Monte Carlo methods for sampling-based Stochastic Optimization

Gersende FORT

LTCI
CNRS & Telecom ParisTech
Paris, France

Joint works with

B. Jourdain, T. Lelièvre, G. Stoltz from ENPC and E. Kuhn from INRA.

A. Schreck and E. Moulines from Telecom ParisTech

P. Priouret from Paris VI

Simulated Annealing (1/2)

- Let U denote the objective function one wants to minimize.

$$\min_{x \in \mathbb{X}} U(x) \Longleftrightarrow \max_{x \in \mathbb{X}} \exp(-U(x)) \Longleftrightarrow \max_{x \in \mathbb{X}} \exp\left(-\frac{U(x)}{T}\right) \quad \forall T > 0$$

- In order to sample from $\pi_{T_\star}$ where

$$\pi_T(x) = \exp\left(-\frac{U(x)}{T}\right)$$

sample successively from a sequence of tempered distributions π_{T_1}, π_{T_2}, $\cdots$ with $T_1 > T_2 > \cdots > T_\star$.

Simulated Annealing (1/2)

- Let U denote the objective function one wants to minimize.

$$\min_{x\in\mathbb{X}} U(x) \Longleftrightarrow \max_{x\in\mathbb{X}} \exp(-U(x)) \Longleftrightarrow \max_{x\in\mathbb{X}} \exp\left(-\frac{U(x)}{T}\right) \quad \forall T > 0$$

- In order to sample from $\pi_{T_\star}$ where

$$\pi_T(x) = \exp\left(-\frac{U(x)}{T}\right)$$

sample successively from a sequence of tempered distributions π_{T_1}, π_{T_2}, $\cdots$ with $T_1 > T_2 > \cdots > T_\star$.

or sample successively from a sequence of (n_t-iterated) kernels $(P_{T_t}(x,\cdot))_t$ such that $\pi_T P_T = \pi_T$.

Simulated Annealing (2/2)

- Under conditions on $\mathbb{X}$, on the cooling schedule $(T_t)_t$, on the kernels $(P_t)_t$, on the dominating measure and the set of minima, $\cdots$

Kirkpatrick, Gelatt and Vecchi. Optimization via Simulated Annealing. Science (1983)
Geman and Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Trans. on PAMI. (1984).
Van Laarhoven and Aarts, Simulated Annealing : theory and applications. Mathematics and its Applications, Reidel, Dordrecht (1987).
Chiang and Chow. On the convergence rate of annealing processes. SIAM J. Control Optim. (1988)
Hajek, B. Cooling schedule for optimal annealing. Math. Operat. Res. (1988).
Haario and Saksman. Simulated Annealing process in general state space. Adv. Appl. Probab. (1991)

X_t converges to the minima of U

Sampling a density

Monte Carlo methods are numerical tools to solve some computational problems

- in bayesian statistics, for the exploration of the a posteriori distribution π
- computation of integrals (w.r.t. π)
- stochastic optimization (of U, $\pi \propto \exp(U)$)
- ...

Monte Carlo methods draw points $(X_t)_t$ approximating π

$$\pi \approx \frac{1}{T} \sum_{t=1}^{T} \delta_{X_t}$$

even in difficult situations when perfect sampling under π is not possible

- π known up to a normalization constant
- complex expression of π if explicit
- large dimension of the state space
- ...

Two main strategies : Importance Sampling & MCMC (1/2)

1. Importance Sampling :

- Choose an auxiliary distribution $\pi_\star$
- Draw points approximating $\pi_\star$
- Reweight these draws to approximate π

Ex. $(X_t)_t$ i.i.d. under $\pi_\star$,

$$\pi \approx \frac{1}{T} \sum_{t=1}^{T} \frac{\pi(X_t)}{\pi_\star(X_t)} \delta_{X_t}$$

Main drawback in large dimension :

- not robust at all when the dimension is large : degeneracy of the weights, large and even infinite variance if $\pi_\star$ is not selected in accordance with π. MCMC far more robust to the dimension

Two main strategies : Importance Sampling & MCMC (2/2)

2. Markov Chain Monte Carlo (MCMC) : Sample a Markov chain, with unique invariant distribution π

Ex. Hastings-Metropolis type algorithms :

- Choose an auxiliary transition kernel $q(x, y)$
- Starting from the current point X_t, propose a candidate $Y \sim q(X_t, \cdot)$
- Accept or Reject the candidate

$$X_{t+1} = \begin{cases} Y & \text{with probability } \alpha(X_t, Y) \\ X_t & \text{with probability } 1 - \alpha(X_t, Y) \end{cases}$$

