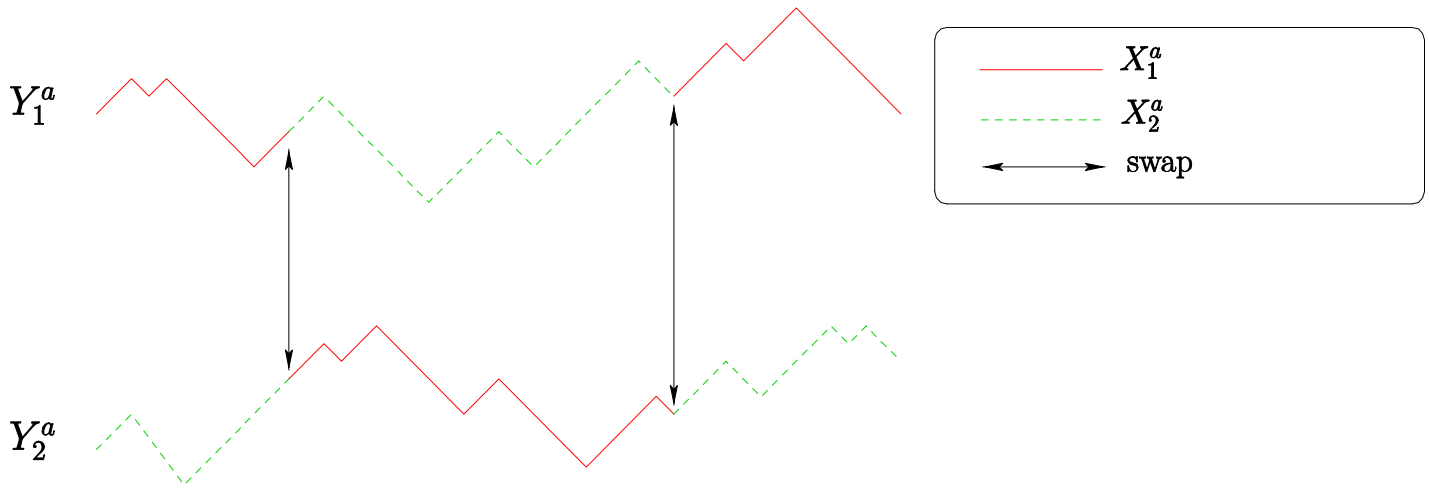
where $Z^a(t)$ jumps from $1 \to 2$ with intensity $ag(Y_1^a(t), Y_2^a(t))$ and from $2 \to 1$ with intensity $ag(Y_2^a(t), Y_1^a(t))$.

# "Temperature swapping "process

Temperature swapping process:

$$dY_1^a = -\nabla V(Y_1^a)dt + \sqrt{2\tau_1 \mathbf{1}(Z^a = 1) + 2\tau_2 \mathbf{1}(Z^a = 2)}dW_1$$
$$dY_2^a = -\nabla V(Y_2^a)dt + \sqrt{2\tau_2 \mathbf{1}(Z^a = 1) + 2\tau_1 \mathbf{1}(Z^a = 2)}dW_2,$$

where $Z^a(t)$ jumps from $1 \to 2$ with intensity $ag(Y_1^a(t), Y_2^a(t))$ and from $2 \to 1$ with intensity $ag(Y_2^a(t), Y_1^a(t))$.



# Infinite swapping limit

Instead of using ordinary empirical measure

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

# Infinite swapping limit

Instead of using ordinary empirical measure

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

use weighted empirical measure $\eta_T^a$:

$$\frac{1}{T}\int_0^T \left[\mathbf{1}(Z^a = 1)\,\delta_{(Y_1^a, Y_2^a)}(\cdot) + \mathbf{1}(Z^a = 2)\,\delta_{(Y_2^a, Y_1^a)}(\cdot)\right] dt.$$

Ergodic theory $\eta_T^a \to \pi$ . $(Y_1^a, Y_2^a, \eta_T^a)$ admits a well defined weak limit $a \to \infty$.

