Sampling multimodal densities in high dimensional sampling space

Gersende FORT

LTCI, CNRS & Telecom ParisTech
Paris, France

Journées MAS
Toulouse, Août 2014

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*

and in the context

- π is multimodal
- large dimension

Research guided by Computational Bayesian Statistics

π: the a posteriori distribution, known up to a normalizing constant

Needed: algorithms to explore π, to compute expectations w.r.t. π, $\cdots$.

Talk based on joint works with


Eric Moulines,
Amandine Schreck
(Telecom ParisTech)

Pierre Priouret (Paris VI)



Benjamin Jourdain,
Tony Lelièvre,
Gabriel Stoltz
(ENPC)

Estelle Kuhn (INRA)

Outline

Usual Monte Carlo samplers

1. Markov chain Monte Carlo (MCMC)
   - Sample a Markov chain $(X_k)_k$ having π as unique invariant distribution
   - Approximation:

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

   Example: Hastings-Metropolis algorithm with proposal kernel $q(x,y)$
   - given X_k, sample $Y \sim q(X_k,\cdot)$
   - accept-reject mecanism

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

Usual Monte Carlo samplers

1. Markov chain Monte Carlo (MCMC)
    - Sample a Markov chain $(X_k)_k$ having π as unique invariant distribution
    - Approximation:

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

    Example: Hastings-Metropolis algorithm with proposal kernel $q(x,y)$
    - given X_k, sample $Y \sim q(X_k, \cdot)$
    - accept-reject mecanism

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

2. Importance Sampling (IS)
    - Sample i.i.d. points $(X_k)_k$ with density q - proposal distribution chosen by the user
    - Approximation:

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

The proposal mecanism: MCMC

- Toy example in the case: Hastings-Metropolis algorithm with Gaussian proposal kernel

$$q(x,y) \propto \exp\left(-\frac{1}{2}(y-x)^T\Sigma^{-1}(y-x)\right)$$

Acceptance-rejection ratio: $1 \wedge \frac{\pi(Y)}{\pi(X_k)}$

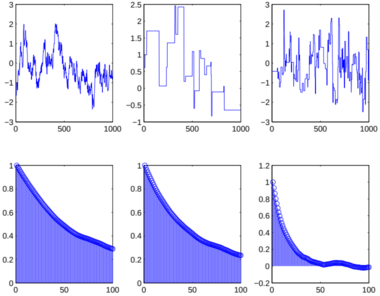


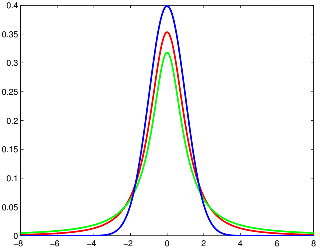
FIG.: For three different values of Σ: [top] Plot of the chain (in $\mathbb{R}$);[bottom] autocorrelation function

The proposal mecanism: Importance Sampling (1/2)

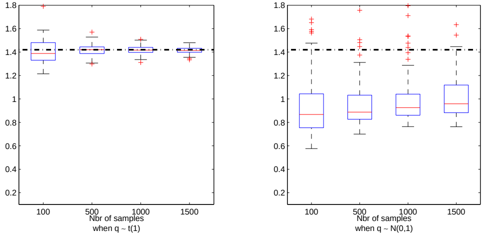
- Toy example:

$$\text{compute} \int_{\mathbb{R}} |x|\pi(x)dx \qquad \text{when} \qquad \pi(x) \sim t(3) \propto \frac{1}{(1+\frac{x^2}{3})^2}$$

Consider in turn the proposal q equal to a Student $t(1)$ and then to a Normal $\mathcal{N}(0,1)$



Plot of the densities q (green, blue) and π (in red)



Boxplot computed from 100 runs of the algorithm

The proposal mecanism: Importance Sampling (2/2)

- The efficiency of the algorithm depends upon the proposal distribution q: if few large weights and the others negligible, the approximation is likely not accurate
- Monitoring the convergence: there exist criteria measuring the proportion of “ineffective draws”:

  Coefficient of Variation
  Effective Sample Size
  Normalized perplexity

Adaptive Monte Carlo samplers

- To fix some design parameters and make the samplers more efficient: adaptive Monte Carlo samplers were proposed
- Adaptive Algorithms:

  - The *optimal* design parameters are defined as the solutions of an optimality criterion. In practice, it can not be solved explicitly.
  - Based on the past history of *the sampler*, solve an approximation of this criterion and compute the design parameters for the *current* run of the samplers.
  - Repeat the scheme: adaption/sampling.

