Dynamiques bien tempérées pour l'exploration Monte Carlo de lois multimodales

Gersende FORT

LTCI, CNRS & Telecom ParisTech
Paris, France

Journée AppliBUGS
Paris, Juin 2015

Context

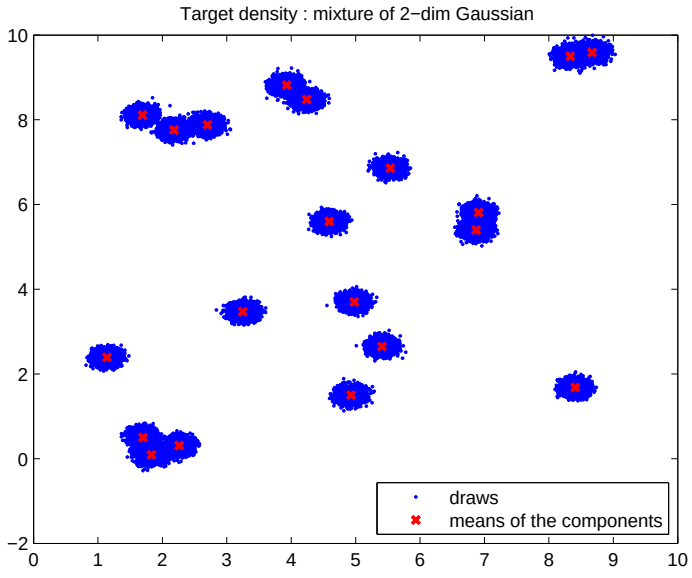
Sample from a target distribution $\pi\, d\lambda$ on $\mathbb{X} \subseteq \mathbb{R}^\ell$,
when π is (possibly) known up to a normalizing constant,

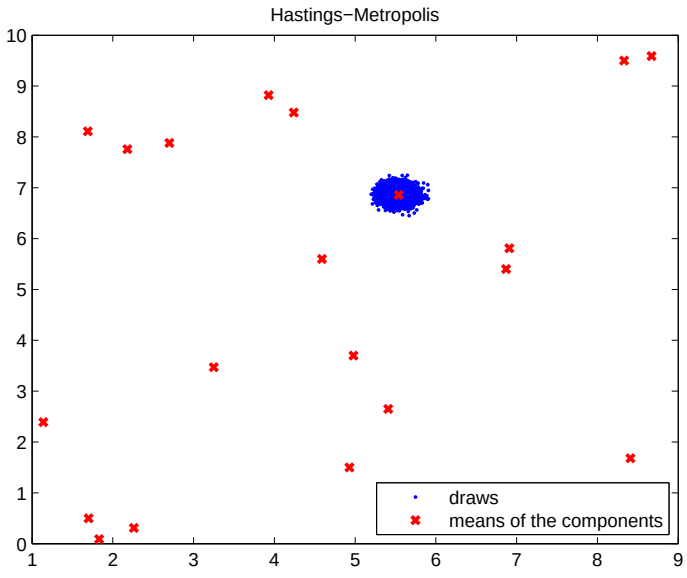
$\hookrightarrow$ *Hereafter, to make the notations simpler, π is assumed to be normalized*

in the context

- π is multimodal
- large dimension

The challenge, on a toy example





π : the target distribution

q : the proposal distribution or proposal transition kernel, chosen by the user.

1. MCMC
   - given X_k, sample $Y \sim q(X_k, \cdot)$
   - accept-reject mecanism

$$X_{k+1} = \begin{cases} Y & \text{with probability } 1 \wedge \frac{\pi(Y)}{\pi(X_k)} \frac{q(Y, X_k)}{q(X_k, Y)} \\ X_k & \text{otherwise} \end{cases}$$

$\hookrightarrow$ Approximation:

$$\pi \approx \frac{1}{n} \sum_{k=1}^{n} \delta_{X_k}$$

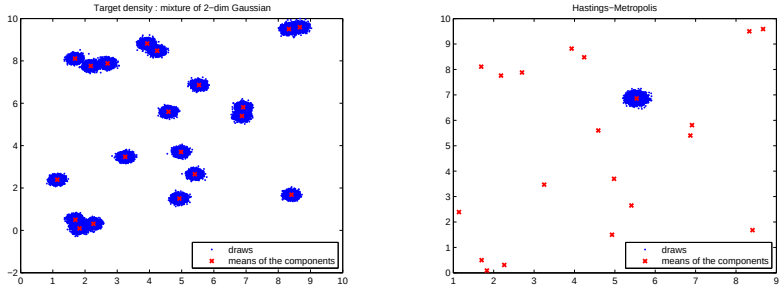
2. Importance Sampling
   - Sample i.i.d. points $(X_k)_k$ with distribution q
   - Associate a weight to each point: $\pi(X_k)/q(X_k)$.

$\hookrightarrow$ Approximation:

$$\pi \approx \frac{1}{n} \sum_{k=1}^{n} \frac{\pi(X_k)}{q(X_k)} \delta_{X_k}$$

Do classical adaptive samplers answer the question ?

1. Adaptive Hastings-Metropolis algorithm



2. Adaptive Importance Sampling

   Sampling Draw points $(X_k^{(t)})_k$ from the proposal $q^{(t)}$ + importance reweighting

   Adaption Update the sampling distribution

$\hookrightarrow$ How to learn π on domains which are not visited (yet) by the algorithm ?
How to avoid degeneracy of the acceptance / reweighting ratios ?

Talk based on joint works with





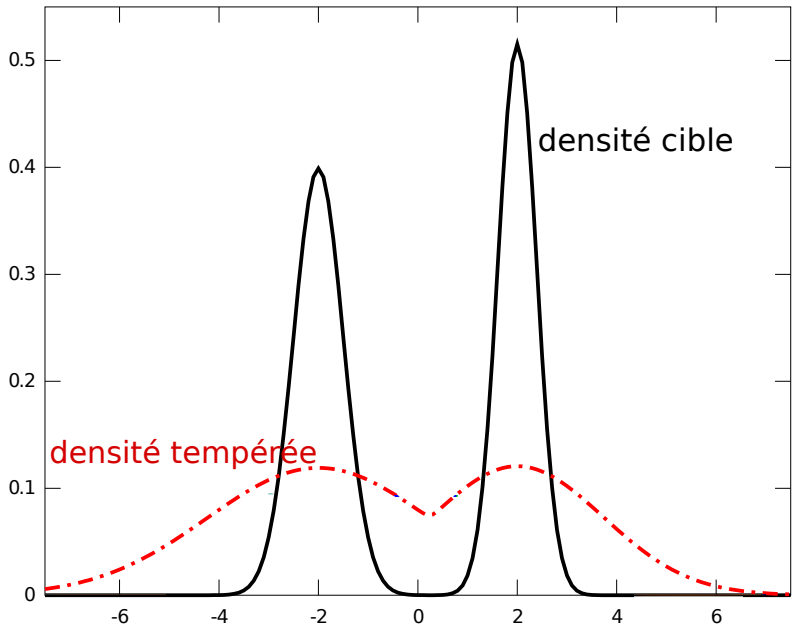

Eric Moulines, (Telecom Paris-Tech)
Pierre Priouret (Paris VI)

Benjamin Jourdain,
Tony Lelièvre,
Gabriel Stoltz
(ENPC)

Estelle Kuhn (INRA)