where

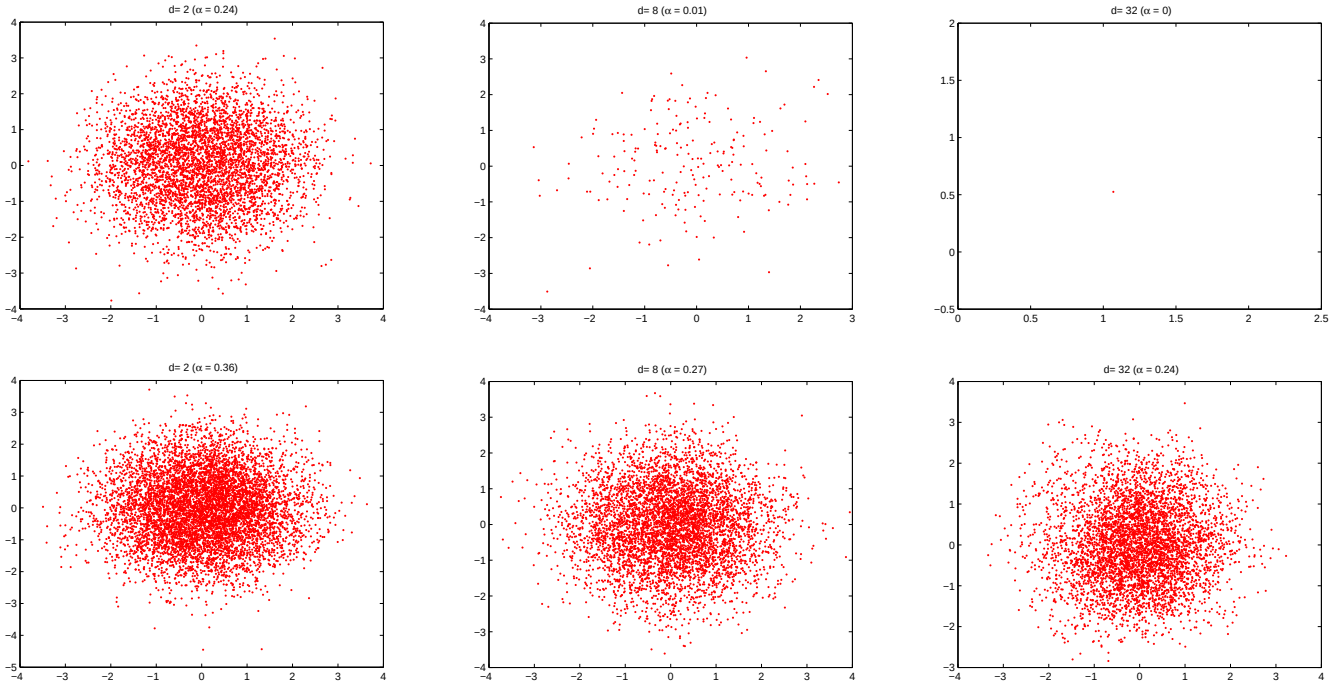
$$\alpha(x, y) = 1 \wedge \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}.$$

Main drawback of classical MCMC samplers for multimodal densities on large dimensional space

- have to scale the size of the proposed moves as a function of the dimension
- remain trapped in some modes, unable to jump and visit the sampling space in a "correct" time.

Example 1

Ex. MCMC - Size of the proposed moves w.r.t. the dimension.



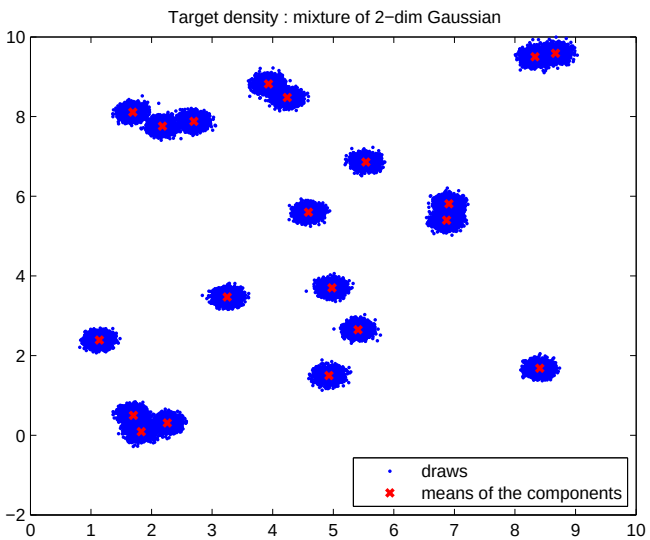
Plot of the first two components of the chain in $\mathbb{R}^d$ *with target* $\pi = \mathcal{N}_d(0, I)$ *for* $d \in \{2, 8, 32\}$ *: the candidate is* $Y = X_t + \mathcal{N}_d(0, \sigma^2 I)$*.* σ *does not depend on* d *(top) and* σ *is of the form* $c/\sqrt{d}$ *(bottom)*

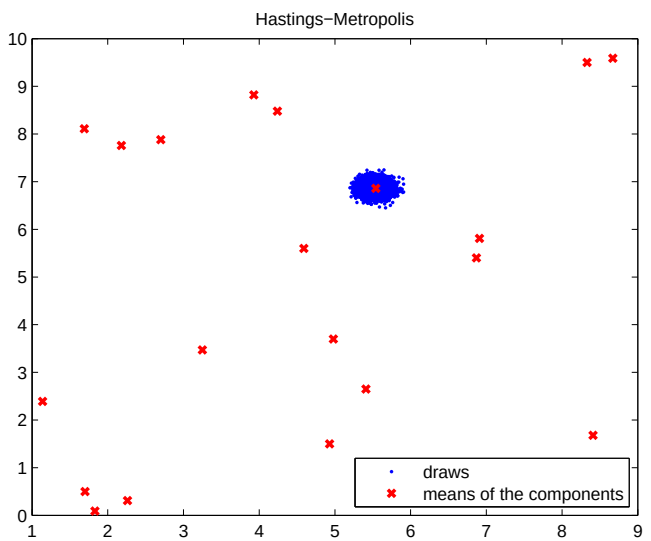
Example 2

The target density π is a mixture of Gaussian in $\mathbb{R}^2$

$$\pi \propto \sum_{i=1}^{20} \mathcal{N}_2(\mu_i, \Sigma_i)$$

We compare N i.i.d. points (left) to N points from a Hastings-Metropolis chain (right)





Classical adaptive MCMC are not robust to the multimodality problem

How to tackle the multimodality question ?

Here are some directions recently proposed in the Statistic literature :

1. Biasing potential approach

- Identify (few) "directions of metastability" $\xi(x)$ and a biasing potential $A(\xi(x))$ such that $\pi_\star(x) \propto \pi(x)\ \exp(-A(\xi(x)))$ has better mixing properties
- Sample under $\pi_\star$ and add a reweighting mecanism to approximate π.

Ex. the Wang-Landau sampler

How to tackle the multimodality question ?