# Infinite swapping limit

Instead of using ordinary empirical measure

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

use weighted empirical measure $\eta_T^a$:

$$\frac{1}{T}\int_0^T \left[\mathbf{1}(Z^a = 1)\,\delta_{(Y_1^a, Y_2^a)}(\cdot) + \mathbf{1}(Z^a = 2)\,\delta_{(Y_2^a, Y_1^a)}(\cdot)\right] dt.$$

Ergodic theory $\eta_T^a \to \pi$ . $(Y_1^a, Y_2^a, \eta_T^a)$ admits a well defined weak limit $a \to \infty$. Define state dependent weight

$$\begin{aligned}
\rho_1(x_1, x_2) &\doteq \frac{\pi_{\tau_1}(x_1)\,\pi_{\tau_2}(x_2)}{\pi_{\tau_1}(x_1)\,\pi_{\tau_2}(x_2) + \pi_{\tau_1}(x_2)\,\pi_{\tau_2}(x_1)}, \\
\rho_2(x_1, x_2) &\doteq \frac{\pi_{\tau_1}(x_2)\,\pi_{\tau_2}(x_1)}{\pi_{\tau_1}(x_1)\,\pi_{\tau_2}(x_2) + \pi_{\tau_1}(x_1)\,\pi_{\tau_2}(x_2)}.
\end{aligned}$$

# Infinite swapping limit (cont'd)

The triple has following weak limit

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

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

# Infinite swapping limit (cont'd)

The triple has following weak limit

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

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

Theorem: for any sequence $a_T \uparrow \infty$, $\left\{\eta_T^{a_T}\right\}$ satisfies the uniform large deviations principle (in $T$) with rate $I^\infty$

# Infinite swapping limit (cont'd)

The triple has following weak limit

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

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

Theorem: for any sequence $a_T \uparrow \infty$, $\left\{\eta_T^{a_T}\right\}$ satisfies the uniform large deviations principle (in $T$) with rate $I^\infty$

$$\lim_{a\to\infty} I^a(\nu) = I^\infty(\nu) \doteq \begin{cases} J_0(\nu) & \text{if } \theta(x_1, x_2) = \theta(x_2, x_1) \\ \infty & \text{otherwise} \end{cases}$$

# Implementation issues

- 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

- 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 even greater.

# Implementation issues

- 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 even greater.
- But, coefficients become complex, e.g., $K!$ weights, and each involves many calculations.

# Implementation issues

- 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 even greater.
- But, coefficients become complex, e.g., $K!$ weights, and each involves many calculations.
- Need for computational feasibility leads to *partial infinite swapping*.