Outline

Tempering: the idea

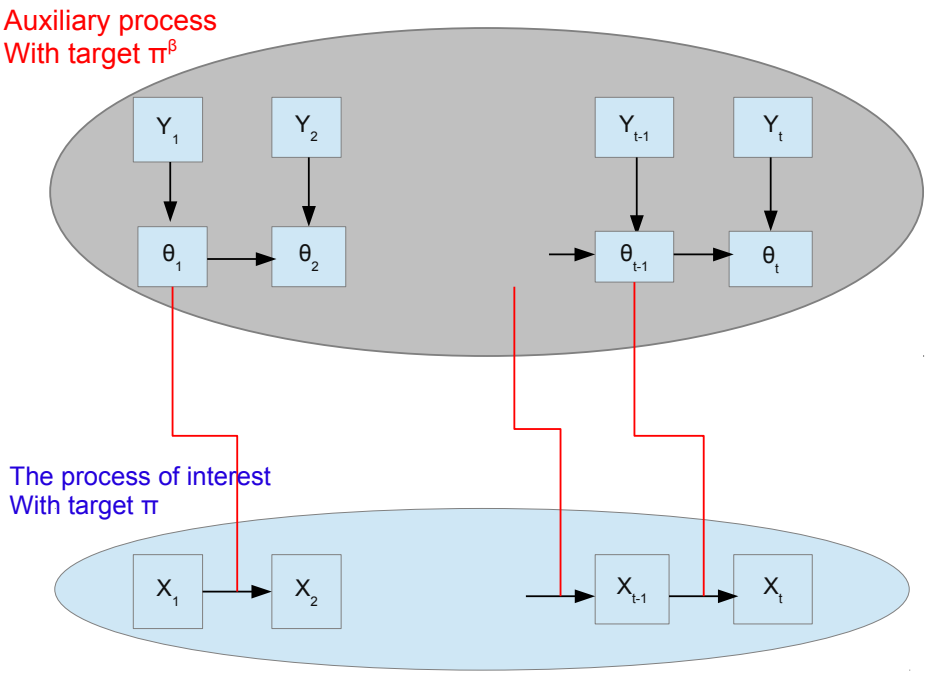


- Learn a well fitted proposal mecanism by considering tempered versions $\pi^{1/T}$ $(T > 1)$ of the target distribution π.
- Hereafter, an example where tempering is plugged in a MCMC sampler.

Equi-Energy sampler (1/2)

Kou, Zhou and Wong (2006)

- In the MCMC proposal mecanism, allow to pick a point from an auxiliary process designed to have better mixing properties.

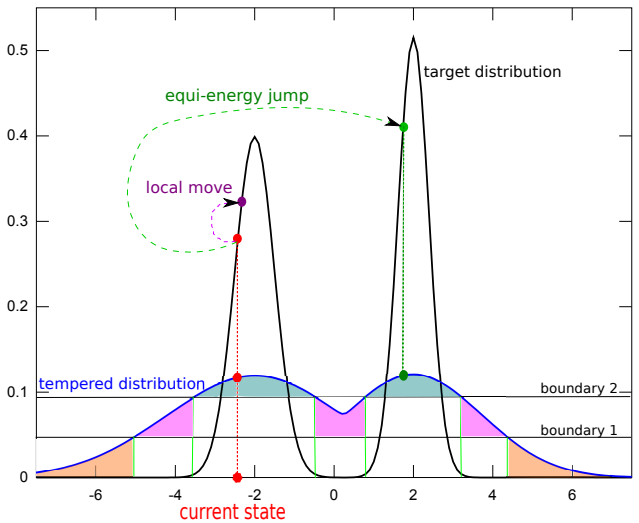


Algorithm: at iteration t, given

the current state X_t

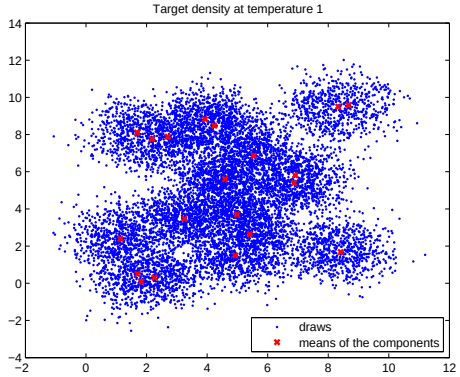
the samples $Y_1, \cdots, Y_t$ from the auxiliary process

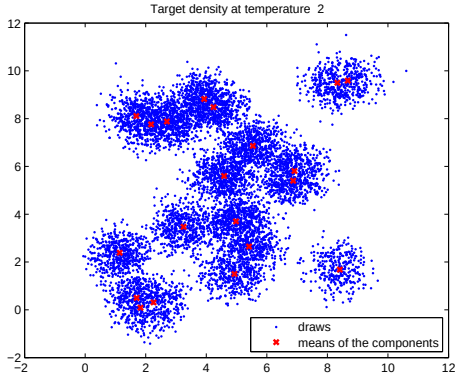
1. with probability $1 - \epsilon$, draw
   $X_{t+1} \sim$ MCMC kernel with invariant distribution π
2. with probability ϵ, choose a point Y_ℓ among the auxiliary samples in the same energy level as X_t and accept/reject the move $X_{t+1} = Y_\ell$.

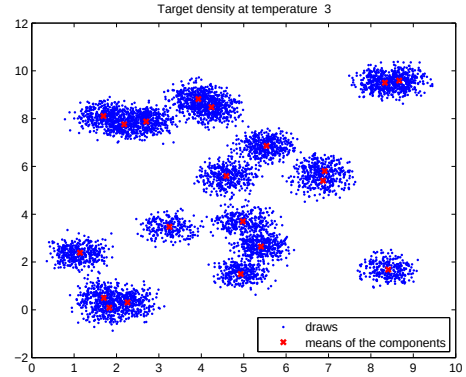


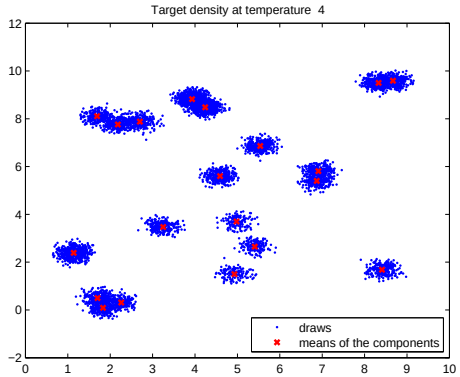
Example 1

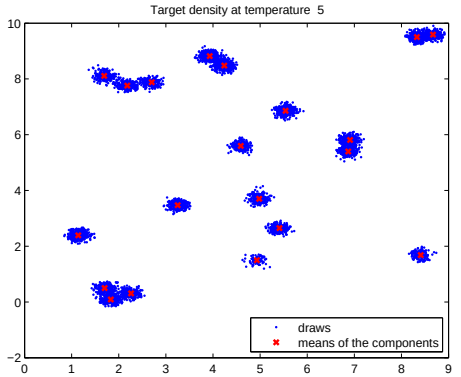
π is a mixture of 20 Gaussian distributions.
With 4 auxiliary processes $\pi^{\beta_4}, \cdots, \pi^{\beta_1}$, $\quad 0 < \beta_4 < \cdots < \beta_1 < 1.$