Here are some directions recently proposed in the Statistic literature :

1. Biasing potential approach

- Identify (few) "directions of metastability" $\xi(x)$ and a biasing potential $A(\xi(x))$ such that $\pi_\star(x) \propto \pi(x)\ \exp(-A(\xi(x)))$ has better mixing properties
- Sample under $\pi_\star$ and add a reweighting mecanism to approximate π.

Ex. the Wang-Landau sampler

2. Tempering methods and Interactions

- Choose a set of inverse temperature $0 < \beta_1 < \cdots < \beta_{K-1} < 1$
- Sample points approximating the tempered densities π^{β_i} by allowing interactions between these points.

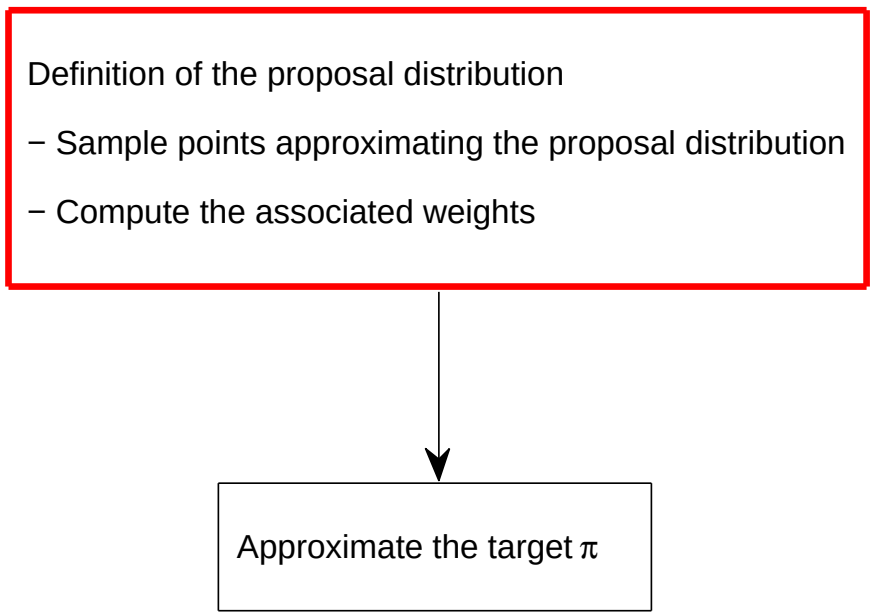
Ex. the Equi-Energy sampler

Outline

The Wang-Landau algorithm

The algorithm was proposed by Wang and Landau in 2001, in the molecular dynamics field.

F.G. Wang and D.P. Landau, Determining the density of states for classical statistical models : A random walk algorithm to produce a flat histogram, Phys. Rev. E 64 (2001).
G. Fort, B. Jourdain, E. Kuhn, T. Lelièvre and G. Stoltz. Convergence of the Wang-Landau algorithm. Accepted for publication in Mathematics of Computation, March 2014.
G. Fort, B. Jourdain, E. Kuhn, T. Lelièvre and G. Stoltz. Efficiency of the Wang-Landau algorithm. Accepted for publication in Applied Mathematics Research Express, February 2014.
L. Bornn, P. Jacob, P. Del Moral and A. Doucet. An Adaptive Wang-Landau Algorithm for Automatic Density Exploration. Journal of Computational and Graphical Statistics (2013).
P. Jacob and R. Ryder. The Wang-Landau algorithm reaches the flat histogram criterion in finite time. Ann. Appl. Probab. (2013).



The proposal distribution (1/3)

- Wang-Landau is an importance sampling algorithm with proposal $\pi_\star$

$$\pi_\star(x) \propto \sum_{i=1}^{d} \frac{\pi(x)}{\pi(\mathbb{X}_i)} \mathbb{1}_{\mathbb{X}_i}(x)$$

  where $\mathbb{X}_1, \cdots, \mathbb{X}_d$ is a partition of the sampling space.

- The proposal distribution $\pi_\star$ consists in reweighting locally π so that

$$\forall i, \qquad \pi_\star(\mathbb{X}_i) = \frac{1}{d}$$

$\hookrightarrow$ This last property will force the sampler $\pi_\star$ to visit all the strata, with the same frequency.

The proposal distribution (2/3)

Unfortunately, $\pi(\mathbb{X}_i)$ is unknown and we can not sample from $\pi_\star$ (even with MCMC)

- The algorithm will use a family of biased distributions

$$\pi_\theta(x) \propto \sum_{i=1}^{d} \frac{\pi(x)}{\theta(i)} \mathbb{1}_{\mathbb{X}_i}(x)$$

where $\theta = (\theta(1), \cdots, \theta(d))$ is a weight vector.

- Key property : $\pi_\star$ is among this family

$$\pi_\star(x) = \pi_{\theta_\star}(x) \qquad \text{with } \theta_\star = \left(\frac{\pi(\mathbb{X}_1)}{Z_\pi}, \cdots, \frac{\pi(\mathbb{X}_d)}{Z_\pi} \right)$$

The algorithm will simultaneously (a) learn the target weight $\theta_\star$ and (b) produce points approximating $\pi_\star$.