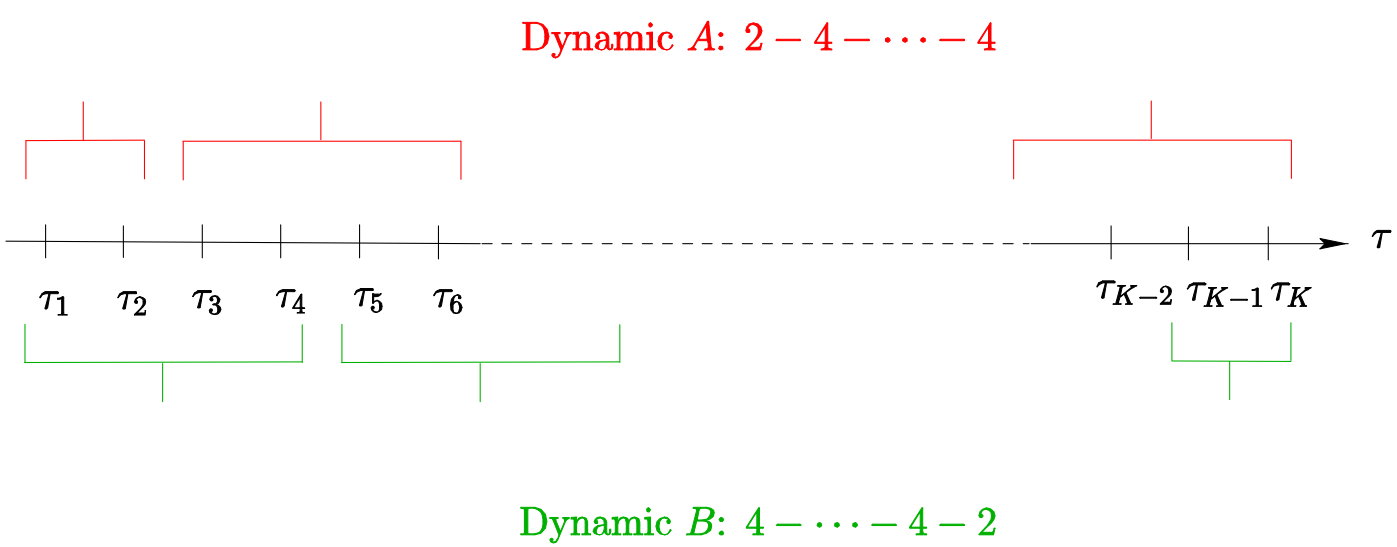
# Partial infinite swapping

# Partial infinite swapping

**Partial infinite swapping.** Instead of instantly symmetrizing all permutations, pick subgroups of the set of permutations (that can generate the whole permutation set) and construct corresponding *partial infinite swapping* dynamics within each group. Then alternate among each dynamics (need certain handoff rule, use proper weight).

