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

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

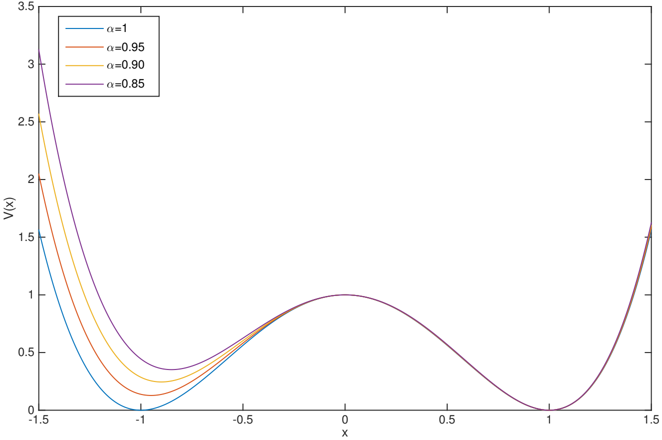
V is the potential of some complicated physical system, λ a reference measure.

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; α determines the level of asymmetry.



For λ Lebesgue measure, can use that μ_1 is the stationary distribution of the solution of

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

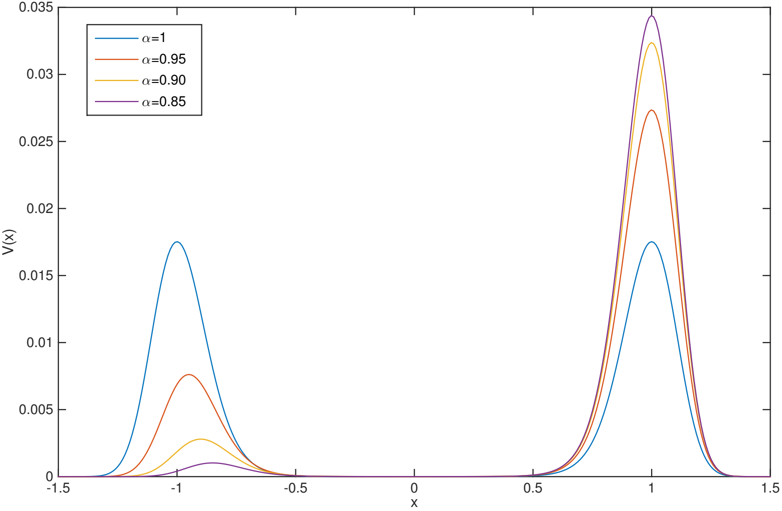
For λ counting measure (finite state space S) can define Glauber dynamics with stationary distribution

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

Consider a continuous-time Markov process $X(t)$ with μ_1 as invariant distribution. Under ergodicity a numerical approximation to μ_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 $\tau_1 = 0.1$.



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 $\Gamma^1_{x,y}$ and $\Gamma^2_{x,y}$ (rate matrices) corresponding to τ_1 and τ_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:

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

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

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

Straightforward to check - e.g., detailed balance - that μ 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:

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

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

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

Straightforward to check - e.g., detailed balance - that μ remains the invariant measure.

How to choose the swap rate a to ensure fast convergence? **Large deviation analysis**.

The empirical measure $\lambda_T^a(\cdot) = \frac{1}{T}\int_0^T \delta_{\mathbf{X}^a(t)}(\cdot)dt$ 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,

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

and

$$J(\nu) = \int_{S^2} g(x_1, x_2) I\left(\sqrt{\frac{\theta(x_2, x_1)}{\theta(x_1, x_2)}}\right)\nu(d\mathbf{x}).$$

The empirical measure $\lambda_T^a(\cdot) = \frac{1}{T}\int_0^T \delta_{\mathbf{X}^a(t)}(\cdot)dt$ 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,

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

and

$$J(\nu) = \int_{S^2} g(x_1,x_2) I\left(\sqrt{\frac{\theta(x_2,x_1)}{\theta(x_1,x_2)}}\right)\nu(d\mathbf{x}).$$

Monotonicity in a suggests letting $a \to \infty$.

Now: switch temperatures/dynamics between the processes, not locations.

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

The weighted empirical measure

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

has the same distribution as the empirical measure of $\mathbf{X}^a$. By ergodicity η_T^a converges to μ as $T \to \infty$.

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

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

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

where

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

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 η_T^a is

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

Ergodicity $\Rightarrow \eta_T^\infty \to \mu$ as $T \to \infty$.