The proposal distribution (3/3)

The algorithm (step 1) Given the current biasing weight θ_t and the current sample X_t

- sample the new point :

$$X_{t+1} \sim P_{\theta_t}(X_t, \cdot)$$

where P_θ is a Markov kernel s.t. $\pi_\theta P_\theta = \pi_\theta$

- Update the biasing weight : if $X_{t+1} \in \mathbb{X}_i$, penalize the stratum i in order to favor the visits to the other stratum. Since $\pi_\theta(x) \propto \pi(x)/\theta(\ell)$ when $x \in \mathbb{X}_\ell$,

$$\theta_{t+1}(i) > \theta_t(i) \qquad \theta_{t+1}(k) < \theta(k)$$

The proposal distribution (3/3)

The algorithm (step 1) Given the current biasing weight θ_t and the current sample X_t

- sample the new point :

$$X_{t+1} \sim P_{\theta_t}(X_t, \cdot)$$

where P_θ is a Markov kernel s.t. $\pi_\theta P_\theta = \pi_\theta$

- Update the biasing weight : if $X_{t+1} \in \mathbb{X}_i$, penalize the stratum i in order to favor the visits to the other stratum. Since $\pi_\theta(x) \propto \pi(x)/\theta(\ell)$ when $x \in \mathbb{X}_\ell$,

$$\theta_{t+1}(i) > \theta_t(i) \qquad \theta_{t+1}(k) < \theta(k)$$

Ex. of updating strategy :

$$\theta_{t+1}(i) = \theta_t(i) + \gamma_{t+1}\,\theta_t(i)(1-\theta_t(i))$$
$$\theta_{t+1}(k) = \theta_t(k) - \gamma_{t+1}\,\theta_t(i)\theta_t(k)$$

based on a Stochastic Approximation algorithm, with deterministic (non increasing) stepsize sequence $(\gamma_t)_t$

A toy example (1/3)

Target density : $\pi(x_1, x_2) \propto \exp(-\beta\,\mathcal{H}(x_1, x_2))\mathbb{1}_{[-R,R]}(x_1)$

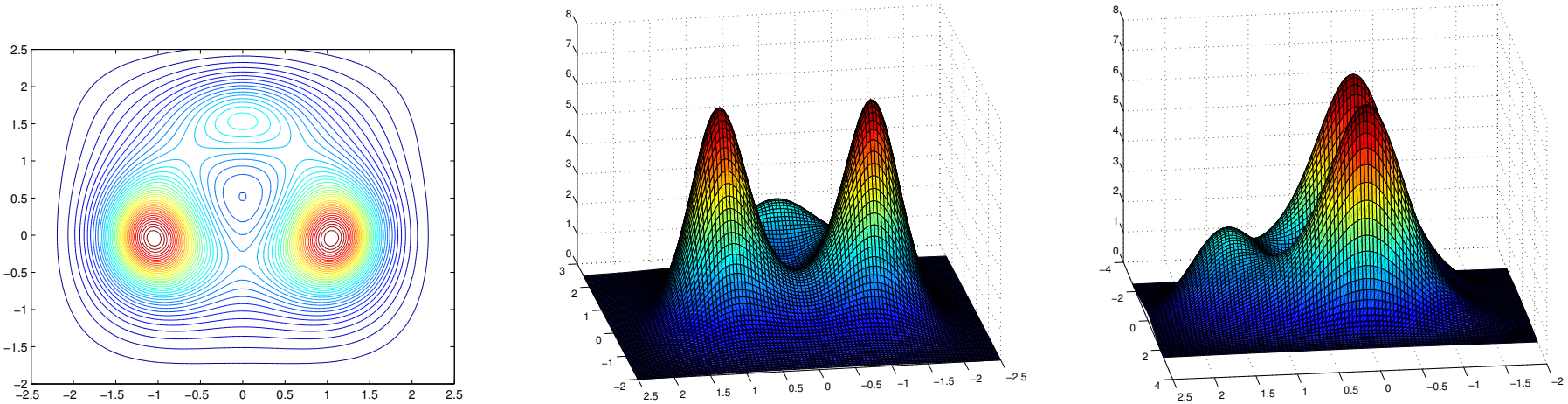
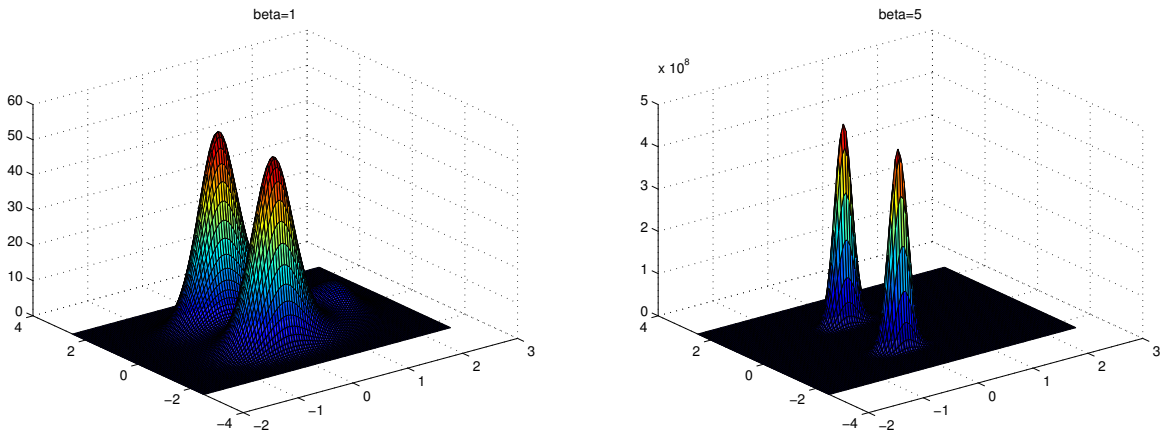
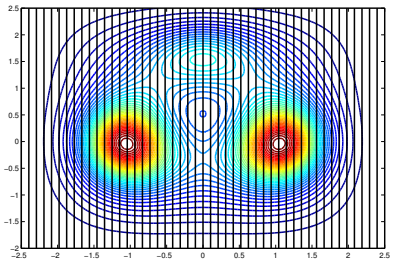


FIGURE: [left] Level curves of the potential $\mathcal{H}$. [center, right] Density π up to a normalizing constant.



The larger β is, the larger is the ratio between the weight of the strata located near to the main metastable states and the weight of the transition region (near $x_1 = 0$)

A toy example (2/3)



$d = 48$ strata, partition along the x-axis.

