Stochastic approximation for adaptive Markov chain Monte Carlo algorithms

Gersende FORT

LTCI / CNRS - TELECOM ParisTech, France

I. Examples of adaptive and interacting MCMC samplers

1. Adaptive Hastings-Metropolis algorithm [Haario et al. 1999]
2. Wang-Landau algorithm [Wang & Landau, 2001]
3. Equi-Energy algorithm [Kou et al. 2006]

Adaptive Hastings-Metropolis algorithm

- Symmetric Random Walk Hastings-Metropolis algorithm
  - Goal: sample a Markov chain with known stationary distribution π on $\mathbb{R}^d$ (known up to a normalizing constant)
  - Iterative mecanism: given the current sample X_n,
    - propose a move to $X_n + Y \qquad Y \sim q(\cdot - X_n)$
    - accept the move with probability

$$\alpha(X_n, X_n + Y) = 1 \wedge \frac{\pi(X_n)}{\pi(X_n + Y)}$$

and set $X_{n+1} = X_n + Y$; otherwise, $X_{n+1} = X_n$.

Adaptive Hastings-Metropolis algorithm

▶ Symmetric Random Walk Hastings-Metropolis algorithm

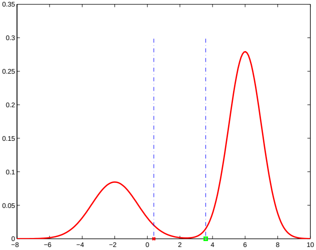
- Goal: sample a Markov chain with known stationary distribution π on $\mathbb{R}^d$ (known up to a normalizing constant)
- Iterative mecanism: given the current sample X_n,
  - propose a move to $X_n + Y \qquad Y \sim q(\cdot - X_n)$
  - accept the move with probability

$$\alpha(X_n, X_n + Y) = 1 \wedge \frac{\pi(X_n)}{\pi(X_n + Y)}$$

  and set $X_{n+1} = X_n + Y$; otherwise, $X_{n+1} = X_n$.

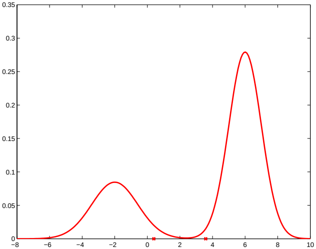
- Design parameter: how to choose the proposal distribution q ?

  For example, in the case $\quad q(\cdot - x) = \mathcal{N}_d(x; \theta) \quad$ how to scale the proposal i.e. how to choose the covariance matrix θ ?



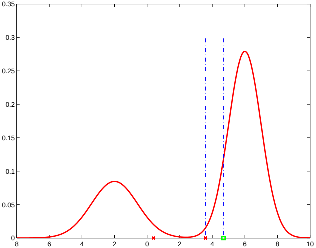
Acceptation-Rejection ratio:

$$= \begin{cases} 1 & \text{if } \pi(X_n) \leq \pi(Y + X_n) \\ \frac{\pi(Y+X_n)}{\pi(X_n)} & \text{otherwise} \end{cases}$$



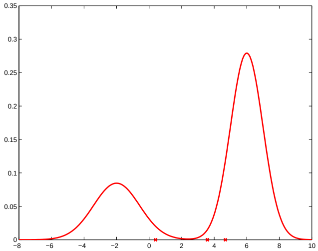
Acceptation-Rejection ratio:

$$= \begin{cases} 1 & \text{if } \pi(X_n) \leq \pi(Y + X_n) \\ \frac{\pi(Y+X_n)}{\pi(X_n)} & \text{otherwise} \end{cases}$$



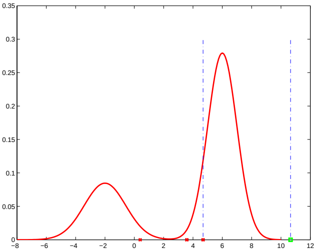
Acceptation-Rejection ratio:

$$= \begin{cases} 1 & \text{if } \pi(X_n) \leq \pi(Y + X_n) \\ \frac{\pi(Y+X_n)}{\pi(X_n)} & \text{otherwise} \end{cases}$$