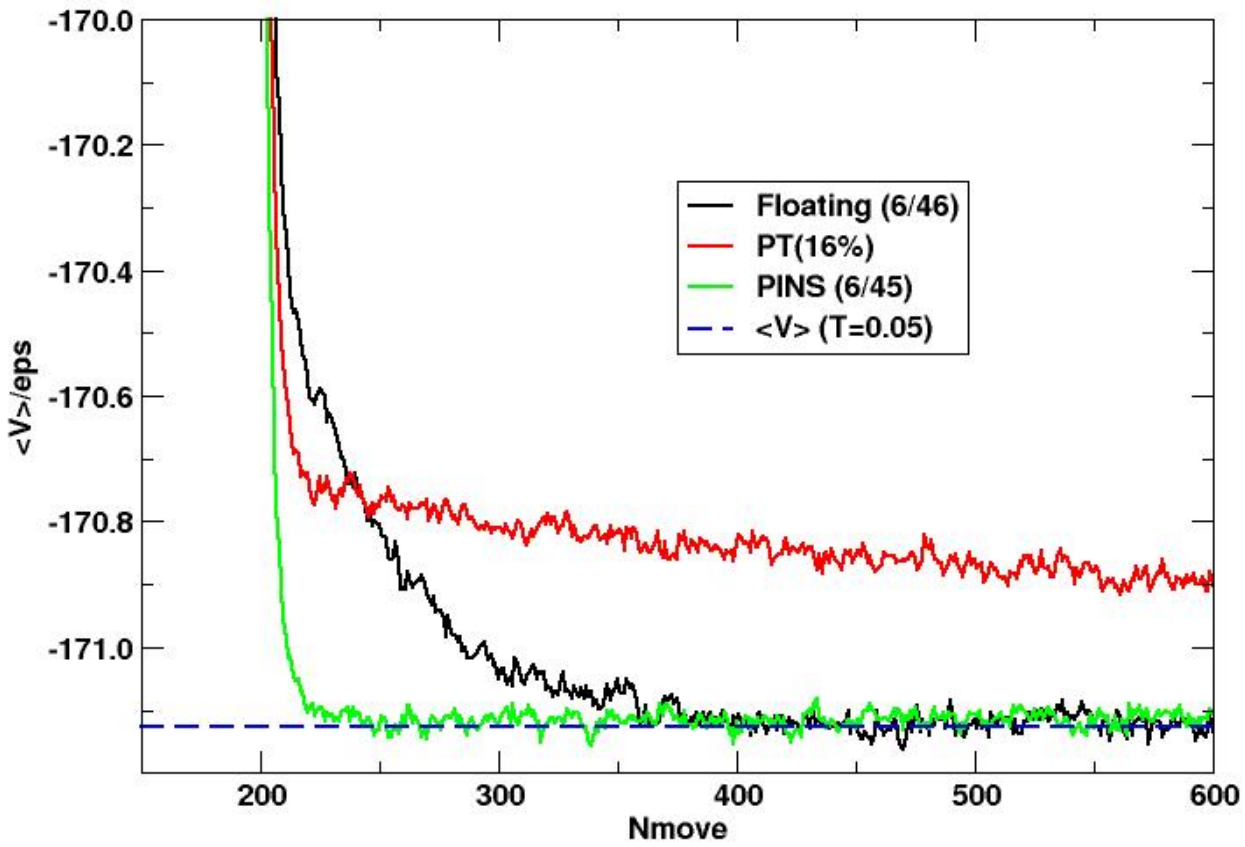
# Partial infinite swapping

**Partial infinite swapping.** Instead of instantly symmetrizing all permutations, pick subgroups of the set of permutations (that can generate the whole permutation set) and construct corresponding *partial infinite swapping* dynamics within each group. Then alternate among each dynamics (need certain handoff rule, use proper weight).
Examples are Dynamics $A$ and $B$ in figure:



# Comparison of PINS and PT

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



# Concluding remarks

References:

- Mathematical paper: "On the infinite swapping limit for parallel tempering", Dupuis, Liu, Plattner and Doll, to be appeared in SIAM J. on MMS
- Applications paper (lots of numerical data): "An infinite swapping approach to the rare-event sampling problem", Plattner, Doll, Dupuis, Wang, Liu and Gubernatis, J. of Chem. Phy. 135, 134111 (2011)

# Concluding remarks

References:

- Mathematical paper: "On the infinite swapping limit for parallel tempering", Dupuis, Liu, Plattner and Doll, to be appeared in SIAM J. on MMS
- Applications paper (lots of numerical data): "An infinite swapping approach to the rare-event sampling problem", Plattner, Doll, Dupuis, Wang, Liu and Gubernatis, J. of Chem. Phy. 135, 134111 (2011)

Many open questions.

- Selection of set of temperatures.

# Concluding remarks

References:

- Mathematical paper: "On the infinite swapping limit for parallel tempering", Dupuis, Liu, Plattner and Doll, to be appeared in SIAM J. on MMS
- Applications paper (lots of numerical data): "An infinite swapping approach to the rare-event sampling problem", Plattner, Doll, Dupuis, Wang, Liu and Gubernatis, J. of Chem. Phy. 135, 134111 (2011)

Many open questions.

- Selection of set of temperatures.
- Selection of "best" subgroups for partial infinite swapping approximations.

# Concluding remarks

References:

- Mathematical paper: "On the infinite swapping limit for parallel tempering", Dupuis, Liu, Plattner and Doll, to be appeared in SIAM J. on MMS
- Applications paper (lots of numerical data): "An infinite swapping approach to the rare-event sampling problem", Plattner, Doll, Dupuis, Wang, Liu and Gubernatis, J. of Chem. Phy. 135, 134111 (2011)

Many open questions.

- Selection of set of temperatures.
- Selection of "best" subgroups for partial infinite swapping approximations.
- Better quantitative understanding of rate of marginals such as $I_1^\infty(\gamma)$

# Concluding remarks

References:

- Mathematical paper: "On the infinite swapping limit for parallel tempering", Dupuis, Liu, Plattner and Doll, to be appeared in SIAM J. on MMS
- Applications paper (lots of numerical data): "An infinite swapping approach to the rare-event sampling problem", Plattner, Doll, Dupuis, Wang, Liu and Gubernatis, J. of Chem. Phy. 135, 134111 (2011)

Many open questions.

- Selection of set of temperatures.
- Selection of "best" subgroups for partial infinite swapping approximations.
- Better quantitative understanding of rate of marginals such as $I_1^\infty(\gamma)$
- Application to other problems such as function minimization.