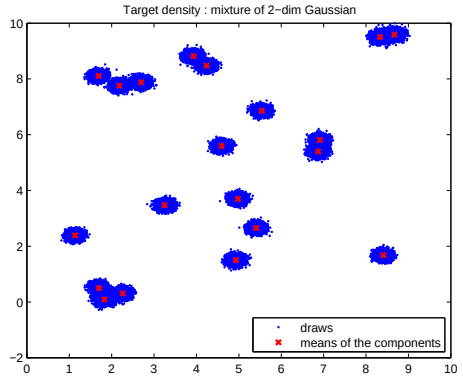








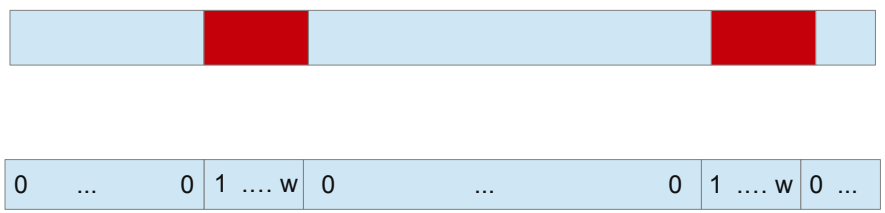




Example 2

Schreck, F. and Moulines (2013)

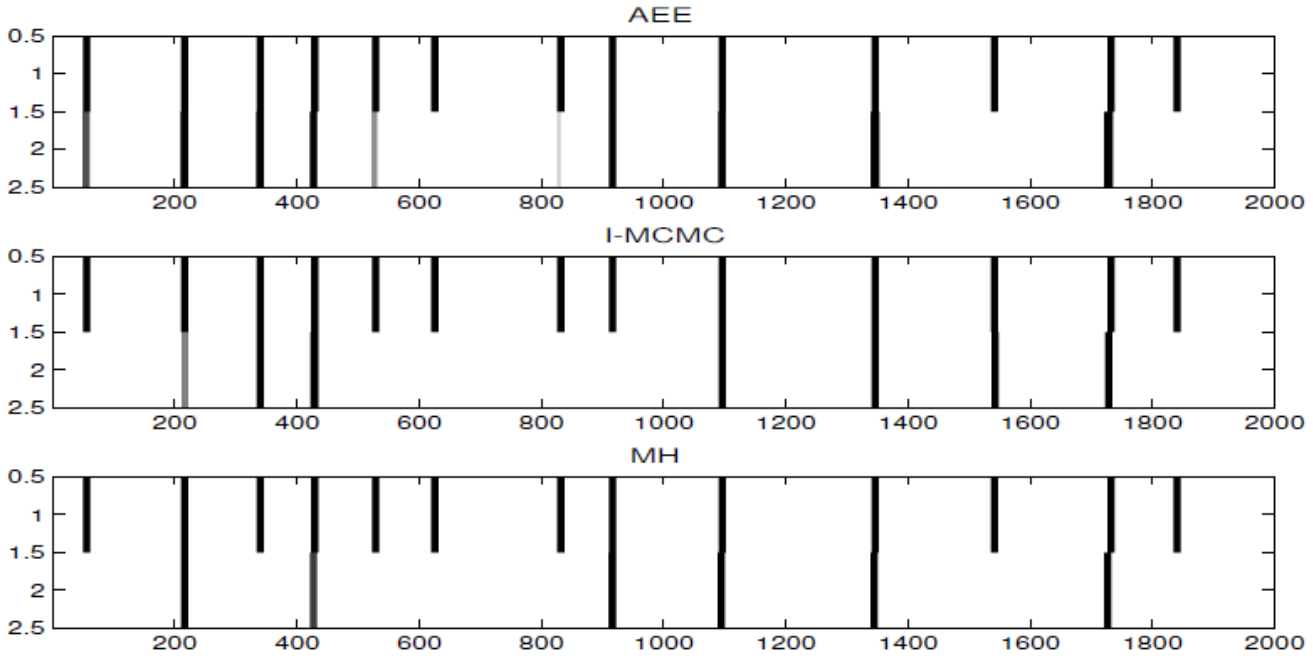
- Problem: Motif sampling in biological sequences
  - objective: find where motifs (a subsequence of length $w = 12$) are, in a ADN sequence of length $L = 2000$.
  Observation: $(s_1, \cdots, s_L)$ with $s_l \in \{A, C, G, T\}$.
  Quantity of interest: motifs position collected in $(a_1, \cdots, a_L)$ with $a_j \in \{0, \cdots, w\}$



Example 2

Schreck, F. and Moulines (2013)

- Result: EE with 4 auxiliary chains and 3 energy rings



Design parameters

- Design parameters
  - the probability of interaction ϵ
  - the number of auxiliary processes and the scale of the β_i
  - the energy rings
  - the MCMC kernels for the local moves

Does it converge ?

- Transition kernel:

$$P_{\theta_t}(X_t,A) = (1-\epsilon)P(X_t,A) + \epsilon \int \underbrace{1 \wedge \frac{\pi(y)\, g(y,X_t)\ \pi^{\beta}(X_t)}{\pi(X_t) g(X_t,y)\ \pi^{\beta}(y)}}_{\text{acceptance-rejection}} \underbrace{g(X_t,y)\theta_t(dy)}_{\text{proposition with selection}}$$

where

$$\theta_t = \frac{1}{t}\sum_{k=1}^{t} \delta_{Y_k}$$

`Convergence Analysis:` Andrieu, Jasra, Doucet and Del Moral (2007,2008); F., Moulines, Priouret (2012); F., Moulines, Priouret and Vandekerkhove (2013)

`Adaptive version of Equi Energy Sampler:` Schreck, F. and Moulines (2013); Baragatti, Grimaud and Pommeret (2013)

Outline

The idea

- Among the *Importance Sampling* Monte Carlo sampler

$$\pi \approx \frac{1}{n} \sum_{t=1}^{n} \frac{\pi(X_t)}{q_\star(X_t)} \delta_{X_t} \qquad \text{where } (X_t)_t \text{ approximates } q_\star$$

- Idea from the molecular dynamics field; see e.g. Chopin, Lelièvre and Stoltz (2012) for the first extensions to Computational Statistics

  Choose a proposal distribution of the form

$$q_\star(x) = \pi(x) \exp\left(-A(\xi(x))\right)$$

  where

  1. $A(\xi(x))$ is a biasing potential depending on few "directions of metastability" $\xi(x)$
  2. and such that q is "less multimodal" than π.

Wang-Landau samplers

Wang and Landau (2001) - very popular algorithm in the molecular dynamics field

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

- Choose a partition of $\mathbb{X}$ in d strata: $\mathbb{X}_1, \cdots, \mathbb{X}_d$
- set $\xi(x) = i$ for any $x \in \mathbb{X}_i$

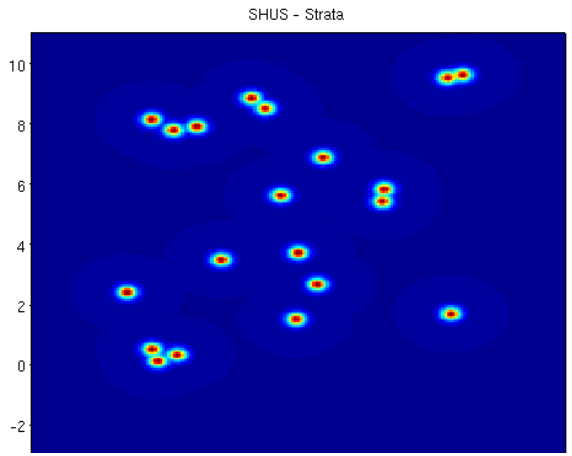
$$\begin{aligned} q_\star(x) &= \sum_{i=1}^{d} \mathbb{I}_{\mathbb{X}_i}(x)\pi(x)\ \exp(-A(i)) \\ &= \sum_{i=1}^{d} \mathbb{I}_{\mathbb{X}_i}(x)\frac{\pi(x)}{\theta_\star(i)} \end{aligned}$$