P_θ are Hastings-Metropolis kernels with proposal distribution $\mathcal{N}(0, (2R/d)^2 I)$ and target π_θ. $X_0 = (-1, 0)$. $R = 2.4$.

The stepsize sequence is $\gamma_t \sim c/t^{0.8}$.

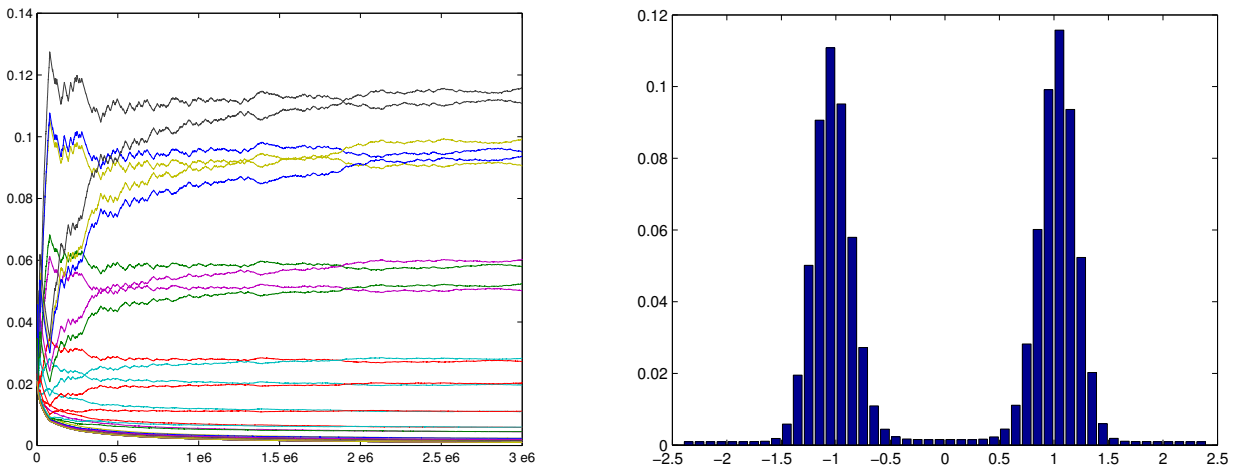


FIGURE: [left] The sequences $(\theta_t(i))_t$. [right] The limiting value $\theta_\star(i)$

A toy example (3/3)

Path of the x_1-component of $(X_t)_t$, when X_t is the WL chain (left) and the Hastings-Metropolis chain (right).

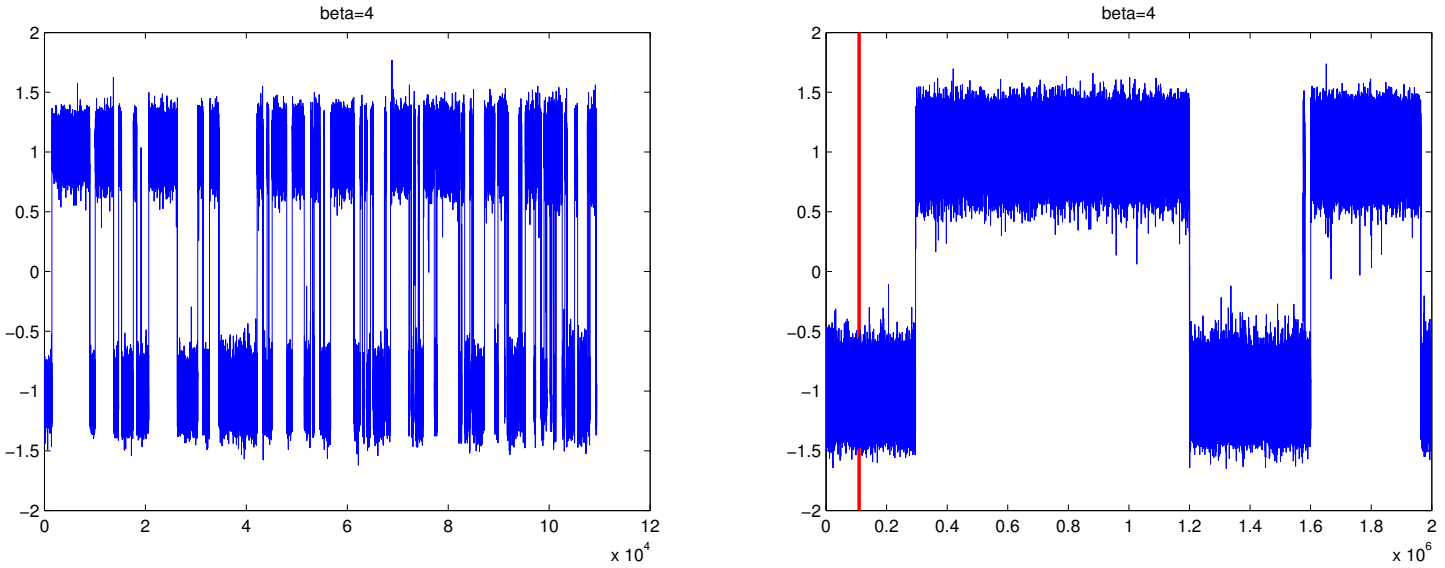
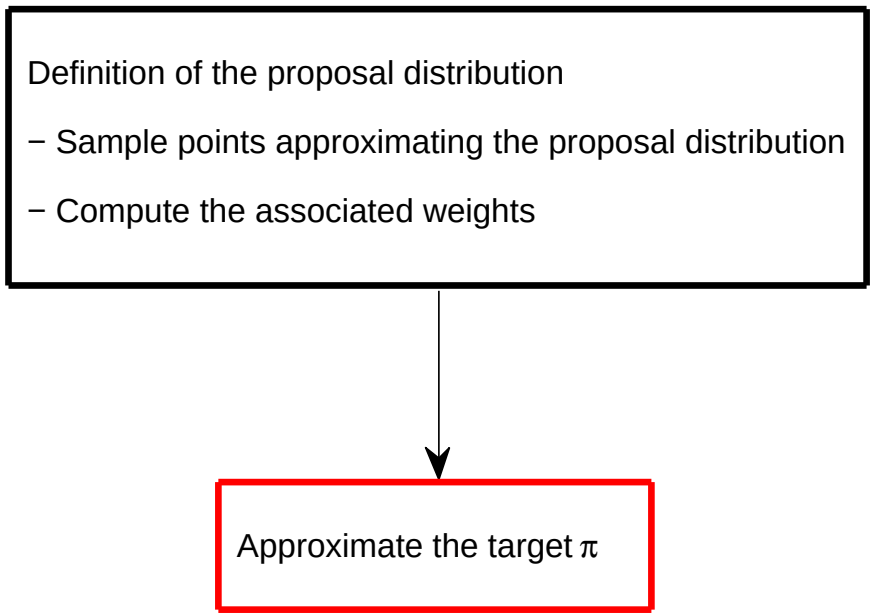


