A large deviation analysis of some properties of parallel tempering and infinite swapping algorithms

Pierre Nyquist

Division of Applied Mathematics
Brown University

YEP XIII
EURANDOM, March 8 2016.

joint work with J. Doll and P. Dupuis.
Monte Carlo methods are typically both easy to understand and implement. However, often suffer from the *rare-event sampling problem*:
Monte Carlo methods are typically both easy to understand and implement. However, often suffer from the *rare-event sampling problem*:

- The ergodic problem: Computing expected values (“thermodynamic properties”) with respect to a stationary distribution.
Monte Carlo methods are typically both easy to understand and implement. However, often suffer from the *rare-event sampling problem*:

- **The ergodic problem:** Computing expected values (“thermodynamic properties”) with respect to a stationary distribution.
- **Transition rate problems:** Computing the probability of transitions over $[0,T]$, exit locations, and mean exit times with respect to metastable states.
Monte Carlo methods are typically both easy to understand and implement. However, often suffer from the *rare-event sampling problem*:

- The ergodic problem: Computing expected values ("thermodynamic properties") with respect to a stationary distribution.
- Transition rate problems: Computing the probability of transitions over \([0,T]\), exit locations, and mean exit times with respect to metastable states.
- Functionals that depend heavily on the "tail" of the distribution (risk measures).
Today's focus: parallel tempering and infinite swapping.

**Example problem:** Compute average potential energy or other functionals w.r.t. a Gibbs measure

\[ \mu_1(dx) = e^{-V(x)/\tau_1} \lambda(dx)/Z(\tau_1). \]

\( V \) is the potential of some complicated physical system, \( \lambda \) a reference measure.

Think of \( V \) as having many local minima. A representative quantity of interest is

\[ \int_S V(x)e^{-V(x)/\tau_1} \lambda(dx)/Z(\tau). \]
“Real” problems can have thousands of local minima.

For examples: Two well potential; $\alpha$ determines the level of asymmetry.
For $\lambda$ Lebesgue measure, can use that $\mu_1$ is the stationary distribution of the solution of

$$dX(t) = -\nabla V(X(t))dt + \sqrt{2\tau_1}dW(t).$$

For $\lambda$ counting measure (finite state space $S$) can define Glauber dynamics with stationary distribution

$$\mu_1(x) = e^{-V(x)/\tau_1}/Z(\tau_1).$$

Consider a continuous-time Markov process $X(t)$ with $\mu_1$ as invariant distribution. Under ergodicity a numerical approximation to $\mu_1$ is given by the empirical measure

$$\eta_T(\cdot) = \frac{1}{T} \int_0^T \delta_{X(t)}(\cdot) dt.$$
Densities we are attempting to sample for $\tau_1 = 0.1$. 

![Graph showing densities for different values of $\alpha$.]
**Idea:** To accelerate convergence use *parallel tempering* (or *replica exchange*).

The idea is to use multiple temperatures $\tau_1 < \tau_2 < \ldots$.

Define Glauber dynamics $\Gamma_{x,y}^1$ and $\Gamma_{x,y}^2$ (rate matrices) corresponding to $\tau_1$ and $\tau_1$, respectively. Running two independent Markov processes, $X_1$ and $X_2$ according to these dynamics produces a Monte Carlo approximation to

$$\mu = \mu_1 \times \mu_2 \text{ on } S^2$$
Next, introduce swaps (at random times) between $X_1$ and $X_2$. State-dependent intensity:

$$ag(x_1, x_2) = a \left( 1 \wedge \frac{\mu(x_2, x_1)}{\mu(x_1, x_2)} \right), \ a > 0.$$
Next, introduce swaps (at random times) between $X_1$ and $X_2$. State-dependent intensity:

$$ag(x_1, x_2) = a \left(1 \wedge \frac{\mu(x_2, x_1)}{\mu(x_1, x_2)}\right), \quad a > 0.$$  

$X^a = (X_1^a, X_2^2)$: two-component process with swap rate $a$. Generator:

$$\mathcal{L}^a f(x_1, x_2) = \mathcal{L}^0 f(x_1, x_2) + ag(x_1, x_2) [f(x_2, x_1) - f(x_1, x_2)].$$

Straightforward to check - e.g., detailed balance - that $\mu$ remains the invariant measure.

