Parrellel Tempering

theory and applications

Mateusz Łącki

Uniwersytet Warszawski

24 April 2013

Today's Agenda

1. What? Why? Who?
2. Metropolis-Hasting algorithm
3. Metropolis-Hasting localness
   - Notation
   - Ideology
   - random walk kernel
   - swap kernel
4. Theory of Parallel Tempering
5. Swapping Strategies
6. Bibliography

What is Parallel Tempering? Nutshell view.

- Stochastic simulation algorithm
- a.k.a. replica exchange Monte Carlo
  - sampling method
- Extension to Metropolis-Hastings algorithm ...
- ...or rather Metropolis-Hastings-Green algorithm
  - more abstract version: general kernels
  - more freedom: discrete kernels, continous kernels, different dimensions

Why Parallel Tempering?

- Tool to fight *multimodality*
  - "allows good mixing with multimodial target distributions"
  - response to Metropolis-Hastings localness
  - better estimation of integrals $\int_{\mathbb{R}^d} f(x)\pi(x)\mathrm{d}x$
- In certain physical models: thermodynamic interpretation
  - Gibbs random-field model

Who might be interested in parallel tempering?

- anyone having problems with locality of simulations
- researchers facing model selection problems while exploring the posteriori distribution over some space of models: the g-priors

The usual Metropolis-Hastings algorithm revised

We are given a measure π with density $h : \mathbb{R}^d \mapsto \mathbb{R}_+$.

As. h need not be normalised

$$\int_{\mathbb{R}^d} h(x)\mathrm{d}\,x \in (0, \infty)$$

As. We can evaluate $h(x)$ for all $x \in \mathbb{R}^d$

As. Transitional probability density $q : \underbrace{\mathbb{R}^d}_{currentstate} \times \overbrace{\mathbb{R}^d}^{proposal} \mapsto \mathbb{R}_+$

The usual Metropolis-Hastings algorithm revised

As. For all x, $q(x, \circ)$ is a normalised probability density

- $\underset{x}{\forall}$ we can simulate $y \sim q(x, \circ)$
- $\underset{x,y}{\forall}$ we can evaluate $q(x, y)$

then the Markov chain $X \equiv \{X^{[k]}\}_{k \geq 0}$ generated by procedure

1. $x = X^{[k-1]}$
2. $y \sim q(x, \circ)$
3. evaluate $R(x, y) = \frac{h(y)q(y,x)}{h(x)q(x,y)}$
4. reject y with probability $\alpha(x, y) = 1 \wedge R(x, y)$
   - If rejected $X^{[k]} = x$
   - Otherwise $X^{[k]} = y$

is reversible: its kernel preserves π.

What's a kernel?

A regular version of $\mathbb{E}\Big(\mathbb{I}_A | X = x\Big)$

- measurable function with x for A fixed
- probability distribution with A for any fixed x (stronger than almost everywhere)

in standard MH

$$P(x, A) = \int_A p(x, y) \mathrm{d}\, y + \Big(1 - \int_{\mathbb{R}^d} p(x, y) \mathrm{d}\, y\Big) \mathcal{I}(x, A)$$

where $p(x, y) = \alpha(x, y) q(x, y)$

and $\mathcal{I}(x, A) = \mathbb{I}_A(x)$ - identity kernel

What's reversibility?

An integral equation

$$\int_A \pi(\mathrm{d}\,x) P(x, B) = \int_B \pi(\mathrm{d}\,x) P(x, A)$$

This assures that the chain preserves π

$$\int_{\mathbb{R}^d} \pi(\mathrm{d}\,x) P(x, B) = \int_B \pi(\mathrm{d}\,x) = \pi(B)$$

or $\pi P = \pi$

That's Functional Analysis

Reversibility is P self-adjointess:

$$\underset{f,g\in\mathbb{L}^2(\pi)}{\forall} \int \pi(\mathrm{d}\,x)P(x,\mathrm{d}\,y)f(x)g(y) = \int \pi(\mathrm{d}\,x)P(x,\mathrm{d}\,y)g(x)f(y)$$