FIGURE: [left] Wang Landau, $T = 110\,000$. [right] Hastings Metropolis, $T = 2\ 10^6$; the red line is at $x = 110\,000$

Approximation of π (1/2)



Approximation of π (2/2)

- It is expected

$$\pi_\star \approx \frac{1}{T}\sum_{t=1}^{T}\delta_{X_t} \qquad \lim_t \theta_t = \left(\frac{\pi(\mathbb{X}_1)}{Z_\pi}, \cdots, \frac{\pi(\mathbb{X}_d)}{Z_\pi}\right)$$

- In addition, by definition of $\pi_\star$

$$x \in \mathbb{X}_i \Longrightarrow \frac{\pi(x)}{Z_\pi\ \pi_\star(x)} = d\ \lim_t \theta_t(i)$$

- Theis yields the algorithm (step 2)

$$\int f\,\frac{d\pi}{Z_\pi} \approx \frac{d}{T}\sum_{t=1}^{T} f(X_t)\left(\sum_{i=1}^{d}\theta_t(i)\mathbb{1}_{\mathbb{X}_i}(X_t)\right)$$

Efficiency of Wang-Landau (1/3)

- We compare the efficiency of the algorithm based on their ability to jump from one mode to another mode in a short time.

- Not possible to explicitly compute this time for general problems. We therefore considered toy examples.
  We report the results for a very simple example, for which explicit computations of the hitting-time is possible.

Efficiency of Wang-Landau (2/3)

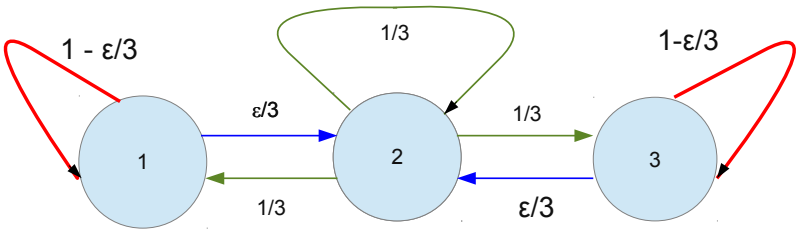
- State space : $\mathbb{X} = \{1, 2, 3\}$
- Target distribution : $\pi(1) \propto 1 \qquad \pi(2) \propto \epsilon \qquad \pi(3) \propto 1$

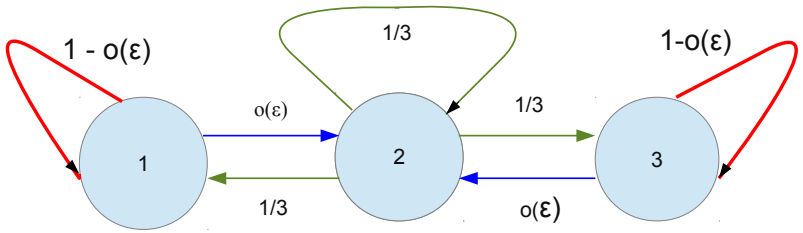
Let us compare

1. **Hastings-Metropolis** P with proposal kernel Q and target π

$$Q = \begin{bmatrix} 2/3 & 1/3 & 0 \\ 1/3 & 1/3 & 1/3 \\ 0 & 1/3 & 2/3 \end{bmatrix}$$

2. **Wang-Landau** where the kernels P_θ are Hastings-Metropolis kernels with proposal Q and target π_θ





(left) Transition matrix P ; (right) Behavior of the transition matrix P_θ when $\epsilon \to 0$ (θ fixed)

Efficiency of Wang-Landau (3/3)

- Comparison based on the hitting time

  $T_{1\to 3}$: hitting-time of state 3, given the chain started from state 1

  and how it behaves when $\epsilon \to 0$.

- When $\epsilon \to 0$, we obtain $T_{1\to 3}$ *scales like*

  For Hastings-Metropolis : $6/\epsilon$

$$\lim_{\epsilon \to 0} \frac{\epsilon}{6} \mathbb{E}\left[T_{1\to 3}\right] = 1$$

$$\frac{\epsilon}{6} T_{1\to 3} \to \mathcal{E}(1) \text{ in distribution}$$

  For Wang-Landau applied with $\gamma_t = \gamma_\star / t^a$ and $1/2 < a < 1$:

$$C(a, \gamma_\star) \; |\ln \epsilon|^{1/(1-a)}$$

  For Wang-Landau applied with $\gamma_t = \gamma_\star / t$

$$\epsilon^{-1/(1+\gamma_\star)}$$

Outline

Adaptive and Interacting MCMC (1/2)

- In the two previous examples, the conditional distribution of the points $(X_t)_t$ satisfies

$$\mathbb{E}\left[h(X_{t+1})|\text{past}_t\right] = \int h(y)\, P_{\theta_t}(X_t, dy)$$

where

P_θ is a Markov transition kernel
$(\theta_t)_t$ is a random process.