How to choose the swap rate $a$ to ensure fast convergence?
Next, introduce swaps (at random times) between $X_1$ and $X_2$. 
State-dependent intensity:

$$ag(x_1, x_2) = a \left( 1 \wedge \frac{\mu(x_2, x_1)}{\mu(x_1, x_2)} \right), \quad a > 0.$$  

$X^a = (X_1^a, X_2^a)$: two-component process with swap rate $a$. Generator:

$$\mathcal{L}^a f(x_1, x_2) = \mathcal{L}^0 f(x_1, x_2) + ag(x_1, x_2) \left[ f(x_2, x_1) - f(x_1, x_2) \right].$$

Straightforward to check - e.g., detailed balance - that $\mu$ remains the invariant measure.

How to choose the swap rate $a$ to ensure fast convergence? **Large deviation analysis.**
The empirical measure $\lambda^a_T(\cdot) = \frac{1}{T} \int_0^T \delta_{\mathbf{x}^a(t)}(\cdot) dt$ satisfies an LDP ($T \to \infty$) with rate function

$$I^a(\nu) = I^0(\nu) + aJ(\nu),$$

where, if $\theta = d\nu/d\mu$ and $q$ is the jump intensity associated with the uncoupled dynamics,

$$I^0(\nu) = \int_{S^2} q(x)\nu(dx) - \int_{S^2 \times S^2} \sqrt{\theta(x)\theta(y)}\Gamma_{x,y}\mu(dx),$$

and

$$J(\nu) = \int_{S^2} g(x_1, x_2)l \left(\sqrt{\frac{\theta(x_2, x_1)}{\theta(x_1, x_2)}}\right) \nu(dx).$$
The empirical measure \( \lambda^a_T(\cdot) = \frac{1}{T} \int_0^T \delta_{X^a_t}(\cdot) dt \) satisfies an LDP \( (T \to \infty) \) with rate function

\[
I^a(\nu) = I^0(\nu) + aJ(\nu),
\]

where, if \( \theta = d\nu/d\mu \) and \( q \) is the jump intensity associated with the uncoupled dynamics,

\[
I^0(\nu) = \int_{S^2} q(x)\nu(dx) - \int_{S^2 \times S^2} \sqrt{\theta(x)\theta(y)}\Gamma_{x,y}\mu(dx),
\]

and

\[
J(\nu) = \int_{S^2} g(x_1, x_2) l \left( \frac{\sqrt{\theta(x_2, x_1)}}{\theta(x_1, x_2)} \right) \nu(dx).
\]

Monotonicity in \( a \) suggests letting \( a \to \infty \).
Now: switch temperatures/dynamics between the processes, not locations.

Take $Y^a = (Y_1^a, Y_2^a)$ to be the *temperature swapped* version of $X^a$. Consider the Markov process $(Y^a, Z^a)$, where $Z^a = \{Z^a(t)\}$ is a jump process that indicates temperature configuration at time $t$.

The weighted empirical measure

$$
\eta^a_T(\cdot) = \frac{1}{T} \int_0^T \left[ 1\{Z^a(t)=0\} \delta(Y_1^a(t), Y_2^a(t))(\cdot) + 1\{Z^a(t)=1\} \delta(Y_2^a(t), Y_1^a(t))(\cdot) \right] dt.
$$

has the same distribution as the empirical measure of $X^a$. By ergodicity $\eta^a_T$ converges to $\mu$ as $T \to \infty$.

Also, there is now hope for a limit in the swap rate $a \to \infty$. 