the inner product convention

$$\langle\, Pf \,|\, h \,\rangle = \langle f \,|\, Ph \,\rangle$$

standard spectral analysis tools applicable

Metropolis-Hastings-Green algorithm

$(S, \mathfrak{G})$ - measurable space

As. Take a positive measure

$$\rho : \mathfrak{G} \mapsto \mathbb{R}_+$$

not necessarily probabilistic

$$\rho(S) \in (0, \infty)$$

As. $\rho = \rho(S)\pi$

proposal kernel $Q(x, A)$

As. we know how to generate $y \sim Q(x, \circ)$

As. For all $x, y \in S$ the Hastings ratio

$$R(x, y_k) \equiv \frac{\rho(\mathrm{d}y)\,Q(y, \mathrm{d}x)}{\rho(\mathrm{d}x)\,Q(x, \mathrm{d}y)}$$

known ans possible to evaluate for any x and y

Chain generation

Given $x = X^{[k-1]}$

1. Simulate $y \sim Q(x, \circ)$.
2. Calculate $R(x, y)$.
3. Accept y with probability $\alpha(x, y) = 1 \wedge R(x, y)$.

Green's result

Then by the Green theorem X's kernel

$$P(x, A) \equiv \int_A \alpha(x, y) Q(x, \mathrm{d}y) + \delta_x(A)\left(1 - \int_\Omega \alpha(x, y) Q(x, \mathrm{d}y)\right)$$

is reversibile with respect to π

Note: kernel is stochastic.

Problems with MH

To assert proposal's adequacy

generate $U \sim \mathfrak{U}(0,1)$
accept y if $U \leq \frac{h(y)q(y,x)}{h(x)q(x,y)} \wedge 1$
if $q(x,y) = q(y,x)$

$$U \leq \frac{h(y)}{h(x)} \wedge 1$$

Problem: if π is multimodial we get stuck in certain region of the state-space

Liang and Wong 2001

$$f(x) = \sum_{i=1}^{20} \frac{\omega_i}{\sigma_i\sqrt{2\pi}} \exp\left(- \frac{(x-\mu_i)'(x-\mu_i)}{2\sigma_i^2} \right)$$

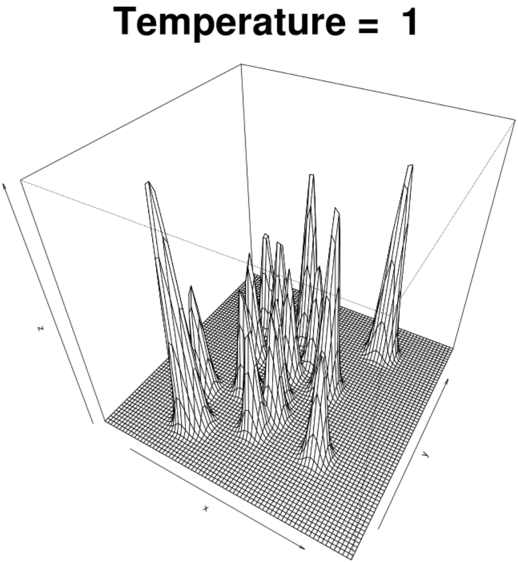
where $\sigma_1 = \cdots = \sigma_{20} = 0.1, \omega_1 = \cdots = \omega_{20} = 0.05$

and the means μ_i are given by

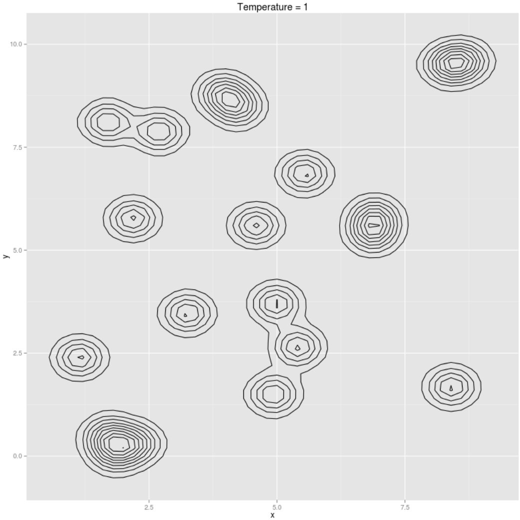
1	2	3	4	5	6	7	8	9	10
2.18	8.67	4.24	8.41	3.93	3.25	1.70	4.59	6.91	6.87
5.76	9.59	8.48	1.68	8.82	3.47	0.50	5.60	5.81	5.40