Adaptive MC sampler: example of adaptive MCMC (1/2)

Adaptive Hastings-Metropolis algorithm with Gaussian proposal distribution

$$q_{\Sigma}(x,y) \propto \exp\left(-\frac{1}{2}(y-x)^T \Sigma^{-1}(y-x)\right)$$

- Design parameters: the covariance matrix Σ
- Optimal criterion: by using the *scaling* approach for Markov Chains, it is advocated pioneering work: Roberts, Gelman, Gilks (1997)

$$\Sigma = \frac{(2.38)^2}{\ell} \times \text{covariance of } \pi$$

- Iterative algorithm Haario, Saksman, Tamminen (2001)

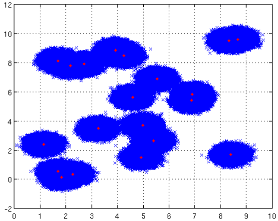
  Adaption Update the covariance matrix

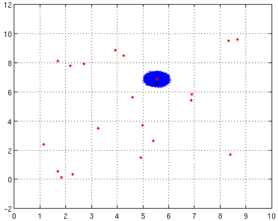
$$\Sigma_t = \frac{(2.38)^2}{\ell} \times \widehat{\Sigma}_t^{(\pi)}$$

  Sampling one step of a Hastings-Metropolis algorithm with proposal q_{Σ_t} to sample X_{t+1}.

Adaptive MC sampler: example of adaptive MCMC (2/2)

- Nevertheless, this receipe is not designed for any context.
- Example: multimodality





Target distribution: mixture of 20 Gaussian in $\mathbb{R}^2$. The means of the Gaussians are indicated with a red cross.
$5\,10^6$ i.i.d. draws

Adaptive Hastings Metropolis: $5\,10^6$ draws

Adaptive MC sampler: example of Adaptive Importance Sampling (1/2)

- Design parameter: the proposal distribution
- Optimal criterion: choose the proposal density q among a (parametric) family $\mathcal{Q}$ as the solution of

$$\mathrm{argmin}_{q \in \mathcal{Q}} \int \log \left(\frac{\pi(x)}{q(x)} \right) \; \pi(x) \, \lambda(dx) \Longleftrightarrow \mathrm{argmax}_{q \in \mathcal{Q}} \int \log q(x) \; \pi(x) \, \lambda(dx)$$

- Iterative algorithm: O. Cappé, A. Guillin, J.M. Marin, C.Robert (2004)

  Adaption Update the sampling distribution

$$q_t = \mathrm{argmax}_{q \in \mathcal{Q}} \frac{1}{n} \sum_{k=1}^{n} \log q(X_k^{(t-1)}) \; \frac{\pi(X_k^{(t-1)})}{q_{t-1}(X_k^{(t-1)})}$$

  Sampling Draw points $(X_k^{(t)})_k$ + importance reweighting

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

Adaptive MC sampler: example of Adaptive Importance Sampling (2/2)

- Nevertheless, it is known that *such* Importance Sampling techniques are not robust to the dimension: when sampling on $\mathbb{R}^\ell$ with $\ell > 15$, the degeneracy of the importance ratios

$$\frac{\pi(X_k)}{q(X_k)}$$

can not be avoided.