Nyquist (Brown)
The limit process $Y^\infty = (Y_1^\infty, Y_2^\infty)$ is a pure-jump Markov process with generator

$$\mathcal{L}^\infty f(x_1, x_2) = \sum_{(y_1, y_2) \in S^2} [f(y_1, y_2) - f(x_1, x_2)] \Gamma^\infty_{x,y},$$

where

$$\Gamma^\infty_{x,y} = \begin{cases} 
\rho(x_1, x_2) \Gamma^1_{x_1,y_1} + \rho(x_2, x_1) \Gamma^2_{x_1,y_1}, & y_1 \neq x_1, y_2 = x_2, \\
\rho(x_1, x_2) \Gamma^2_{x_2,y_2} + \rho(x_2, x_1) \Gamma^1_{x_2,y_2}, & y_1 = x_1, y_2 \neq x_2, \\
0, & \text{otherwise},
\end{cases}$$

and

$$\rho(x_1, x_2) = \frac{\mu(x_1, x_2)}{\mu(x_1, x_2) + \mu(x_2, x_1)}. $$
The limit of the weighted empirical measure $\eta_T$ is

$$
\eta_T^\infty(\cdot) = \frac{1}{T} \int_0^T \left[ \rho(Y_1^\infty(t), Y_2^\infty(t)) \delta(Y_1^\infty(t), Y_2^\infty(t))(\cdot) \\
+ \rho(Y_2^\infty(t), Y_1^\infty(t)) \delta(Y_2^\infty(t), Y_1^\infty(t))(\cdot) \right] dt.
$$

Ergodicity $\Rightarrow \eta_T^\infty \rightarrow \mu$ as $T \rightarrow \infty$. 
The limit of the weighted empirical measure $\eta_T^a$ is

$$\eta^\infty_T(\cdot) = \frac{1}{T} \int_0^T \left[ \rho(Y_1^\infty(t), Y_2^\infty(t))\delta(Y_1^\infty(t), Y_2^\infty(t))(\cdot) \
+ \rho(Y_2^\infty(t), Y_1^\infty(t))\delta(Y_2^\infty(t), Y_1^\infty(t))(\cdot) \right] dt.$$ 

Ergodicity $\Rightarrow \eta^\infty_T \rightarrow \mu$ as $T \rightarrow \infty$.

**Infinite swapping:** Simulate $Y^\infty$ and use $\eta^\infty_T$ for numerical approximations of $\mu$. 
The limit process $\mathbf{Y}^\infty$ has invariant measure

$$\bar{\mu}(x_1, x_2) = \frac{1}{2} [\mu(x_1, x_2) + \mu(x_2, x_1)].$$

Connectedness of the density much improved compared to PT.
The limit process $Y^\infty$ has invariant measure

$$\bar{\mu}(x_1, x_2) = \frac{1}{2} [\mu(x_1, x_2) + \mu(x_2, x_1)].$$

Connectedness of the density much improved compared to PT.
Large deviations: Let $\nu_T$ be the empirical measure associated with $Y^\infty$,

$$\nu_T(\cdot) = \frac{1}{T} \int_0^T \delta_{Y^\infty(t)}(\cdot) dt.$$ 

The sequence $\{\nu_T\}$ satisfies an LDP as $T \to \infty$ with rate function

$$I^\infty(\nu) = \int_{S^2} q^\infty(x) \nu(dx) - \int_{S^2 \times S^2} \sqrt{\theta(x)\theta(y)} \Gamma^{\infty}_{x,y} \bar{\mu}(dx),$$

where $\theta = d\nu/d\bar{\mu}$ and $q^\infty$ is the intensity associated with $\Gamma^\infty$. 
Large deviations: Let $\nu_T$ be the empirical measure associated with $Y^\infty$, 
\[ \nu_T(\cdot) = \frac{1}{T} \int_0^T \delta_{Y^\infty(t)}(\cdot) \, dt. \]

The sequence $\{\nu_T\}$ satisfies an LDP as $T \to \infty$ with rate function
\[ I^\infty(\nu) = \int_{S^2} q^\infty(x) \nu(dx) - \int_{S^2 \times S^2} \sqrt{\theta(x)\theta(y)} \Gamma^\infty_{x,y} \bar{\mu}(dx), \]
where $\theta = d\nu/d\bar{\mu}$ and $q^\infty$ is the intensity associated with $\Gamma^\infty$.