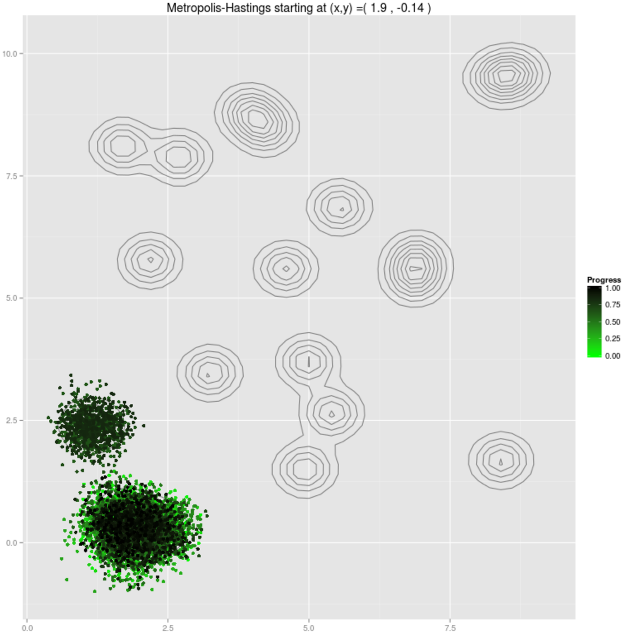
11	12	13	14	15	16	17	18	19	20
5.41	2.70	4.98	1.14	8.33	4.93	1.83	2.26	5.54	1.69
2.65	7.88	3.70	2.39	9.50	1.50	0.09	0.31	6.86	8.11

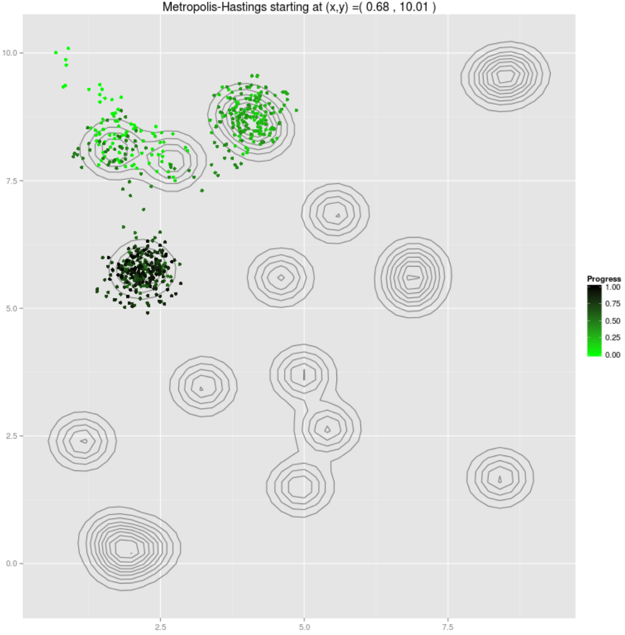
TOY-EXAMPLE VISUALISED



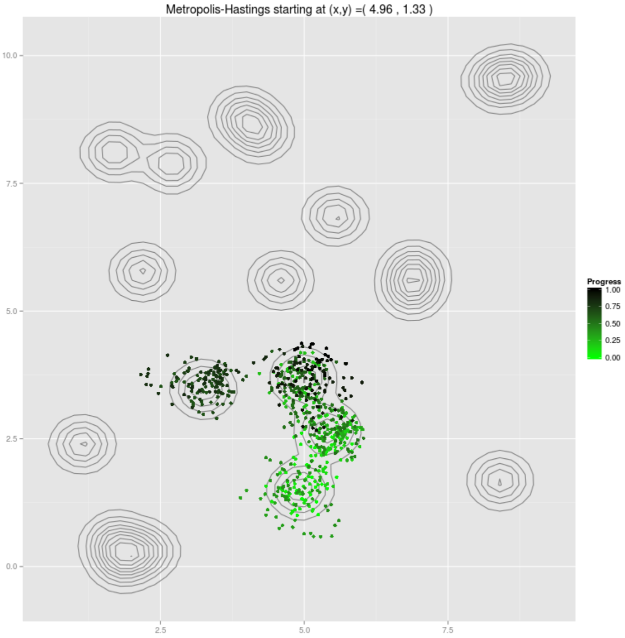
TOY-EXAMPLE VISUALISED AS A CONTOUR PLOT