Conclusion

- Usual adaptive Monte Carlo samplers are not robust (enough) to the context of
    - multimodality of the target distribution π: how to jump from modes to modes.
    - large dimension of the sampling space

Importance Sampling: $\frac{\pi(x)}{q(x)}$

MCMC: $1 \wedge \frac{\pi(y)q(y,x)}{\pi(x)q(x,y)} = 1 \wedge \frac{\pi(y)}{\pi(x)}$ when q is a symetric kernel

- New Monte Carlo samplers combine

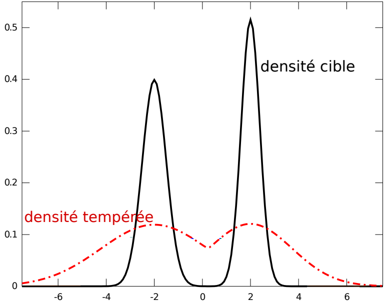
  tempering techniques and/or biasing potential techniques

  and

  sampling steps.

Outline

Tempering: the idea

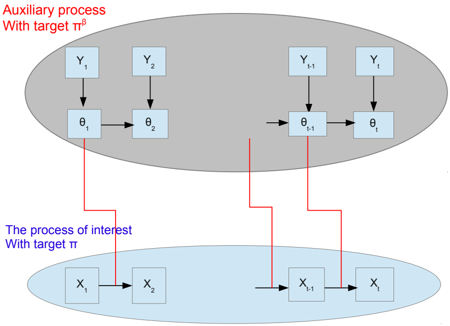


- Learn a well fitted proposal mecanism by considering tempered versions $\pi^{1/T}$ $\quad {\scriptstyle (T > 1)}$ of the target distribution π.
- Hereafter, an example where tempering is plugged in a MCMC sampler.

Example: Equi-Energy sampler (1/6)

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.



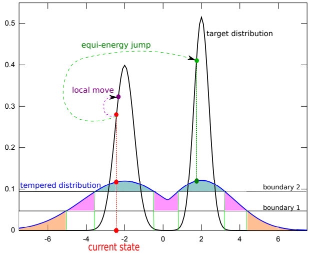
Example: Equi-Energy sampler (2/6)

Algorithm: at iteration t, given

the current state X_t

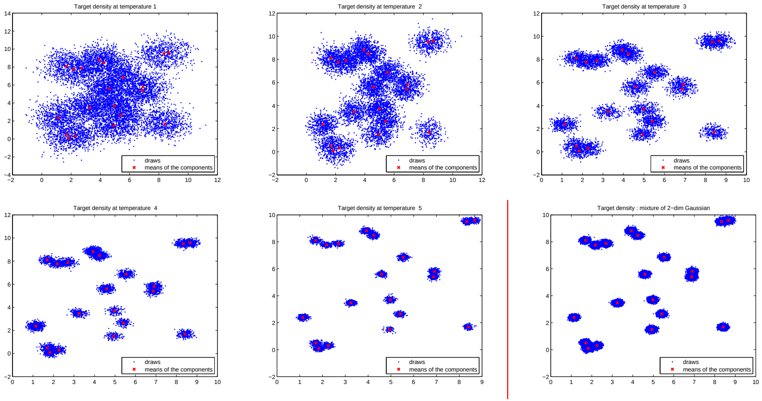
the samples $Y_1, \cdots, Y_t$ from the auxiliary process

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: Equi-Energy sampler (3/6), numerical illustration

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: Equi-Energy sampler (4/6), numerical illustration

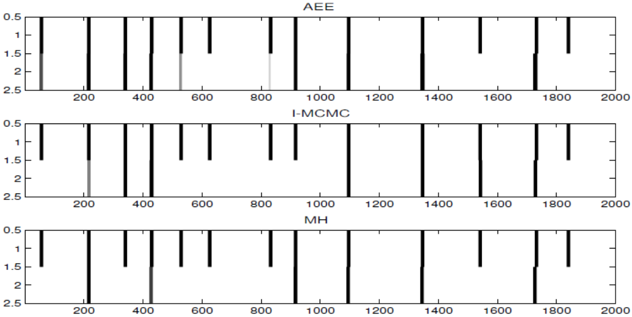
Schreck, F. and Moulines (2013)