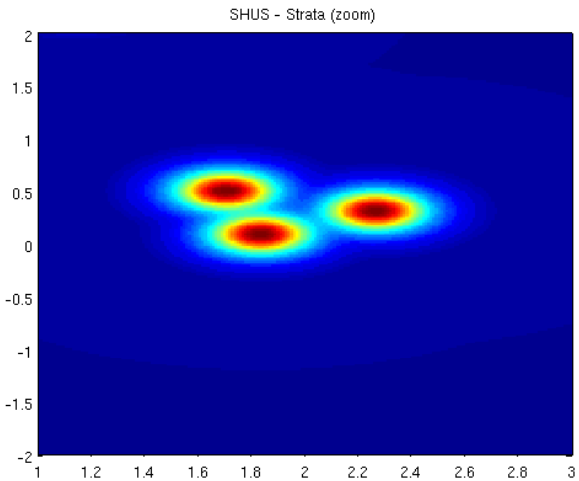
- Choose the weight $\theta_\star(i)$ so that

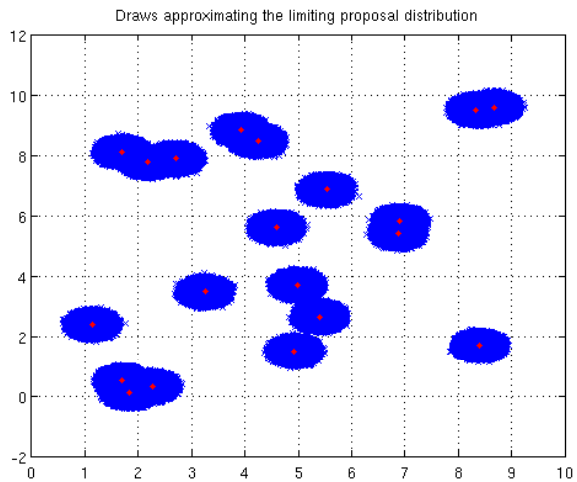
$$\int_{\mathbb{X}_i} q_\star(x)\ \mathrm{d}x = \frac{1}{d} \Longleftrightarrow \theta_\star(i) = d\int_{\mathbb{X}_i} \pi(x)\mathrm{d}x.$$

A toy example

Target distribution: mixture of 20 Gaussian in $\mathbb{R}^2$. The means of the Gaussians are indicated with a red cross
Wang Landau algorithm: 50 strata, otained by partitioning the energy levels.



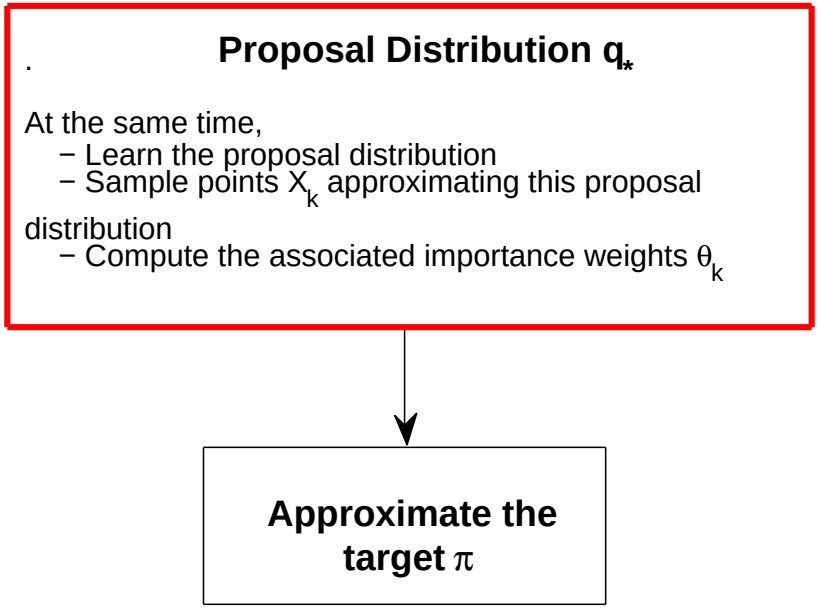




$5\,10^6$ draws approximating $q_\star$: the sampler was able to jump the deep valleys and draw points around all the modes.

An adaptive Importance Sampler

In practice, $q_\star$ is not available since $\theta_\star(i)$ are unknown $\rightarrow$ Adaptive Sampler



At iteration t:

- approximation q_t of $q_\star$
- draw X_t approximating q_t, and compute its associated importance weight $\pi(X_t)/q_t(X_t)$.

$$\pi \approx \frac{1}{n}\sum_{t=1}^{n} \frac{\pi(X_t)}{q_t(X_t)}\delta_{X_t}$$

The algorithm

Wang-Landau algorithm: at iteration t, given

the current point X_t

the current bias $\theta_t = (\theta_t(1), \cdots, \theta_t(d))$

1. Draw a new point

$$X_{t+1} \sim \text{MCMC with invariant distribution } q_t(x) \propto \sum_{i=1}^{d} \frac{\pi(x)}{\theta_t(i)} \mathbb{I}_{\mathbb{X}_i}(x)$$

2. Update the bias θ_{t+1} (see next slides)
3. In parallel, update the approximation of π

$$\pi \propto \frac{1}{n} \sum_{t=1}^{n} \left(d \sum_{i=1}^{d} \theta_t(i) \mathbb{I}_{\mathbb{X}_i}(X_t) \right) \delta_{X_t}$$

The update mecanism (1/2)

To learn $\theta_\star$ on the fly: many strategies in the literature, based on Stochastic Approximation algorithms with controlled Markov chain dynamics $(X_t)_t$

$$\theta_{t+1}(i) = \theta_t(i) + \gamma_{t+1}\, \mathcal{H}_i(\theta_t, X_{t+1})$$

where $\mathcal{H}_i$ is chosen so that

- penalize the stratum currently visited: $\mathcal{H}_i(\theta_t, X_{t+1}) > 0$ iff $X_{t+1} \in \mathbb{X}_i$
- the mean field function $\theta \mapsto \int \mathcal{H}(\theta, x)\, q_\star(x) dx$ admits $\theta_\star$ as the unique root.

and

- $X_{t+1} \sim$ MCMC with invariant distribution $q_{\theta_t}(x)$
- the stepsize sequence $\{\gamma_t, t \geq 0\}$ is deterministic or random but satisfies

$$\sum_t \gamma_t = \infty \text{ (a.s.)} \qquad \sum_t \gamma_t^2 < \infty \text{ (a.s.)}$$

The update mecanism (2/2)

$$\forall i \in \{1, \cdots, d\}, \qquad \theta_{n+1}(i) = \frac{\tilde{\theta}_{n+1}(i)}{\sum_{\ell=1}^{d} \tilde{\theta}_{n+1}(\ell)}$$

1. Wang-Landau algorithm:

$$\tilde{\theta}_{n+1}(i) = \tilde{\theta}_n(i) + \gamma_{n+1} \tilde{\theta}_n(i) \mathbb{1}_{\mathbb{X}_i}(X_{n+1})$$

2. the SHUS algorithm: (Self Healing Umbrella Sampling)

$$\tilde{\theta}_{n+1}(i) = \tilde{\theta}_n(i) + \gamma \theta_n(i) \mathbb{1}_{\mathbb{X}_i}(X_{n+1})$$
$$= \tilde{\theta}_n(i) + \gamma_{n+1} \tilde{\theta}_n(i) \mathbb{1}_{\mathbb{X}_i}(X_{n+1}) \qquad \gamma_{n+1} = \frac{\gamma}{\sum_{\ell=1}^{d} \tilde{\theta}_n(\ell)}$$