The limit of the weighted empirical measure η_T^a is

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

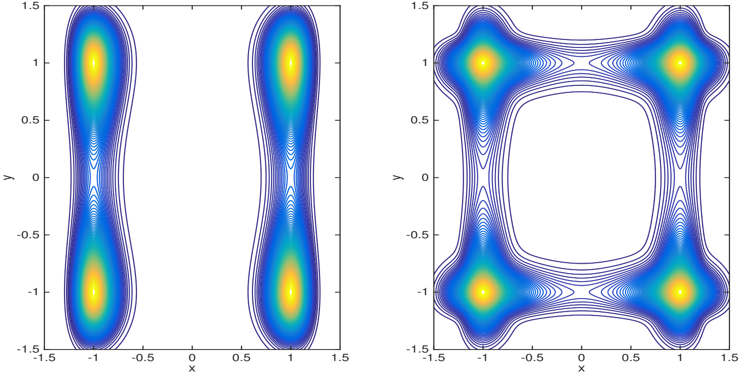
Ergodicity $\Rightarrow \eta_T^\infty \to \mu$ as $T \to \infty$.

Infinite swapping: Simulate $\mathbf{Y}^\infty$ and use η_T^∞ for numerical approximations of μ.

The limit process $\mathbf{Y}^\infty$ has invariant measure

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

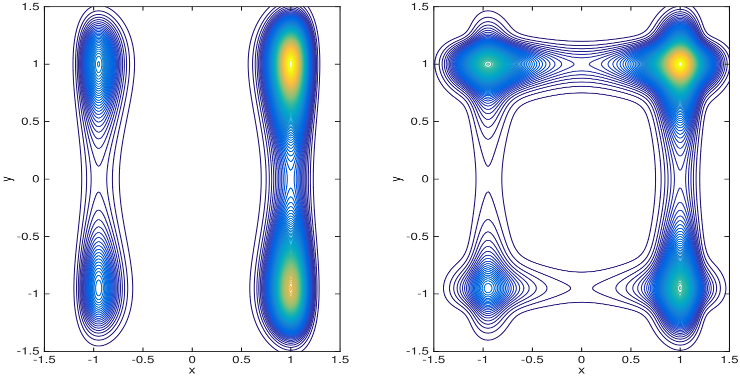
Connectedness of the density much improved compared to PT.



The limit process $\mathbf{Y}^{\infty}$ has invariant measure

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

Connectedness of the density much improved compared to PT.



Large deviations: Let ν_T be the empirical measure associated with $\mathbf{Y}^\infty$,

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

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

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

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

Large deviations: Let ν_T be the empirical measure associated with $\mathbf{Y}^{\infty}$,

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

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

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

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

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 μ). By the contraction principle we retrieve the LDP for $\{\eta^{\infty}_T\}$.

Large deviations: Let ν_T be the empirical measure associated with $\mathbf{Y}^\infty$,

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

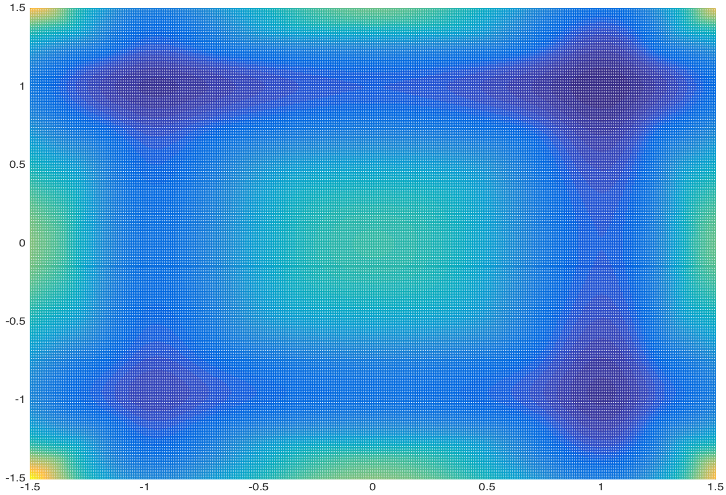
The sequence $\{\nu_T\}$ satisfies an LDP as $T \to \infty$ with rate function

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

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

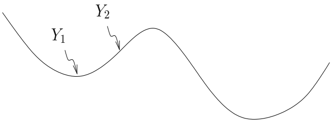
Let $M : \mathcal{P}(S^2) \to \mathcal{P}(S^2)$ be the mapping for which $M\nu_T = \eta^\infty_T$ (also takes $\bar{\mu}$ to μ). By the contraction principle we retrieve the LDP for $\{\eta^\infty_T\}$.
Aim: Use this LDP to investigate properties of infinite swapping algorithms.