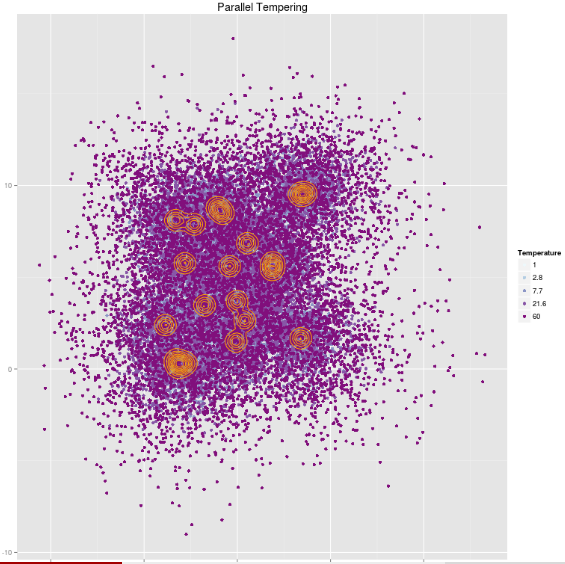

Metropolis-Hastings starting at (x,y) =(1.9 , -0.14)
10.0
7.5
5.0
2.5
0.0
0.0
2.5
5.0
7.5
Progress
1.00
0.75
0.50
0.25
0.00



Metropolis-Hastings starting at (x,y) =(0.68 , 10.01)
10.0
7.5
5.0
2.5
0.0
2.5
5.0
7.5
Progress
1.00
0.75
0.50
0.25
0.00


Metropolis-Hastings starting at (x,y) =(4.96 , 1.33)
10.0
7.5
5.0
2.5
0.0
2.5
5.0
7.5
Progress
1.00
0.75
0.50
0.25
0.00

Parallel Tempering in action.



Getting inside Parallel Tempering

$(\Omega, \mathfrak{F})$	measurable space
Ω	subset of a Polish space
$\mathfrak{F}$	Borel subsets countably generated
$\mathcal{I} = [0, 1]$	unit interval
$\pi : \mathfrak{F} \mapsto \mathcal{I}$	measure

Assumptions and further notation

As. π has density w.r. to Lebesgue measure

- π - the density
  - know up to its proportionality factor
  - unnormalised

$$\int_{\Omega} \pi(x) \mathrm{d}\, x \in (0, \infty)$$

Solution's ideology

Space for chains

$$(\Omega^L, \mathfrak{F}^{\otimes L}, \pi_\beta)$$

where $\mathfrak{F}^{\otimes L} \equiv \underbrace{\mathfrak{F} \otimes \cdots \otimes \mathfrak{F}}_{L \text{ times}}$

$$\pi_\beta \propto \pi^{\beta_1} \times \cdots \times \pi^{\beta_L}$$

$\beta = (\beta_1, \ldots, \beta_L)$ - inverse temperatures $\beta_i = T_i^{-1}$
and $1 = T_1 < \cdots < T_L < \infty$