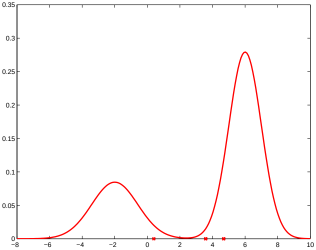
Acceptation-Rejection ratio:

$$= \begin{cases} 1 & \text{if } \pi(X_n) \leq \pi(Y + X_n) \\ \frac{\pi(Y+X_n)}{\pi(X_n)} & \text{otherwise} \end{cases}$$



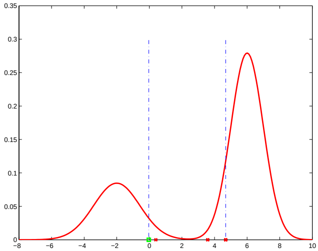
Acceptation-Rejection ratio:

$$= \begin{cases} 1 & \text{if } \pi(X_n) \leq \pi(Y + X_n) \\ \frac{\pi(Y+X_n)}{\pi(X_n)} & \text{otherwise} \end{cases}$$



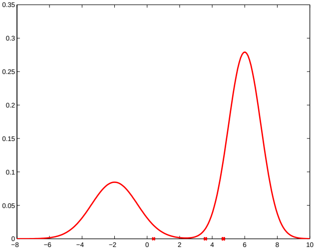
Acceptation-Rejection ratio:

$$= \begin{cases} 1 & \text{if } \pi(X_n) \leq \pi(Y + X_n) \\ \frac{\pi(Y+X_n)}{\pi(X_n)} & \text{otherwise} \end{cases}$$



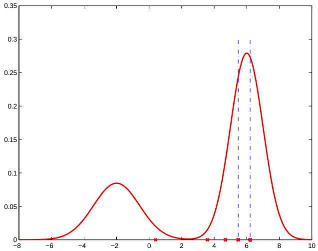
Acceptation-Rejection ratio:

$$= \begin{cases} 1 & \text{if } \pi(X_n) \leq \pi(Y + X_n) \\ \frac{\pi(Y+X_n)}{\pi(X_n)} & \text{otherwise} \end{cases}$$



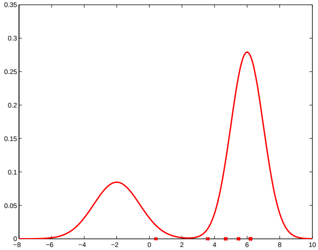
Acceptation-Rejection ratio:

$$= \begin{cases} 1 & \text{if } \pi(X_n) \leq \pi(Y + X_n) \\ \frac{\pi(Y+X_n)}{\pi(X_n)} & \text{otherwise} \end{cases}$$



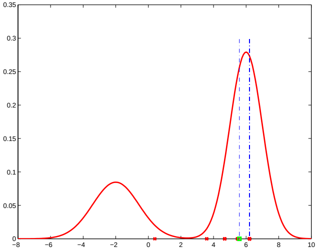
Acceptation-Rejection ratio:

$$= \begin{cases} 1 & \text{if } \pi(X_n) \leq \pi(Y + X_n) \\ \frac{\pi(Y+X_n)}{\pi(X_n)} & \text{otherwise} \end{cases}$$



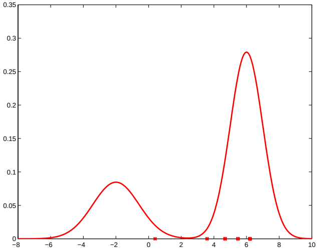
Acceptation-Rejection ratio:

$$= \begin{cases} 1 & \text{if } \pi(X_n) \leq \pi(Y + X_n) \\ \frac{\pi(Y+X_n)}{\pi(X_n)} & \text{otherwise} \end{cases}$$



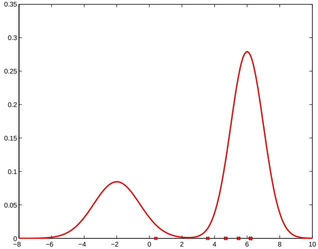
Acceptation-Rejection ratio:

$$= \begin{cases} 1 & \text{if } \pi(X_n) \leq \pi(Y + X_n) \\ \frac{\pi(Y+X_n)}{\pi(X_n)} & \text{otherwise} \end{cases}$$