The impact of asymmetry: The process $\mathbf{Y}^{\infty}$ moves in a potential landscape with four metastable points.



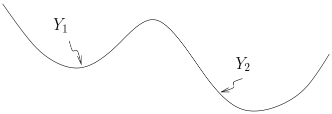
Turns out that asymmetry in the potential landscape is a hindrance to convergence of parallel tempering and infinite swapping.

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 $\approx \Gamma_1$

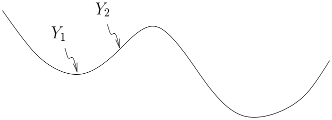
intensity matrix of Y_2 is $\approx \Gamma_2$



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.



intensity matrix of Y_1 is $\approx \Gamma_1$ | intensity matrix of Y_1 is $\approx \Gamma_2$

intensity matrix of Y_2 is $\approx \Gamma_2$ | intensity matrix of Y_2 is $\approx \Gamma_1$

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

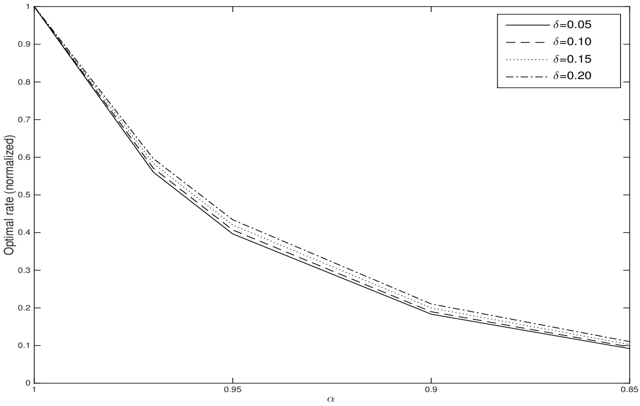
$$\inf\{I^{\infty}(\nu): (M\nu)((-\infty,0]\times S) = \mu((-\infty,0])(1-\delta)\}$$

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

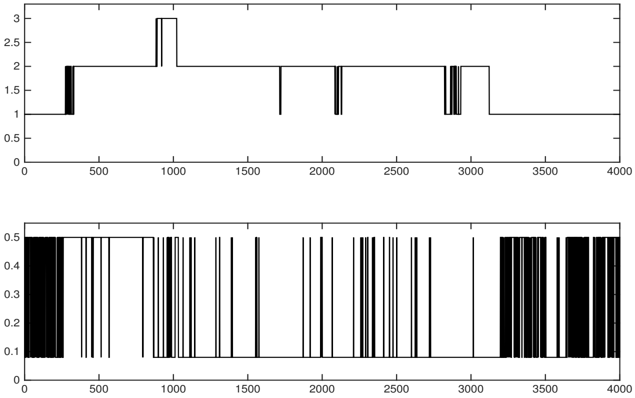
	α				
δ	1	0.97	0.95	0.90	0.85
0.05	1	0.5605	0.3965	0.1833	0.09188
0.10	1	0.5709	0.4070	0.1898	0.09634
0.15	1	0.5833	0.4200	0.1997	0.1026
0.20	1	0.5964	0.4341	0.2105	0.1103

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 (τ_1, τ_2) and (τ_2, τ_1).

The convergence

$$\rho_T \to \left(\frac{1}{2}, \frac{1}{2} \right), \; T \to \infty,$$

provides a possible diagnostic for convergence of η_T^∞.

Joint LDP: If μ_1 and μ_2 are the unique invariant distributions of Γ^1 and Γ^2, then $\{(\eta_T^{\infty}, \rho_T)\}$ satisfies an LDP ($T \to \infty$) on $\mathcal{P}(S^2) \times \mathcal{P}(\Sigma_2)$, with rate function

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

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

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

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

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

Interpretation: Regardless of the target measure γ, the most likely measure ν 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 ϵ-neighborhood of w in $\mathcal{P}(\Sigma_2)$ and similarly for $\mathcal{N}_{\delta}(\mu)$ for μ in $\mathcal{P}(S^2)$ (weak topologies).

A second result: Let $\mathcal{N}_{\epsilon}(w)$ denote an open ϵ-neighborhood of w in $\mathcal{P}(\Sigma_2)$ and similarly for $\mathcal{N}_{\delta}(\mu)$ for μ in $\mathcal{P}(S^2)$ (weak topologies).

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

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

Interpretation: The particle-temperature associations must converge to $(1/2, 1/2)$ if the empirical measure η_T^{∞} is to converge to the stationary distribution μ.

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...)