- 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: Equi-Energy sampler (4/6), numerical illustration

Schreck, F. and Moulines (2013)

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



Example: Equi-Energy sampler (5/6), 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

`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)

Example: Equi-Energy sampler (6/6), transition kernel

- Let us describe the conditional distribution of X_{t+1} given the past:

$$\begin{aligned} P_{\theta_t}(X_t, A) &= (1-\epsilon) P(X_t, A) \\ &\quad + \epsilon \int \cdots \underbrace{g(X_t, y)\theta_t(dy)}_{\text{proposition with selection}} \end{aligned}$$

where

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

Example: Equi-Energy sampler (6/6), transition kernel

- Let us describe the conditional distribution of X_{t+1} given the past:

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

where

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

Example: Equi-Energy sampler (6/6), transition kernel

- Let us describe the conditional distribution of X_{t+1} given the past:

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

where

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

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 extension to Computational Statistics Choose a proposal distribution of the form

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

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

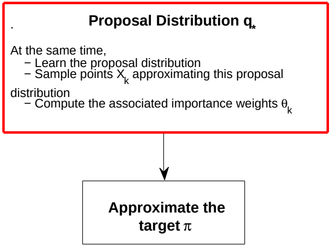
- Example:

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

Example: Wang-Landau algorithms (1/4)

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

- Wang-Landau type algorithms: learn adaptively the proposal distribution



At iteration t:

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

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

Wang-Landau algorithms (2/4)

- Key idea: $q_\star$ is obtained by *locally biasing* the target distribution

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

where

- $\mathbb{X}_1, \cdots, \mathbb{X}_d$ is a partition of the sampling space $\mathbb{X}$.
- the weights $\theta_\star(i)$ are given by

$$\theta_\star(i) = \int_{\mathbb{X}_i} \pi(x) \, dx.$$

With this biasing strategy, the proposal distribution visits each stratum $\mathbb{X}_i$ with the same frequency

$$\int_{\mathbb{X}_i} q_\star(x) \, dx = \frac{1}{d}.$$

Wang-Landau algorithms (2/4)

- Key idea: $q_\star$ is obtained by *locally biasing* the target distribution

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

where

- $\mathbb{X}_1, \cdots, \mathbb{X}_d$ is a partition of the sampling space $\mathbb{X}$.
- the weights $\theta_\star(i)$ are given by

$$\theta_\star(i) = \int_{\mathbb{X}_i} \pi(x) \; dx.$$

With this biasing strategy, the proposal distribution visits each stratum $\mathbb{X}_i$ with the same frequency

$$\int_{\mathbb{X}_i} q_\star(x) \; dx = \frac{1}{d}.$$

- Unfortunately,
  - $\theta_\star(i)$ are unknown and have to be learnt on the fly.
  - exact sampling under $q_\star$ is not possible, but it can be replaced by a MCMC step.

Wang-Landau algorithms (3/4)

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

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

2. Update the bias θ_{t+1}.

3. In parallel, update the approximation of π

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

Wang-Landau algorithms (4/4)

To learn $\theta_\star$ on the fly:

- Different 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.

Wang-Landau algorithms (4/4)

To learn $\theta_\star$ on the fly:

- Different 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.

- Two examples of updating rules:
  1. if $X_{t+1} \in \mathbb{X}_i$

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

  2.

$$S_{t+1}(j) = S_t(j) + \gamma\ \theta_t(j)\ \mathbb{1}_{\mathbb{X}_j}(X_{t+1})$$
$$\theta_{t+1}(j) = \frac{S_{t+1}(j)}{\sum_{r=1}^{d} S_{t+1}(r)}$$

Transition kernel

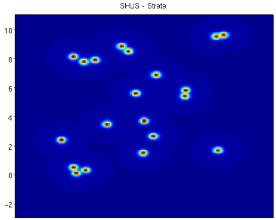
- The conditional distribution of X_{t+1} given the past is a MCMC kernel with invariant distribution q_t, denoted by P_{θ_t}
- Example: HM with Gaussian proposal distribution