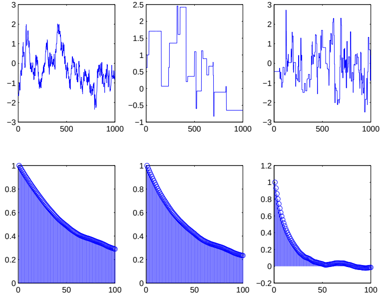
Acceptation-Rejection ratio:

$$= \begin{cases} 1 & \text{if } \pi(X_n) \leq \pi(Y + X_n) \\ \frac{\pi(Y+X_n)}{\pi(X_n)} & \text{otherwise} \end{cases}$$



Acceptation-Rejection ratio:

$$= \begin{cases} 1 & \text{if } \pi(X_n) \leq \pi(Y + X_n) \\ \frac{\pi(Y+X_n)}{\pi(X_n)} & \text{otherwise} \end{cases}$$



"goldilock principle"

Too small, too large, better variance

▶ Adaptive Hastings-Metropolis algorithm(s)

Based on theoretical results [Roberts et al. 1997; $\cdots$] when the proposal is Gaussian $\mathcal{N}_d(x, \theta)$, choose θ

- as the covariance structure of π [Haario et al. 1999]: $\theta \propto \Sigma_\pi$. In practice, Σ_π is unknown and this quantity is computed "online" with the past samples of the chain

$$\theta_{n+1} = \frac{n}{n+1}\theta_n + \frac{1}{n+1}\left\{(X_{n+1} - \mu_{n+1})(X_{n+1} - \mu_{n+1})^T + \kappa \ \mathrm{Id}_d\right\}$$

where μ_{n+1} is the empirical mean. $\kappa > 0$, prevent from badly scaled matrix

▶ Adaptive Hastings-Metropolis algorithm(s)

Based on theoretical results [Roberts et al. 1997; $\cdots$] when the proposal is Gaussian $\mathcal{N}_d(x, \theta)$, choose θ

- as the covariance structure of π [Haario et al. 1999]: $\theta \propto \Sigma_\pi$. In practice, Σ_π is unknown and this quantity is computed "online" with the past samples of the chain

$$\theta_{n+1} = \frac{n}{n+1}\theta_n + \frac{1}{n+1}\left\{(X_{n+1} - \mu_{n+1})(X_{n+1} - \mu_{n+1})^T + \kappa \; \mathrm{Id}_d\right\}$$

  where μ_{n+1} is the empirical mean. $\kappa > 0$, prevent from badly scaled matrix

- OR such that the mean acceptance rate converges to $\alpha_\star$ [Andrieu & Robert 2001]. In practice this θ is unknown and so this parameter is adapted during the run of the algorithm

$$\theta_n = \tau_n \mathrm{Id} \quad \text{with} \quad \log \tau_{n+1} = \log \tau_n + \eta_{n+1} \left(\alpha_n - \alpha_\star\right)$$

  where α_n is the mean acceptance rate.

- OR $\cdots$

▶ In practice, simultaneous adaptation of the design parameter and simulation.

Given the current value of the chain X_n and the design parameter θ_n

- Draw the next sample X_{n+1} with the transition kernel $P_{\theta_n}(X_n, \cdot)$.
- Update the design parameter: $\theta_{n+1} = \Xi_{n+1}(\theta_n, X_{n+1}, \cdot)$.

▶ In practice, simultaneous adaptation of the design parameter and simulation.

Given the current value of the chain X_n and the design parameter θ_n

- Draw the next sample X_{n+1} with the transition kernel $P_{\theta_n}(X_n, \cdot)$.
- Update the design parameter: $\theta_{n+1} = \Xi_{n+1}(\theta_n, X_{n+1}, \cdot)$.

▶ In this MCMC context, we are interested in the behavior of the chain $\{X_n, n \geq 0\}$ e.g.