no normalisation of π_β coordinates

...enters Markov

Markov Chain $X \equiv \{X^{[k]}\}_{k\geq 0}$

- Ω^L state-space for X

$$X^{[k]} = (X_1^{[k]}, \ldots, X_L^{[k]})$$

High temperature

Low temperatures coordinates

NON-SWEDISH WAY TO EQUIDISTRIBUTION

As. $\pi > 0$ somewhere between modes

- then $\pi^{\beta_k} > \pi$ for $k \geq 2$

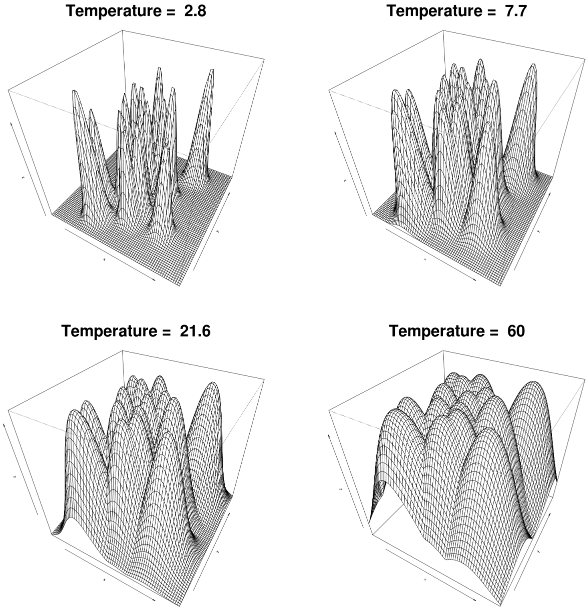
So that if $\pi(y) < \pi(x)$ then

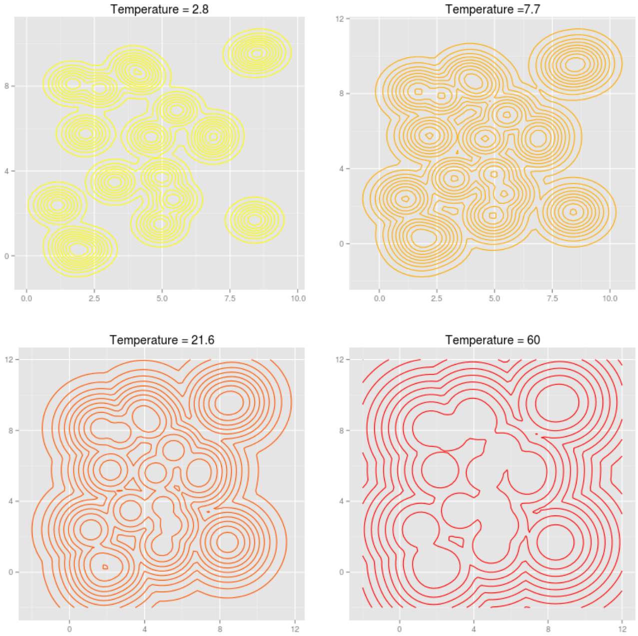
$$\alpha_{\beta_1}(x, y) = 1 \wedge \frac{\pi(y)}{\pi(x)} = \frac{\pi(y)}{\pi(x)} < \left(\frac{\pi(y)}{\pi(x)}\right)^{\beta_k} = 1 \wedge (\frac{\pi(y)}{\pi(x)})^{\beta_k} = \alpha_{\beta_k}(x, y)$$

- proposal accepted more often in higher temperatures when

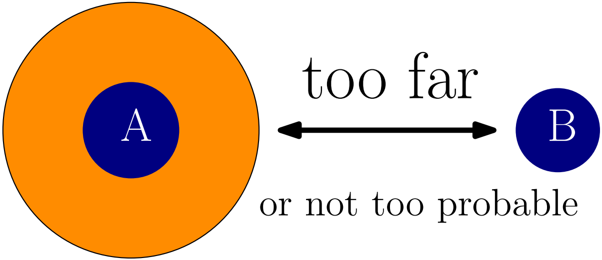
$$\frac{\pi(y)}{\pi(x)} < 1$$

- we enlarge regions from which proposals get accepted


Temperature = 2.8
Temperature = 7.7
Temperature = 21.6
Temperature = 60


Temperature = 2.8
0.0
2.5
5.0
7.5
10.0
0
4
8
Temperature =7.7
0.0
2.5
5.0
7.5
10.0
0
4
8
12
Temperature = 21.6
0
4
8
12
Temperature = 60
0
4
8
12