3. the Well-Tempered algorithm: $a \in (0,1]$

$$\tilde{\theta}_{n+1}(i) = \tilde{\theta}_n(i) + \gamma \left(\theta_n(i)\right)^a \mathbb{1}_{\mathbb{X}_i}(X_{n+1})$$
$$= \tilde{\theta}_n(i) + \gamma_{n+1} \tilde{\theta}_n(i) \mathbb{1}_{\mathbb{X}_i}(X_{n+1}) \qquad \gamma_{n+1} = \frac{\gamma \sum_{\ell=1}^{d} \theta_n^{a-1}(\ell) \mathbb{1}_{\mathbb{X}_\ell}(X_{n+1})}{\sum_{\ell=1}^{d} \tilde{\theta}_n(\ell)}$$

Transition kernel

- The conditional distribution of X_{t+1} given the past is a MCMC kernel with invariant distribution q_t, denoted by P_{θ_t}
- Example: HM with Gaussian proposal distribution

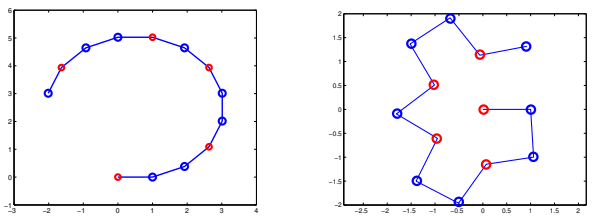
$$P_\theta(x,A) = \int_A \left(1 \wedge \frac{\pi(y)}{\pi(x)} \frac{\theta(\mathsf{str}(x))}{\theta(\mathsf{str}(y))}\right) \mathcal{N}(x,\Sigma)[dy] + \delta_x(A) \int 1 - \left(1 \wedge \frac{\pi(y)}{\pi(x)} \frac{\theta(\mathsf{str}(x))}{\theta(\mathsf{str}(y))}\right) \mathcal{N}(x,\Sigma)[dy]$$

where $\mathsf{str}(x) = i$ iff $x \in \mathbb{X}_i$.

Structure of a protein

In biophysics, structure of a protein from its sequence.

AB model: two types of monomers A (hydrophobic) and B (hydophilic), linked by rigidbonds of unit length to form (2D) chains. Given a sequence, what is the optimal shape of the N monomers?



Minimize the energy function $\mathcal{H}(x)$ on

$$x = (x_{1,2}, x_{2,3}, \cdots, x_{N-2,N-1}) \in [-\pi,\pi]^{N-2}$$

where

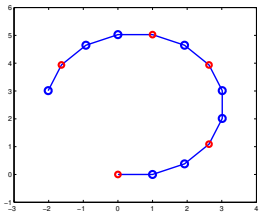
$$\mathcal{H}(x) = \frac{1}{4}\sum_{i=1}^{N-2}(1-\cos(x_{i,i+1})) + 4\sum_{i=1}^{N-2}\sum_{j=i+2}^{N}\left(\frac{1}{r_{ij}^{12}} - \frac{C(\sigma_i,\sigma_j)}{r_{ij}^{6}}\right)$$

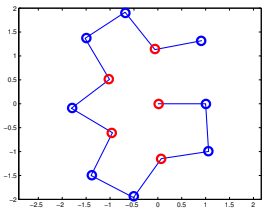
$x_{i,j}$ is the angle between i-th and j-th bond vector
r_{ij} is the distance between monomers i,j
$C(\sigma_i,\sigma_j) = 1$ (resp. $1/2$ and $-1/2$) between monomers AA (resp. BB and AB).

Structure of a protein

In biophysics, structure of a protein from its sequence.

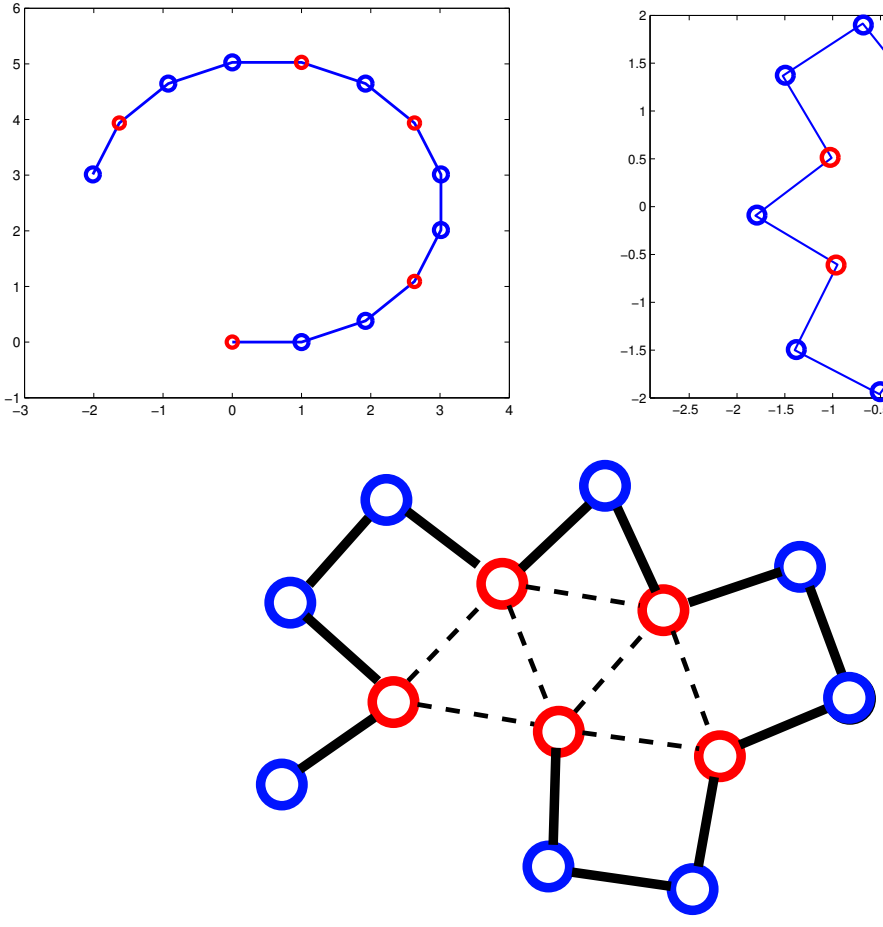
AB model: two types of monomers A (hydrophobic) and B (hydophilic), linked by rigidbonds of unit length to form (2D) chains. Given a sequence, what is the optimal shape of the N monomers?