- Convergence of the marginals: $\mathbb{E}\left[f(X_n)\right] \to \pi(f)$ for f bounded.
- Law of large numbers: $n^{-1} \sum_{k=1}^{n} f(X_k) \to \pi(f)$ (a.s. or $\mathbb{P}$)
- Central limit theorem

but not necessarily in the stability / convergence of the adaptation process $\{\theta_n, n \geq 0\}$.

▶ In practice, simultaneous adaptation of the design parameter and simulation.

Given the current value of the chain X_n and the design parameter θ_n

- Draw the next sample X_{n+1} with the transition kernel $P_{\theta_n}(X_n, \cdot)$.
- Update the design parameter: $\theta_{n+1} = \Xi_{n+1}(\theta_n, X_{n+1}, \cdot)$.

▶ In this MCMC context, we are interested in the behavior of the chain $\{X_n, n \geq 0\}$ e.g.

- Convergence of the marginals: $\mathbb{E}\left[f(X_n)\right] \to \pi(f)$ for f bounded.
- Law of large numbers: $n^{-1} \sum_{k=1}^{n} f(X_k) \to \pi(f)$ (a.s. or $\mathbb{P}$)
- Central limit theorem

but not necessarily in the stability / convergence of the adaptation process $\{\theta_n, n \geq 0\}$.

Note that in this example $\pi P_\theta = \pi$ for any θ: the convergence of θ_n is NOT crucial for the convergence of $\{X_n, n \geq 0\}$.

Equi-Energy sampler

▶ Proposed by Kou et al. (2006) for the simulation of multi-modal density π.

In a Hastings-Metropolis algorithm, how to choose a proposal distribution q that both allows

- local moves for a local exploration of the density.
- and large jumps in order to visit other modes of the target ?

Equi-Energy sampler

▶ Proposed by Kou et al. (2006) for the simulation of multi-modal density π.

In a Hastings-Metropolis algorithm, how to choose a proposal distribution q that both allows

- local moves for a local exploration of the density.
- and large jumps in order to visit other modes of the target ?

▶ Idea: (a) build an auxiliary process that moves between the modes far more easily and (b) define the process of interest

- by running a "classical" Hastings-Metropolis algorithm
- and sometimes, choose a value of the auxiliary process as the new value of the process of interest: draw a point at random + acceptation-rejection mecanism

Equi-Energy sampler

▶ Proposed by Kou et al. (2006) for the simulation of multi-modal density π.

In a Hastings-Metropolis algorithm, how to choose a proposal distribution q that both allows

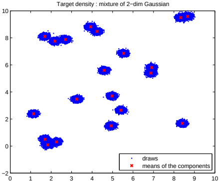
- local moves for a local exploration of the density.
- and large jumps in order to visit other modes of the target ?

▶ Idea: (a) build an auxiliary process that moves between the modes far more easily and (b) define the process of interest

- by running a "classical" Hastings-Metropolis algorithm
- and sometimes, choose a value of the auxiliary process as the new value of the process of interest: draw a point at random + acceptation-rejection mecanism

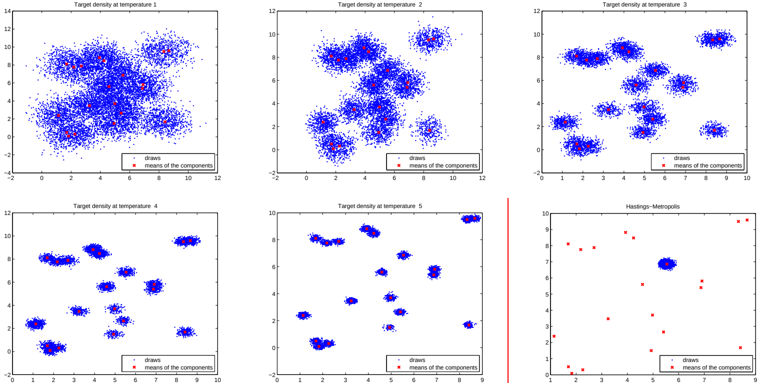
How to define such an auxiliary process ? Ans.: as a process with stationary distribution π^{β} $(\beta \in (0,1))$, a tempered version of the target π.