Potential Problems



Potential Problems

- π might be ok
- because of computer's finite arithmetic it is not ok anymore

Theoretical Details

algorithm reached n^{th} step - $X^{[n]}$

we act with two kernels

$$X^{[n]} \overset{\mathcal{S}_\beta}{\to} \widetilde{X}^{[n+1]} \overset{M_{\Sigma,\beta}}{\to} X^{[n+1]}.$$

- $\mathcal{S}_\beta$ - swap kernel
- $M_{\Sigma,\beta}$ - random walk kernel

- Their reversibility is assured by the MHG algorithm reversibility.

  so we get π-preservation

$$\mathcal{S}_\beta M_{\Sigma,\beta}\pi = \mathcal{S}_\beta \pi = \pi$$

RANDOM WALK $M_{\Sigma,\beta}$

take $A_i \in \mathfrak{F}$ and $x \in \Omega^L$

then

$$M_{\Sigma,\beta}(x, A_1 \times \cdots \times A_L) = \prod_{l=1}^{L} M_{\Sigma_l,\beta_l}(x_l, A_l)$$

where $M_{\Sigma_l,\beta_l}(x_l, A_l)$ is equal to

$$\int_A \alpha_{\beta_l}(x_l, y_l) q_{\Sigma_l}(y_l - x_l) \mathrm{d}y_l + \delta_x(A) \int [1 - \alpha_{\beta_l}(x_l, y_l)] q_{\Sigma_l}(y_l - x_l) \mathrm{d}y_l$$

where α_{β_l} is the acceptance level

and q_{Σ_l} is proposal distribution

Observations and assumptions

As. q_{Σ_l} - density of $\mathcal{N}(0, \Sigma_l)$

Symmetry $q(x_l, y_l) = q(y_l, x_l)$ implies

$$\alpha_{\beta_l}(x_l, y_l) \equiv 1 \wedge \frac{\pi^{\beta_l}(y_l)}{\pi^{\beta_l}(x_l)}$$

for the chain to preserve π_β.

Implementation: independent simulation of M_{Σ_l, β_l} for each $\widetilde{X}_l^{[n-1]}$

Interlacing independent chains with $\mathcal{S}_\beta$

Swaps

- Less-tempered chains placed in unusual places
- a pair of coordinates $(\widetilde{X}_l^{[n-1]}, \widetilde{X}_k^{[n-1]})$ drawn at random

As. only one pair per turn

- different strategies possible

Introduce swap operation

$$S_{ij}x = (x_1, \dots, x_{i-1}, x_j, x_{i+1}, \dots, x_{j-1}, x_i, x_{j+1}, \dots, x_L)$$

Precise kernel form of $\mathcal{S}_\beta$

For any $x \in \Omega^L$ and $A \in \mathfrak{F}^{\otimes L}$

$$\mathcal{S}_\beta(x, A) \equiv$$

$$\sum_{i<j} p_{ij}(x)\alpha_{\text{swap}}(x, S_{ij}x)\mathbb{I}_A(S_{ij}x) + \Big(1 - \sum_{i<j} p_{ij}(x)\alpha_{\text{swap}}(x, S_{ij}x)\Big)\mathcal{I}(x, A)$$

where

$$\alpha_{\text{swap}}(x, S_{ij}x) = \left[\left(\frac{\pi(x_j)}{\pi(x_i)}\right)^{\beta_i - \beta_j} \frac{p_{ij}(S_{ij}x)}{p_{ij}(x)}\right] \wedge 1$$

is the acceptance level for swaps,

$p_{ij}(x)$ - probability function of a swap given x.