Minimize the energy function $\mathcal{H}(x)$ on

$$\min \mathcal{H}(x) \Longleftrightarrow \max \pi_n(x) \propto \exp(-\beta_n \mathcal{H}(x)) \qquad \beta_n > 0$$



(top left) WL: initial config with energy 0.1945; (top right) WL: optimal config with energy -3.2925; (bottom) optimal config in the literature with energy -3.2941

Convergence results

`Convergence analysis:` Liang (2005); Liang, Liu and Carroll (2007); Atchadé and Liu (2010); Jacob and Ryder (2012); F., Jourdain, Kuhn, Lelièvre and Stoltz (2014a); F., Jourdain, Lelièvre and Stoltz (submitted: 2014, 2015)

`Efficiency analysis:` F., Jourdain, Kuhn, Lelièvre and Stoltz (2014b); F., Jourdain, Lelièvre and Stoltz (2014)

`Adaptive Wang Landau:` Bornn, Jacob, Del Moral and Doucet (2012)

Design parameters (1/6)

Many design parameters:

stepsize γ_n in the Stochastic Approximation update step

number of strata d

transition kernels P_θ

which may play a role on

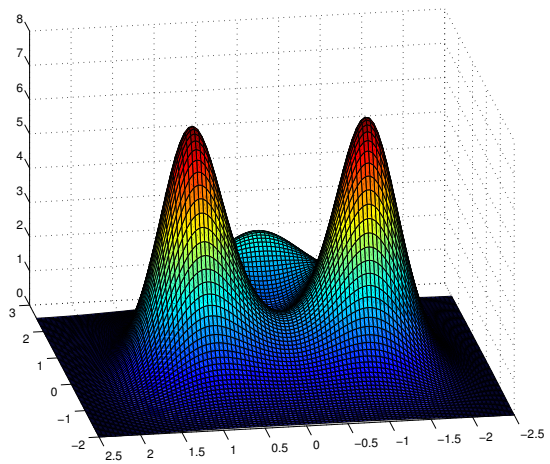
- the limiting behavior of the sampler
- the transient phase of the sampler: for example, how long is the exit time from a mode?

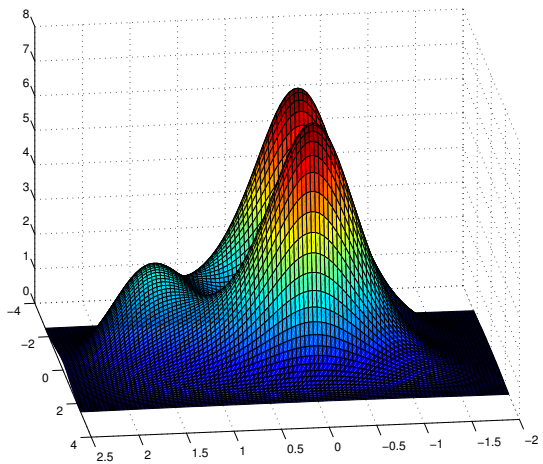
$\hookrightarrow$ Let us discuss the role of γ_n and d.

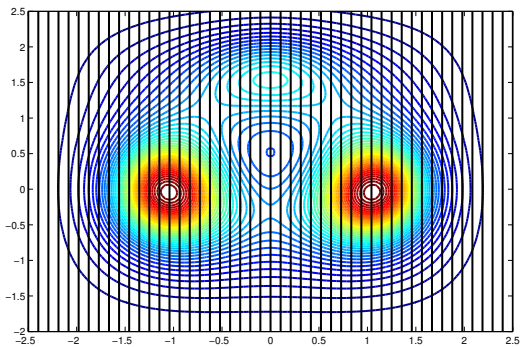
Design parameters (2/6)

Through an example:

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







d strata (see the right plot); the chains are initialised at $(-1,0)$

We compute the mean time over M runs to move from the left mode to the center.

Design parameters (3/6)

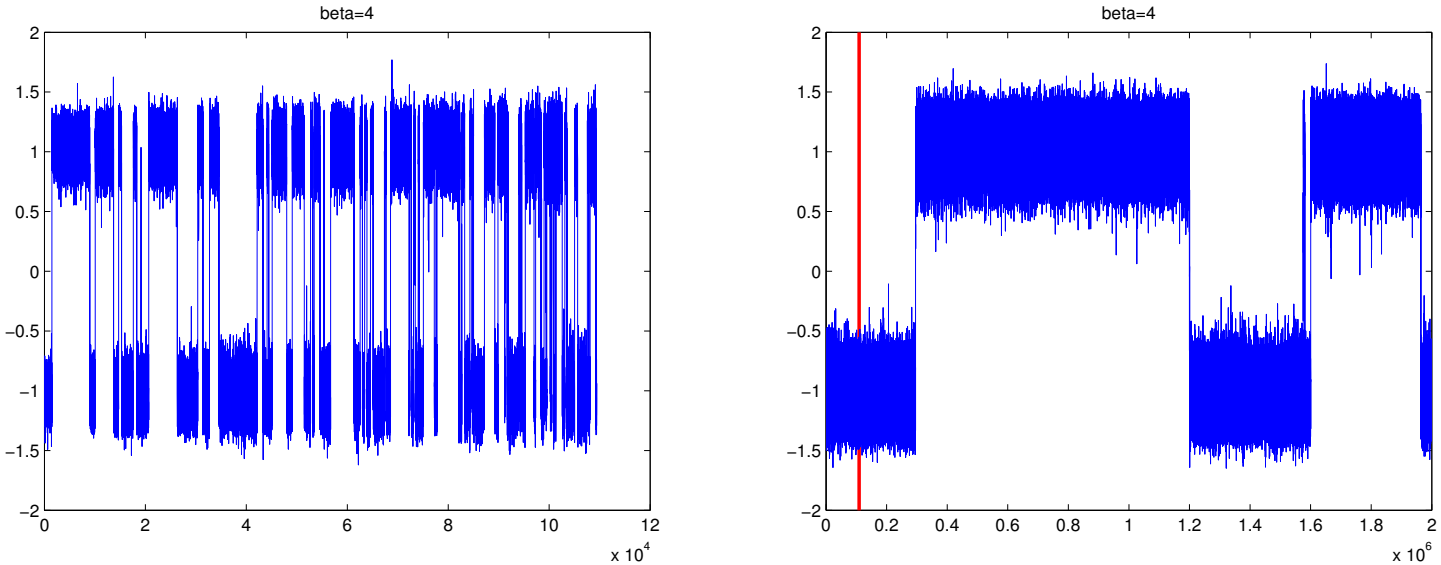


FIG.: [left] Wang Landau, $T = 110\,000$ and $d = 48$. [right] Hastings Metropolis, $T = 2\ 10^6$; the red line is at $x = 110\,000$

Design parameters (4/6)

For the long-time behavior, the theory establishes

Wang-Landau

- Convergence for any $\gamma_n \sim \frac{\gamma}{n^\alpha}$ $\qquad \alpha \in]1/2,1]$
- When estimating $\theta_\star$, the rate of convergence is optimal with $\gamma_n = \frac{d}{n}$.