▶ On an example: a K-stage Equi-Energy sampler.



- target density: $\pi = \sum_{i=1}^{20} \mathcal{N}_2(\mu_i, \Sigma_i)$
- K auxiliary processes: with targets π^{1/T_i}

  $T_1 > T_2 > \cdots > T_{K+1} = 1$



- An example of interacting MCMC (2 stages)

Repeat:

- Update the adaptation process

$$\theta_n = \frac{1}{n} \sum_{k=0}^{n-1} \delta_{Y_k}$$

  where $\{Y_n, n \geq 0\}$ is the auxiliary process with stationary distribution π^β.

- Update the process of interest with transition : $X_{n+1} \sim P_{\theta_n}(X_n, \cdot)$ where

$$P_{\theta_n}(x, A) = (1-\epsilon)P(x, A) + \epsilon \left\{ \int_A \underbrace{\alpha(x, y)}_{\text{accept/reject mecanism}} \theta_n(dy) + \delta_x(A) \int (1 - \alpha(x, y)) \theta_n(dy) \right\}$$

- An example of interacting MCMC (2 stages)

Repeat:

- Update the adaptation process

$$\theta_n = \frac{1}{n} \sum_{k=0}^{n-1} \delta_{Y_k}$$

where $\{Y_n, n \geq 0\}$ is the auxiliary process with stationary distribution π^β.

- Update the process of interest with transition : $X_{n+1} \sim P_{\theta_n}(X_n, \cdot)$ where

$$P_{\theta_n}(x, A) = (1-\epsilon)P(x, A) + \epsilon \left\{ \int_A \underbrace{\alpha(x, y)}_{\text{accept/reject mecanism}} \theta_n(dy) + \delta_x(A) \int (1 - \alpha(x, y))\theta_n(dy) \right\}$$

P_θ is such that when $\theta_n \propto \pi^\beta$, $\pi P_{\pi^\beta} = \pi$: asymptotically, when θ_n "is" π^β, the process of interest $\{X_n, n \geq 0\}$ behaves like a Markov chain with invariant distribution π.

▶ An example of interacting MCMC (2 stages)

Repeat:

- Update the adaptation process

$$\theta_n = \frac{1}{n} \sum_{k=0}^{n-1} \delta_{Y_k}$$

where $\{Y_n, n \geq 0\}$ is the auxiliary process with stationary distribution π^β.

- Update the process of interest with transition : $X_{n+1} \sim P_{\theta_n}(X_n, \cdot)$ where

$$P_{\theta_n}(x, A) = (1-\epsilon)P(x, A) + \epsilon \left\{ \int_A \underbrace{\alpha(x, y)}_{\text{accept/reject mecanism}} \theta_n(dy) + \delta_x(A) \int (1 - \alpha(x, y))\theta_n(dy) \right\}$$

P_θ is such that when $\theta_n \propto \pi^\beta$, $\pi P_{\pi^\beta} = \pi$: asymptotically, when θ_n "is" π^β, the process of interest $\{X_n, n \geq 0\}$ behaves like a Markov chain with invariant distribution π.

In this MCMC context, we are again interested in the behavior of $\{X_n, n \geq 0\}$ but convergence of θ_n is crucial since the algorithm is designed to "sample from" π only when $\theta_n = \pi^\beta$.

▶ Proposed by Wang & Landau ( ) to favor the moves between elements of a partition of the state space, when the weights of these elements is unknown.

- Context:
  - Partition $\{\mathsf{X}_i, i \leq d\}$ of the state space X.
  - $\theta_\star(i) \stackrel{\text{def}}{=} \int_{\mathsf{X}_i} \pi(x) dx$ is unknown.

▶ Proposed by Wang & Landau ( ) to favor the moves between elements of a partition of the state space, when the weights of these elements is unknown.

- Context:
  - Partition $\{\mathsf{X}_i, i \leq d\}$ of the state space X.
  - $\theta_\star(i) \stackrel{\mathrm{def}}{=} \int_{\mathsf{X}_i} \pi(x) dx$ is unknown.