- Even in the simple situation when

there exists π such that $\pi P_\theta = \pi$ for any θ and

$$\lim_n \|P_\theta^n(x, \cdot) - \pi\| = 0$$

Is π the stationary distribution of the process $(X_t)_t$?

Adaptive and Interacting MCMC (2/2)

Consider the following adapted chain :

- Fix $t_0, t_1 \in (0,1)$. Define an adapted chain as follows :

$$X_{k+1}|\text{past}_k \sim \begin{cases} P_{t_0}(X_k, \cdot) & \text{if } X_k = 0 \\ P_{t_1}(X_k, \cdot) & \text{if } X_k = 1 \end{cases}$$

where

$$P_{t_\ell} = \begin{pmatrix} 1 - t_\ell & t_\ell \\ t_\ell & 1 - t_\ell \end{pmatrix}$$

- P_{t_0} and P_{t_1} both converge to the stationary distribution $\pi = \begin{pmatrix} 1/2 \\ 1/2 \end{pmatrix}$.
- Then, $(X_k)_k$ is a Markov chain, with transition matrix

$$\begin{pmatrix} 1 - t_0 & t_0 \\ t_1 & 1 - t_1 \end{pmatrix}$$

and it converges to the distribution $\tilde{\pi} \propto \begin{pmatrix} t_1 \\ t_0 \end{pmatrix} \neq \pi$.

Adaptive and Interacting MCMC (2/2)

Consider the following adapted chain :

- Fix $t_0, t_1 \in (0,1)$. Define an adapted chain as follows :

$$X_{k+1}|\text{past}_k \sim \begin{cases} P_{t_0}(X_k, \cdot) & \text{if } X_k = 0 \\ P_{t_1}(X_k, \cdot) & \text{if } X_k = 1 \end{cases}$$

where

$$P_{t_\ell} = \begin{pmatrix} 1-t_\ell & t_\ell \\ t_\ell & 1-t_\ell \end{pmatrix}$$

- P_{t_0} and P_{t_1} both converge to the stationary distribution $\pi = \begin{pmatrix} 1/2 \\ 1/2 \end{pmatrix}$.
- Then, $(X_k)_k$ is a Markov chain, with transition matrix

$$\begin{pmatrix} 1-t_0 & t_0 \\ t_1 & 1-t_1 \end{pmatrix}$$

and it converges to the distribution $\tilde{\pi} \propto \begin{pmatrix} t_1 \\ t_0 \end{pmatrix} \neq \pi$.

Adaption cas destroy the convergence !

Sufficient conditions for the convergence (1/3)

The literature provides sufficient conditions so that

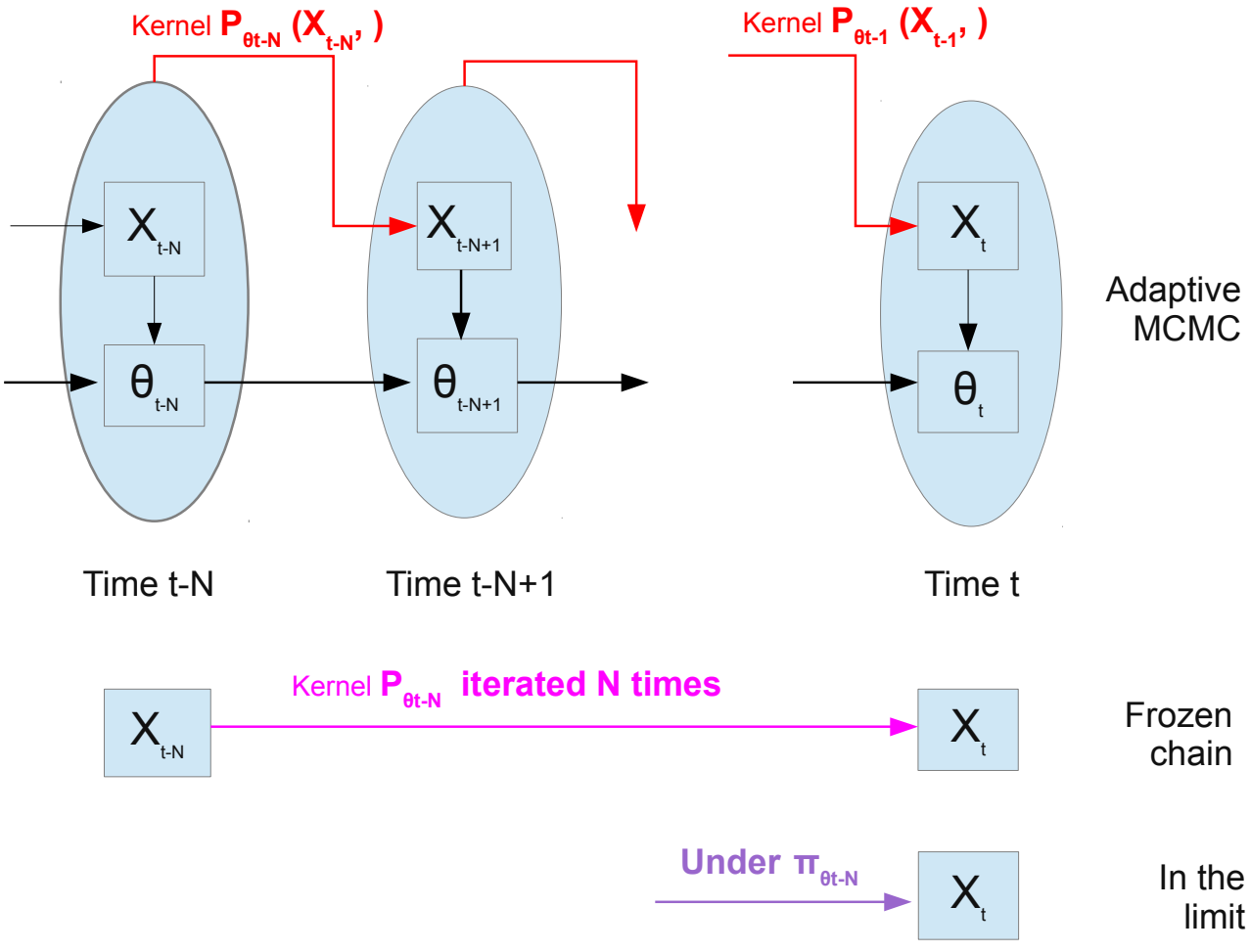
- convergence in distribution of $(X_t)_t$
- Strong law of large numbers for $(X_t)_t$
- Central Limit Theorem for $(X_t)_t$