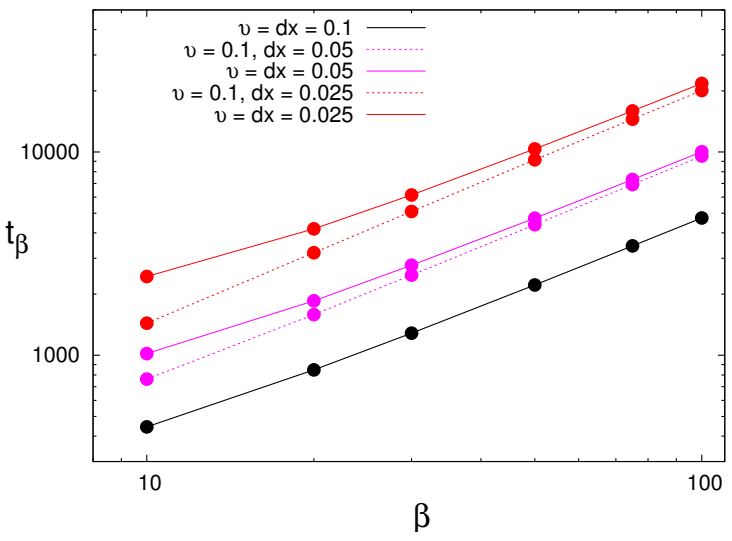
SHUS

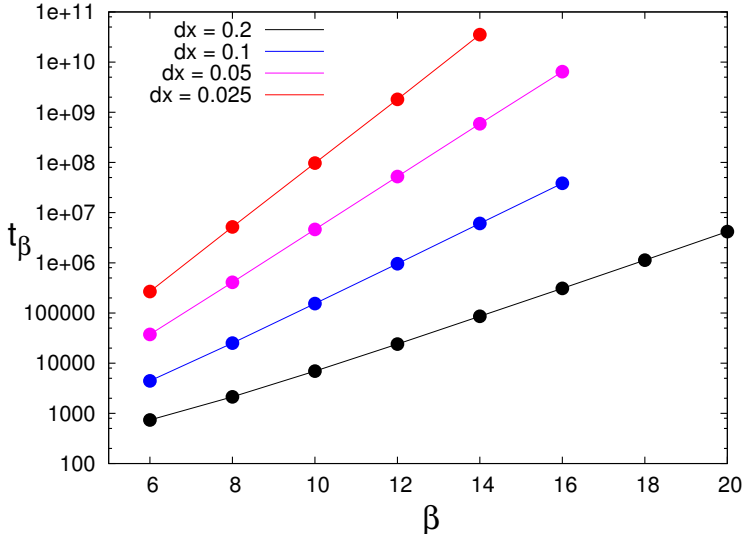
- When $n \to \infty$, w.p.1 $\lim_n \gamma_n = \frac{d}{n}$.
- Simple modifications of the updating mecanism yield: w.p.1 $\lim_n \gamma_n = \frac{\gamma_\alpha}{n^\alpha}$.

Design parameters (5/6)

For the short-time behavior: F., Jourdain, Kuhn, Lelièvre, Stoltz (2014)

(mean) exit times t_β from the left mode for different values of α. Here: Wang Landau (idem for the modified SHUS)



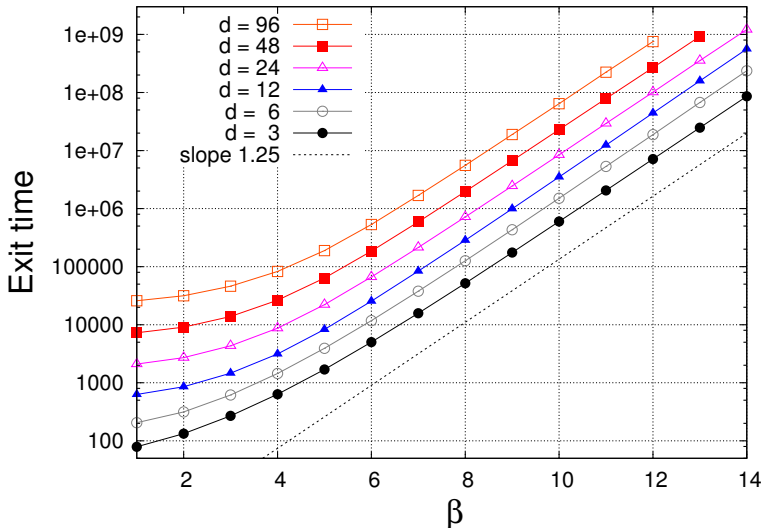


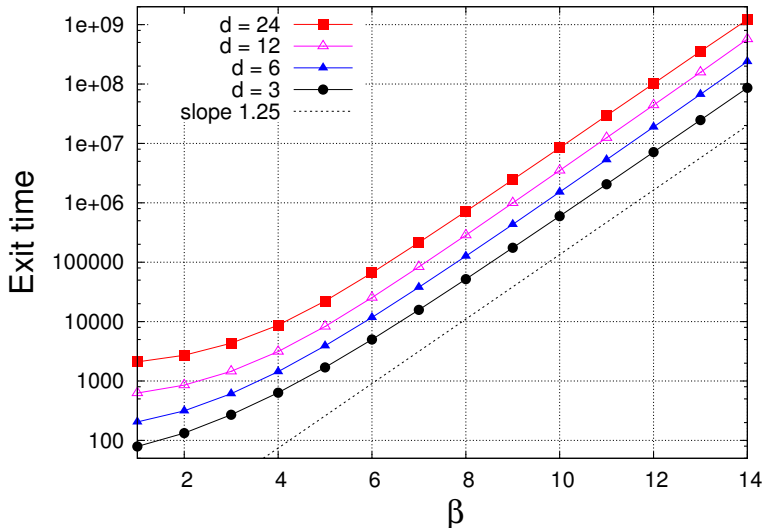
$$t_\beta = C(\beta,\sigma,d)\ t^{1/(1-\alpha)} \qquad t_\beta = C(\beta,\sigma,d)\ \exp(\beta\mu)$$

Design parameters (6/6)

For the short-time behavior: F., Jourdain, Lelièvre, Stoltz (2014)

(mean) When $\gamma_n = \gamma/n$, exit times t_β from the left mode for different values of d. and [left] a fixed proposal scale σ in the MCMC samplers; [right] a proposal scale $\sigma \propto 1/d$ in the MCMC samplers. Here: SHUS





$$t_\beta = C(\beta,\sigma,d)\ \exp(\beta\mu)$$ with a slope μ independent of d and σ

Outline

Controlled Markov chains (1/2)

These new samplers combine adaption/interaction and sampling: the draws $(X_t)_t$ are from a controlled Markov chain

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

where $(P_\theta, \theta \in \Theta)$ is a family of Markov kernels having an invariant distribution π_θ.

Controlled Markov chains (2/2)

- Question: let $(P_\theta, \theta \in \Theta)$ be a family of Markov kernels having the same invariant distribution π. Let $(\theta_t)_t$ be some $\mathcal{F}_t$-adapted random processes and draw

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

  Does $(X_t)_t$ converges (say in distribution) to π?

Controlled Markov chains (2/2)

- Question: let $(P_\theta, \theta \in \Theta)$ be a family of Markov kernels having the same invariant distribution π. Let $(\theta_t)_t$ be some $\mathcal{F}_t$-adapted random processes and draw

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

  Does $(X_t)_t$ converges (say in distribution) to π?

- Answer: No. For example:

$$X_{t+1} \sim \begin{cases} P_0(X_t, \cdot) & \text{if } X_t = 0 \\ P_1(X_t, \cdot) & \text{if } X_t = 1 \end{cases}$$