- Goal:
  - build a chain on $\prod_{i=1}^d (\mathsf{X}_i \times \{i\})$ with stationary distribution

$$\Pi(A_i \times \{i\}) = \frac{1}{d} \int_{A_i} \frac{\pi(x)}{\theta_\star(i)} \mathbb{1}_{\mathsf{X}_i}(x)\ dx\ ,$$

  - and/ or estimate the normalizing constants $\theta_\star(i)$.

▶ Proposed by Wang & Landau ( ) to favor the moves between elements of a partition of the state space, when the weights of these elements is unknown.

- Context:
  - Partition $\{\mathsf{X}_i, i \leq d\}$ of the state space X.
  - $\theta_\star(i) \stackrel{\mathrm{def}}{=} \int_{\mathsf{X}_i} \pi(x) dx$ is unknown.

- Goal:
  - build a chain on $\prod_{i=1}^d (\mathsf{X}_i \times \{i\})$ with stationary distribution
$$\Pi(A_i \times \{i\}) = \frac{1}{d} \int_{A_i} \frac{\pi(x)}{\theta_\star(i)} \mathbb{1}_{\mathsf{X}_i}(x) \; dx \;,$$
  - and/ or estimate the normalizing constants $\theta_\star(i)$.

- Tool :
  - A family of transition kernels P_θ on $\prod_{i=1}^d (\mathsf{X}_i \times \{i\})$
  - where $\theta = (\theta(1), \cdots, \theta(d))$ is a probability on $\{1, \cdots, d\}$
  - with invariant distribution known up to a normalizing constant
$$\Pi_\theta(A_i \times \{i\}) = \left(\sum_{j=1}^d \frac{\theta_\star(j)}{\theta(j)} \right)^{-1} \int_{A_i} \frac{\pi(x)}{\theta(i)} \mathbb{1}_{\mathsf{X}_i}(x) \; dx \;,$$

- Algorithm: repeat
  - Draw $(X_{n+1}, I_{n+1}) \sim P_{\theta_n}((X_n, I_n), \cdot)$
  - Update the adaptation process

$$\theta_{n+1}(i) \propto \theta_n(i) + \gamma_{n+1}\theta_n(i)\mathbb{1}_{I_{n+1}}(i)$$

- Algorithm: repeat
  - Draw $(X_{n+1}, I_{n+1}) \sim P_{\theta_n}((X_n, I_n), \cdot)$
  - Update the adaptation process

$$\theta_{n+1}(i) \propto \theta_n(i) + \gamma_{n+1}\theta_n(i)\mathbb{1}_{I_{n+1}}(i)$$

- In this MCMC context, we are also interested in the convergence of the sequence $\{\theta_n, n \geq 0\}$: at a first order,

$$\theta_{n+1}(i) \approx \theta_n(i) + \gamma_{n+1}\theta_n(i)\left(\mathbb{1}_{I_{n+1}}(i) - \theta_n(I_{n+1})\right)$$

and when $(X_n, I_n) \sim \Pi_{\theta_n}$

$$\mathbb{E}\left[\theta_n(i)\left(\mathbb{1}_{I_{n+1}}(i) - \theta_n(I_{n+1})\right)|\mathcal{F}_n\right] = XXXX$$

i.e. $\{\theta_n, n \geq 0\}$ should converge to $\theta_\star$!

Conclusion

In adaptive MCMC,

- given a family of transition kernels $\{P_\theta, \theta \in \Theta\}$
- with invariant distribution π_θ

we define a bivariate process $\{(X_n, \theta_n), n \geq 0\}$ such that

$$\mathbb{P}\left(X_{n+1} \in \cdot | \mathcal{F}_n\right) = P_{\theta_n}(X_n, \cdot)$$

What kind of conditions on the adaptation mecanism, for the convergence of the process $\{X_n, n \geq 0\}$ to a target distribution π ?

In the sequel, "convergence" means " convergence of the marginals"

$$\mathbb{E}\left[f(X_n)\right] \to \pi(f) \qquad f \text{ bounded}$$