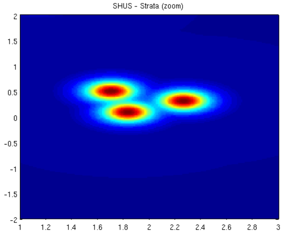
$$P_{\theta}(x,A) = \int_A \left(1 \wedge \frac{\pi(y)}{\pi(x)} \frac{\theta(\mathsf{str}(x))}{\theta(\mathsf{str}(y))}\right) \mathcal{N}(x,\Sigma)[dy] \\ + \delta_x(A) \int 1 - \left(1 \wedge \frac{\pi(y)}{\pi(x)} \frac{\theta(\mathsf{str}(x))}{\theta(\mathsf{str}(y))}\right) \mathcal{N}(x,\Sigma)[dy]$$

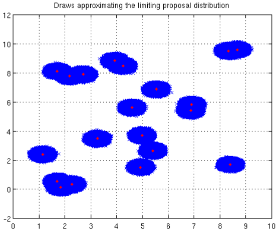
Numerical illustration, a toy example

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, obtained by partitioning the energy levels.





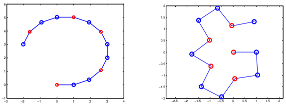


$5\,10^6$ draws approximating $q_\star$: the sampler was able to jump the deep valleys and draw points around all the modes.

Numerical illustration: 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

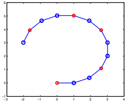
r_{ij} is the distance between monomers i,j

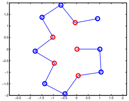
$C(\sigma_i,\sigma_j) = 1$ (resp. $1/2$ and $-1/2$) between monomers AA (resp. BB and AB).

Numerical illustration: 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?





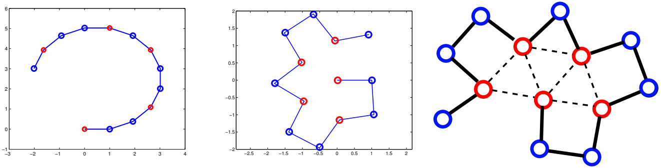
Minimize the energy function $\mathcal{H}(x)$ on

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

Numerical illustration: 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?



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

Design parameters (1/4)

- Choice of the biasing potential $A(\xi(x))$ i.e. in the Wang-Landau algorithms
  - Number of strata and the strata
  - The update strategy for the bias vector θ_t
- The MCMC kernels with target distribution q_t

`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)

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

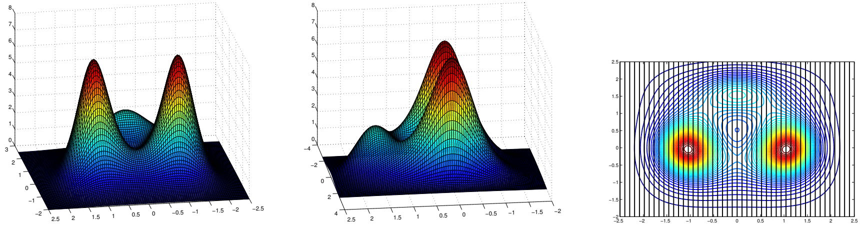
`Adaptive Wang Landau:` Bornn, Jacob, Del Moral and Doucet (2012)

Design parameters (2/4)

- Role on the limiting behavior of the sampler: convergence occurs whatever the number of strata and the strata, for many MCMC samplers and many update strategies of the bias vector.
- Role on the transient phase of the sampler: for example, how long is the exit time from a mode?