Let $M : \mathcal{P}(S^2) \to \mathcal{P}(S^2)$ be the mapping for which $M \nu_T = \eta^\infty_T$ (also takes $\bar{\mu}$ to $\mu$). By the contraction principle we retrieve the LDP for $\{\eta^\infty_T\}$.
Large deviations: Let $\nu_T$ be the empirical measure associated with $Y^\infty$,

$$\nu_T(\cdot) = \frac{1}{T} \int_0^T \delta_{Y^\infty(t)}(\cdot) dt.$$ 

The sequence $\{\nu_T\}$ satisfies an LDP as $T \to \infty$ with rate function

$$I^\infty(\nu) = \int_{S^2} q^\infty(x) \nu(dx) - \int_{S^2 \times S^2} \sqrt{\theta(x) \theta(y)} \Gamma^\infty_{x,y} \bar{\mu}(dx),$$

where $\theta = d\nu / d\bar{\mu}$ and $q^\infty$ is the intensity associated with $\Gamma^\infty$.

Let $M : \mathcal{P}(S^2) \to \mathcal{P}(S^2)$ be the mapping for which $M \nu_T = \eta^\infty_T$ (also takes $\bar{\mu}$ to $\mu$). By the contraction principle we retrieve the LDP for $\{\eta^\infty_T\}$.

**Aim:** Use this LDP to investigate properties of infinite swapping algorithms.
The impact of asymmetry: The process $Y^\infty$ moves in a potential landscape with four metastable points.
Turns out that asymmetry in the potential landscape is a hindrance to convergence of parallel tempering and infinite swapping.
Turns out that asymmetry in the potential landscape is a hindrance to convergence of parallel tempering and infinite swapping.

intensity matrix of $Y_1$ is $\approx \Gamma_1$
intensity matrix of $Y_2$ is $\approx \Gamma_2$
intensity matrix of $Y_1$ is $\approx \Gamma_2$
intensity matrix of $Y_2$ is $\approx \Gamma_1$
Turns out that asymmetry in the potential landscape is a hindrance to convergence of parallel tempering and infinite swapping.

\[
\begin{align*}
\text{intensity matrix of } Y_1 & \approx \Gamma_1 \\
\text{intensity matrix of } Y_2 & \approx \Gamma_2 \\
\text{intensity matrix of } Y_1 & \approx \Gamma_2 \\
\text{intensity matrix of } Y_2 & \approx \Gamma_1
\end{align*}
\]

Causes the process to move easily between three out of four stable points - a “secondary metastability” has been introduced.
The effect of this secondary metastability is detected by the large deviation rate function. Consider the optimization problem

$$\inf \{ I^\infty(\nu) : (M \nu)((-\infty, 0] \times S) = \mu((-\infty, 0])(1 - \delta) \}$$
The effect of this secondary metastability is detected by the large deviation rate function. Consider the optimization problem

$$\inf \{ I^\infty(\nu) : (M\nu)((-\infty, 0] \times S) = \mu((-\infty, 0])(1 - \delta) \}$$

**Table:** Optimal rate normalized to the value for $\alpha = 1$ (symmetric potential)

<table>
<thead>
<tr>
<th>$\delta$</th>
<th>0.05</th>
<th>0.10</th>
<th>0.15</th>
<th>0.20</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha$</td>
<td>0.97</td>
<td>0.5605</td>
<td>0.5709</td>
<td>0.5833</td>
</tr>
<tr>
<td>1</td>
<td>0.90</td>
<td>0.1833</td>
<td>0.1898</td>
<td>0.1997</td>
</tr>
<tr>
<td>0.05</td>
<td>0.90</td>
<td>0.1833</td>
<td>0.1898</td>
<td>0.1997</td>
</tr>
<tr>
<td>0.10</td>
<td>0.90</td>
<td>0.1833</td>
<td>0.1898</td>
<td>0.1997</td>
</tr>
<tr>
<td>0.15</td>
<td>0.90</td>
<td>0.1833</td>
<td>0.1898</td>
<td>0.1997</td>
</tr>
<tr>
<td>0.20</td>
<td>0.90</td>
<td>0.1833</td>
<td>0.1898</td>
<td>0.1997</td>
</tr>
</tbody>
</table>
The effect of this secondary metastability is detected by the large deviation rate function. Consider the optimization problem