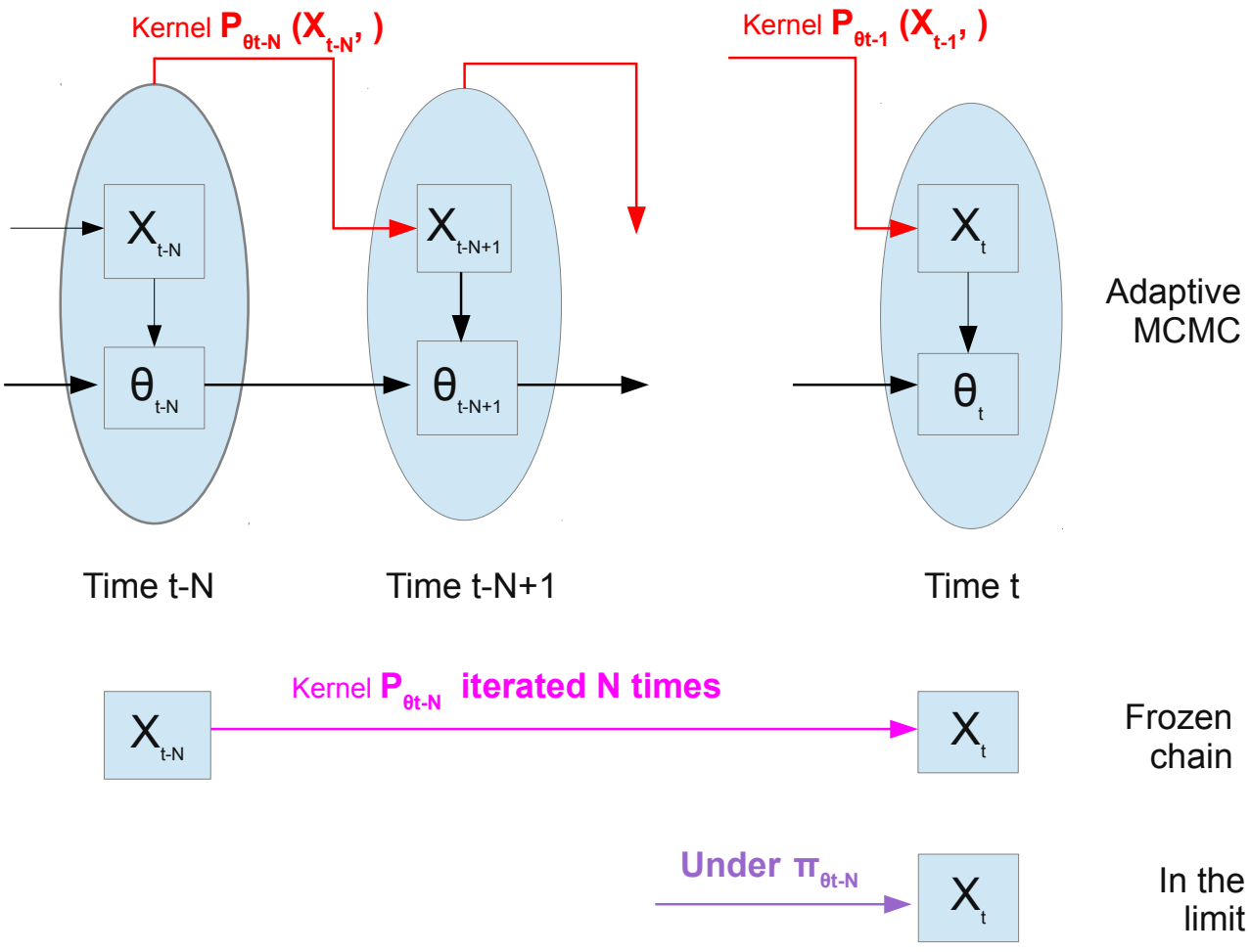
  where

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

  We have $\pi P_\ell = \pi$ with $\pi \propto (1,1)$ but the transition matrix of $(X_t)_t$ is

$$\tilde{P} = \begin{pmatrix} 1 - t_0 & t_0 \\ t_1 & 1 - t_1 \end{pmatrix} \qquad \text{with invariant distribution } \tilde{\pi} \propto (t_1, t_0)$$

Sufficient conditions for the cvg in distribution (1/3)



Sufficient conditions for the cvg in distribution (2/3)

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

- 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

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

Sufficient conditions for the cvg in distribution (3/3)

F., Moulines, Priouret (2012)

Assume

A. (Containment condition)
   - $\exists \pi_\theta$ s.t. $\pi_\theta P_\theta = \pi_\theta$
   - for any $\epsilon > 0$, there exists a non-decreasing positive sequence $\{r_\epsilon(n), n \geq 0\}$ such that $\limsup_{n\to\infty} r_\epsilon(n)/n = 0$ and

$$\limsup_{n\to\infty} \mathbb{E}\left[\| P^{r_\epsilon(n)}_{\theta_{n-r_\epsilon(n)}}(X_{n-r_\epsilon(n)}, \cdot) - \pi_{\theta_{n-r_\epsilon(n)}} \|_{\mathrm{tv}}\right] \leq \epsilon$$

B. (Diminishing adaptation) For any $\epsilon > 0$,

$$\lim_{n\to\infty} \sum_{j=0}^{r_\epsilon(n)-1} \mathbb{E}\left[\sup_x \| P_{\theta_{n-r_\epsilon(n)+j}}(x, \cdot) - P_{\theta_{n-r_\epsilon(n)}}(x, \cdot) \|_{\mathrm{tv}}\right] = 0$$

C. (Convergence of the invariant distributions) $(\pi_{\theta_n})_n$ converges weakly to π almost-surely.

Then for any bounded and continuous function f

$$\lim_n \mathbb{E}\left[f(X_n)\right] = \pi(f)$$

Convergence results for controlled Markov chains

The literature provides sufficient conditions for

- 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.

Conditions successfully applied to establish the convergence of Adaptive Hastings-Metropolis, (adaptive) Equi-Energy, Wang-Landau, $\cdots$

Convergence of Wang Landau (1/2)

Theorem (F., Jourdain, Kuhn, Lelièvre, Stoltz (2014))

Assume $\cdots$ *(see next slide)*
Then for any bounded measurable function f

$$\lim_t \mathbb{E}\left[f(X_t)\right] = \int f(x)\ q_\star(x)\ d\lambda(x)$$

$$\lim_T \frac{1}{T}\sum_{t=1}^{T} f(X_t) = \int f(x)\ q_\star(x)\ \mathrm{d}\lambda(x) \text{ almost-surely}$$

$$\lim_T \frac{1}{T}\sum_{t=1}^{T}\left(d\sum_{\ell=1}^{d}\mathbb{1}_{\mathbb{X}_\ell}(X_t)\theta_t(\ell)\right) f(X_t) = \int f(x)\ \pi(x)\ \mathrm{d}\lambda(x) \text{ almost-surely}$$

Convergence of Wang Landau (1/2)

Theorem (F., Jourdain, Kuhn, Lelièvre, Stoltz (2014))

Assume

1. *The target distribution* $\pi\, d\lambda$ *satisfies* $0 < \inf_{\mathbb{X}} \pi \leq \sup_{\mathbb{X}} \pi < \infty$ *and* $\inf_i \pi(\mathbb{X}_i) > 0$.
2. *For any* θ, P_θ *is a Hastings-Metropolis kernel with invariant distribution*

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

   and proposal distribution $q(x,y)d\lambda(y)$ *such that* $\inf_{\mathbb{X}^2} q > 0$.
3. *The step-size sequence is non-increasing, positive,*

$$\sum_t \gamma_t = \infty \qquad \sum_t \gamma_t^2 < \infty$$

Convergence of Wang Landau (2/2)

Sketch of proof

(1.) The containment condition:

There exist $\rho \in (0,1)$ and C such that

$$\sup_x \sup_\theta \|P_\theta^t(x,\cdot) - \pi_\theta\|_{\mathrm{TV}} \leq C\,\rho^t$$

(2.) The diminishing adaption condition:

There exists C such that for any θ,θ'

$$\sup_x \|P_\theta(x,\cdot) - P_{\theta'}(x,\cdot)\|_{\mathrm{TV}} \leq C \sum_{i=1}^{d} \left|1 - \frac{\theta(i)}{\theta'(i)}\right|$$

The update of the parameter satisfies: there exists C' such that $\forall t$

$$\|\theta_{t+1} - \theta_t\| \leq C'\,\gamma_{t+1}$$

(3.) Convergence of π_{θ_n} Requires to prove the convergence of Stochastic Approximation algorithm with controlled Markov chain dynamics.