Let us illustrate the role of some design parameters on the exit time from a mode when:

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



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

Design parameters (3/4)

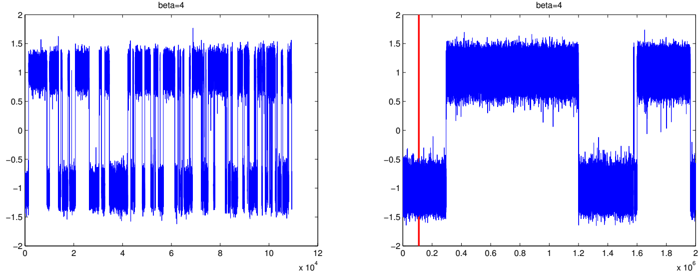


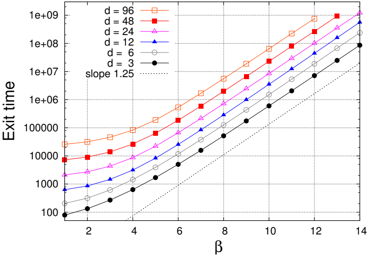
FIG.: [left] Wang Landau, $T = 110\,000$ and $d = 48$. [right] Hastings Metropolis, $T = 2\ 10^6$; the red line is at $x = 110\,000$

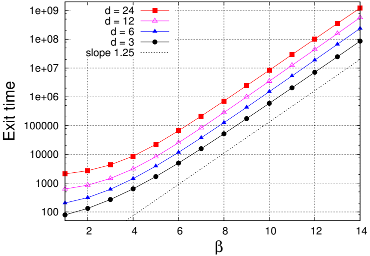
Design parameters (4/4)

F., Jourdain, Lelièvre, Stoltz (2014)

We compute the (mean) exit times t_β from the left mode (time to reach the right mode $x > 1$) 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. We observe

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





Outline

Controlled Markov chains (2/2)

- 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.

- Examples
  1. Wang Landau: the conditional distribution $X_{t+1}|\mathcal{F}_t$ is a MCMC kernel with invariant distribution distribution $q_t \propto \sum_{i=1}^d \frac{\pi(x)}{\theta_t(i)} \mathbb{I}_{\mathbb{X}_i}(x)$. Here, π_θ depends on θ and its expression is known.
  2. Equi-Energy: the conditional distribution $X_{t+1}|\mathcal{F}_t$ is a MCMC kernel indexed by the empirical distribution θ_t of the auxiliary process. Here, π_θ exists but its expression is **un**known.
  3. Adaptive Hastings-Metropolis: the conditional distribution $X_{t+1}|\mathcal{F}_t$ is a MCMC kernel with invariant distribution π and proposal distribution $\mathcal{N}(X_t, \theta_t)$ Here, all the kernels have the same invariant distribution.

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?

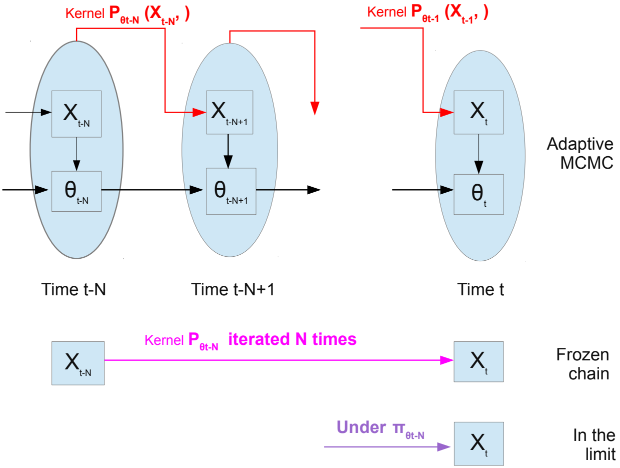
Controlled Markov chains (2/2)

- 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?

No.

- Example:
$$\begin{aligned} &\text{if } X_t = 0 \qquad X_{t+1} \sim P_0(X_t, \cdot) \\ &\text{if } X_t = 1 \qquad X_{t+1} \sim P_1(X_t, \cdot) \end{aligned}$$
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)

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

- 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

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, $\cdots$

Example: Application to Wang Landau (1/2)

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

Assume $\cdots$
Then for any bounded measurable function f

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

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

Example: Application to Wang Landau (1/2)

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

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,*

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

Example: Application to 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.