$$\inf \{ I^\infty (\nu) : (M\nu)((-\infty, 0] \times S) = \mu((-\infty, 0])(1 - \delta) \}$$
The secondary metastability is also illustrated by the swapping of dynamics during a simulation:
Accompanying the process $\mathbf{Y}^\infty$ are the so-called particle-temperature associations:

$$\rho_T = \left( \frac{1}{T} \int_0^T \rho(Y_1^\infty(t), Y_2^\infty(t)) dt, \frac{1}{T} \int_0^T \rho(Y_2^\infty(t), Y_1^\infty(t)) dt \right).$$

Empirical measure on $\Sigma_2 = \{\{1, 2\}, \{2, 1\}\}$ - corresponds to temperature assignments $(\tau_1, \tau_2)$ and $(\tau_2, \tau_1)$.

The convergence

$$\rho_T \rightarrow \left( \frac{1}{2}, \frac{1}{2} \right), \quad T \rightarrow \infty,$$

provides a possible diagnostic for convergence of $\eta^\infty_T$. 
Joint LDP: If $\mu_1$ and $\mu_2$ are the unique invariant distributions of $\Gamma^1$ and $\Gamma^2$, then $\{(\eta^\infty_T, \rho_T)\}$ satisfies an LDP ($T \to \infty$) on $\mathcal{P}(S^2) \times \mathcal{P}(\Sigma_2)$, with rate function

$$I(\gamma, w) = \left\{ I^\infty(\nu) : M\nu = \gamma, \int_{S^2} \rho(x)d\nu(x) = w_1 \right\}.$$ 

Obtained from LDP for $\{\nu_T\}$ via contraction principle.
A first result: Suppose we fix a target measure $\gamma \in \mathcal{P}(S^2)$. Then

$$\inf \{ I^0(\nu) : M\nu = \gamma \}$$

is attained at the symmetric measure

$$\nu_{\text{sym}}(x) = \frac{\gamma(x)}{2\rho(x)}.$$
A first result: Suppose we fix a target measure $\gamma \in \mathcal{P}(S^2)$. Then

$$\inf \{ I^0(\nu) : M\nu = \gamma \}$$

is attained at the symmetric measure

$$\nu_{sym}(x) = \frac{\gamma(x)}{2\rho(x)}.$$

Interpretation: Regardless of the target measure $\gamma$, the most likely measure $\nu$ such that $M\nu = \gamma$ has weights/particle-temperature associations $(1/2, 1/2)$. 
A second result: Let $\mathcal{N}_\epsilon(w)$ denote an open $\epsilon$-neighborhood of $w$ in $\mathcal{P}(\Sigma_2)$ and similarly for $\mathcal{N}_\delta(\mu)$ for $\mu$ in $\mathcal{P}(S^2)$ (weak topologies).
A second result: Let $\mathcal{N}_\epsilon(w)$ denote an open $\epsilon$-neighborhood of $w$ in $\mathcal{P}(\Sigma_2)$ and similarly for $\mathcal{N}_\delta(\mu)$ for $\mu$ in $\mathcal{P}(S^2)$ (weak topologies).

Let $w^* = (1/2, 1/2)$. For each $\epsilon > 0$ there is a $\delta > 0$ such that

$$P(\eta_T^\infty \in \mathcal{N}_\delta(\mu) | \rho_T \in (\mathcal{N}_\epsilon(w^*))^c) \rightarrow 0, \quad T \rightarrow \infty.$$ 

Interpretation: The particle-temperature associations must converge to $(1/2, 1/2)$ if the empirical measure $\eta_T^\infty$ is to converge to the stationary distribution $\mu$.

Proof relies on studying the associated ergodic control problem.
Thank you!

J. Doll, P. Dupuis and P. Nyquist
A large deviation analysis of certain qualitative properties of parallel tempering and infinite swapping algorithms
(On arXiv any day now...)