G.O. Roberts, J.S. Rosenthal. Coupling and Ergodicity of Adaptive Markov chain Monte Carlo algorithms. J. Appl. Prob. (2007)

G. Fort, E. Moulines, P. Priouret. *Convergence of adaptive MCMC algorithms : ergodicity and law of large numbers.* Ann. Stat. 2012

G. Fort, E. Moulines, P. Priouret and P. Vandekerkhove. A Central Limit Theorem for Adaptive and Interacting Markov Chain. Bernoulli, 2013.

Sufficient conditions for the convergence (2/3)



Sufficient conditions for the convergence (3/3)

$$\mathbb{E}\left[h(X_t)|\text{past}_{t-N}\right] - \int h(y)\ \pi_{\theta_\star}(dy) = \mathbb{E}\left[h(X_t)|\text{past}_{t-N}\right] - \int h(y)\ P^N_{\theta_{t-N}}(X_{t-N}, dy)$$
$$+ \int h(y)\ P^N_{\theta_{t-N}}(X_{t-N}, dy) - \int h(y)\ \pi_{\theta_{n-N}}(dy)$$
$$+ \int h(y)\ \pi_{\theta_{n-N}}(dy) - \int h(y)\ \pi_{\theta_\star}(dy)$$

- Diminishing adaption condition Roughly speaking :

$$\text{dist}(P_\theta, P_{\theta'}) \leq \text{dist}(\theta, \theta')$$

  If $\theta_t - \theta_{t-1}$ are close, then the transition kernels P_{θ_t} and $P_{\theta_{t-1}}$ are close also.
- Containment condition Roughly speaking :

$$\lim_{N\to\infty} \text{dist}(P^N_\theta, \pi_\theta) = 0$$

  at some rate depending smoothly on θ.
- Regularity in θ of π_θ so that

$$\lim_t \theta_t = \theta_\star \implies \text{dist}\left(\pi_{\theta_t} - \pi_{\theta_\star}\right) \to 0$$

Convergence of Wang-Landau

Fort et al. (2014) provide sufficient conditions on

the transition kernels P_θ

the target distribution π

the step size $(\gamma_t)_t$ controlling the adaption rate of the weight sequence

implying that almost-surely

$$\lim_t \theta_t = \left(\frac{\pi(\mathbb{X}_1)}{Z_\pi}, \cdots, \frac{\pi(\mathbb{X}_d)}{Z_\pi} \right)$$

$$\lim_T \frac{1}{T} \sum_{t=1}^{T} f(X_t) = \int f \, d\pi_\star$$

$$\lim_T \frac{d}{T} \sum_{t=1}^{T} f(X_t) \sum_{i=1}^{d} \theta_t(i) \mathbb{1}_{\mathbb{X}_i}(X_t) = \int f \frac{d\pi}{Z_\pi}$$

The rate of convergence of $(\theta_t)_t$ is also established.

Outline

WL and simulated annealing (1/2)

Liang, Cheng and Lin. Simulated Stochastic Approximation Annealing for Global Optimization with a Square-Root-Cooling Schedule. JASA (2014)

- Let a cooling schedule $(T_t)_t$ such that $\lim_t \downarrow T_t = T_\star > 0$.
- Choose $\mathbb{X}_i = \{x : u_{i-1} < U(x) \leq u_i\}$.
- Set

$$\pi_{T,\theta}(x) \propto \sum_{i=1}^{d} \frac{1}{\theta(i)} \exp\left(-\frac{U(x)}{T}\right) \mathbb{1}_{\mathbb{X}_i}(x)$$

Algorithm : repeat

- Given the past, draw X_t under a transition kernel P_{T_t,θ_t} with invariant distribution π_{T_t,θ_t}.
- Update the weight parameter θ_t as in the WL algorithm.

WL and simulated annealing (2/2)

Results Liang et al. (2014)

1. Law of large numbers : a.s.

$$\lim_N \frac{1}{N} \sum_{t=1}^{N} f(X_t) = \int f(x) \frac{\pi_{T_\star,\theta_\star}(x)}{Z_\star} \, \mathrm{d}\lambda$$

2. Let $u_\star$ be the minimal value, necessarily reached in stratum $\mathbb{X}_1$

$$\lim_{t\to\infty} \mathbb{P}\left(U(X_t) \leq u_\star + \epsilon | X_t \in \mathbb{X}_1\right) = \mathbb{P}\left(U(Y) \leq u_\star + \epsilon | Y \in \mathbb{X}_1\right)$$

where $Y \sim \exp(-U(x)/T_\star)$.

3. "When $T_\star \to 0$,

$$\mathbb{P}\left(U(Y) \leq u_\star + \epsilon | Y \in \mathbb{X}_1\right) \to 1,$$

thus showing the convergence of the algorithm to the minima of $U(x)$."

But we could imagine other methods to combine WL and Simulates Annealing, or WL and stochastic optimization