Nota bene: given state x, $\mathcal{S}_\beta$ has a finite support

$$\mathfrak{S}_x \equiv \{S_{ij}x : i < j\}$$

Different possible swapping strategies $p_{ij}(x)$

Strategy I

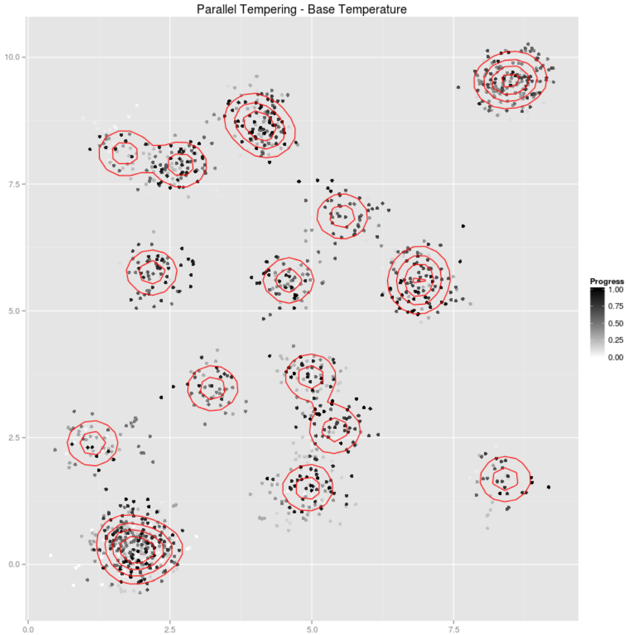
$$p_{ij}(x) \propto \frac{\pi(x_j)}{\pi(x_i)} \wedge \frac{\pi(x_i)}{\pi(x_j)} = \exp\Big(- \,|\log(\pi(x_j)) - \log(\pi(x_i))|\Big)$$

- promotes swaps between coordinates relatively the same, $\pi(x_j) \approx \pi(x_i)$

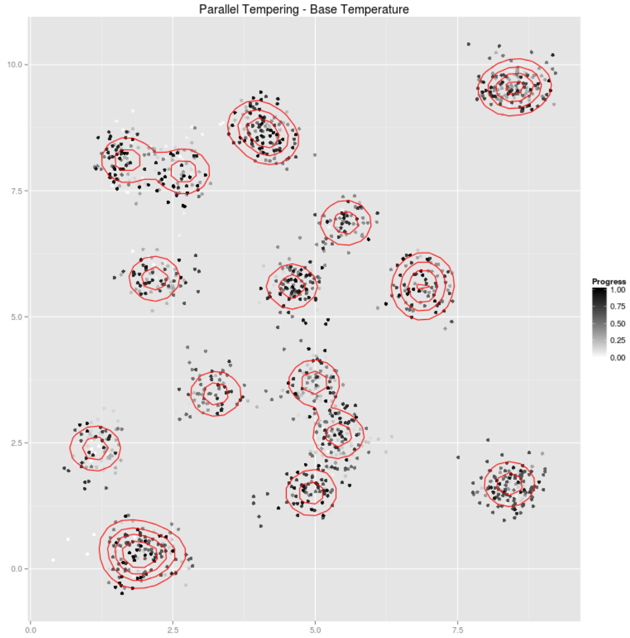
Strategy II

$$p_{ij}(x) \propto \frac{\pi(x_j)}{\pi(x_i)} \wedge 1 = \exp\Big(- (\log(\pi(x_j)) - \log(\pi(x_j)))\Big) \wedge 1$$

- breaks the symmetry of the previous one



Parallel Tempering - Base Temperature
10.0
7.5
5.0
2.5
0.0
0.0
2.5
5.0
7.5
Progress
1.00
0.75
0.50
0.25
0.00


Parallel Tempering - Base Temperature
10.0
7.5
5.0
2.5
0.0
0.0
2.5
5.0
7.5
Progress
1.00
0.75
0.50
0.25
0.00

Different possible swapping strategies $p_{ij}(x)$

Strategy III

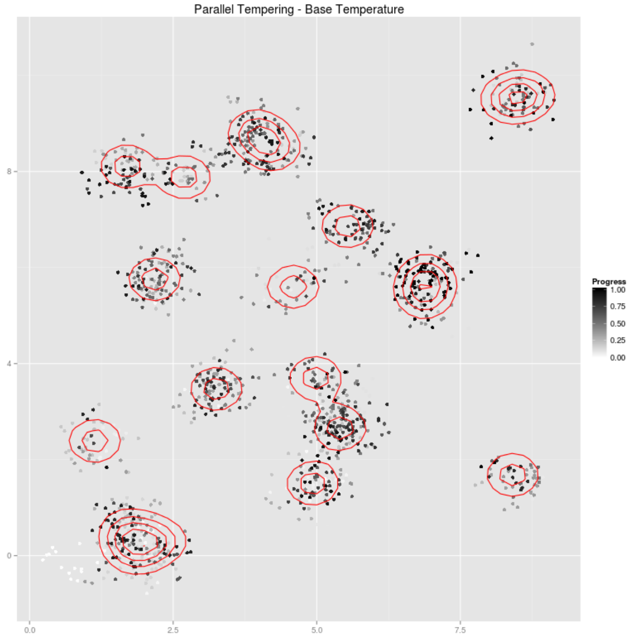
$$p_{ij} \propto \left(\frac{\pi(x_j)}{\pi(x_i)} \wedge \frac{\pi(x_i)}{\pi(x_j)}\right)^{\beta_i-\beta_j} = \exp\Big(-(\beta_i-\beta_j)|\log(\pi(x_j))-\log(\pi(x_i))|\Big)$$

- softens the requirement $\pi(x_j) \approx \pi(x_i)$
- promotes $\beta_i - \beta_j \approx 0$
- promotes swaps between adjacent chains

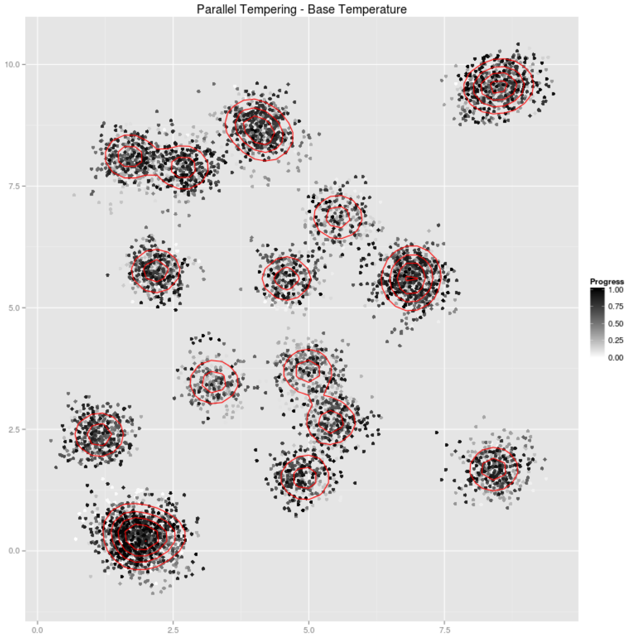
Strategy IV

$$p_{ij} \propto \left(\frac{\pi(x_j)}{\pi(x_i)} \wedge \frac{\pi(x_i)}{\pi(x_j)}\right)^{\frac{\beta_i-\beta_j}{1+\rho(x_i,x_j)}} = \exp\Big(-\frac{(\beta_i-\beta_j)|\log(\pi(x_j))-\log(\pi(x_i))|}{1+\rho(x_i,x_j)}$$

- added a quasi-metric
- ρ does not require symmetry $\rho(x_i, x_j) = \rho(x_j, x_i)$
- could be of use in the Gibbs random-field model


Parallel Tempering - Base Temperature
8
4
0
0.0
2.5
5.0
7.5
Progress
1.00
0.75
0.50
0.25
0.00

Parallel Tempering - Base Temperature
10.0
7.5
5.0
2.5
0.0
2.5
5.0
7.5
Progress
1.00
0.75
0.50
0.25
0.00



Parallel Tempering - Base Temperature
10.0
7.5
5.0
2.5
0.0
0.0
2.5
5.0
7.5
Progress
1.00
0.75
0.50
0.25
0.00

Bibliography

- Błażej Miasojedow, Eric Moulines, Matti Vihola, *Adaptive Parallel Tempering Algorithm*, Arxive.
- Meïli Baragatti, Agnès Grimaud, Denys Pommeret *Parallel Tempering with Equi-Energy Moves.*
- Charles J. Geyer, *Markov Chain Monte Carlo Lecture Notes.*