Adaptive Equi-Energy Sampler : Convergence and Illustration

Amandine Schreck, Gersende Fort and Eric Moulines

Telecom Paristech

April 11th 2012

- Goal : sample a target distribution π known up to a multiplicative constant
- Example : motif sampling in biology
- Problem : for multimodal distributions, some algorithms remain trapped in one of the modes

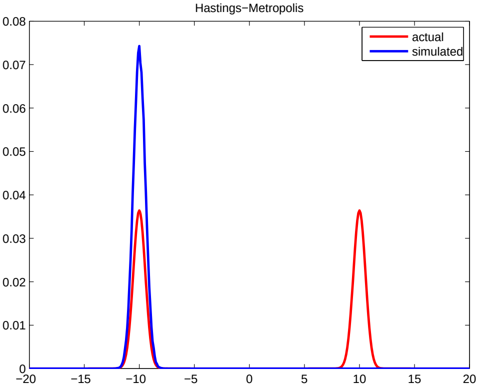


Figure: Random walk Metropolis-Hastings for a mixture of Gaussian distributions

Metropolis-Hastings algorithm :

- Sample X_0 under any initial distribution μ
- Knowing the current state X_n, sample Y_{n+1} under $Q(X_n, .)$
- Compute the acceptation-rejection probability : $\alpha(X_n, Y_{n+1}) = \min\left(1, \frac{\pi(Y_{n+1})q(Y_{n+1}, X_n)}{\pi(X_n)q(X_n, Y_{n+1})}\right)$
- Set $X_{n+1} = Y_{n+1}$ with probability $\alpha(X_n, Y_{n+1})$ and $X_{n+1} = X_n$ with probability $1 - \alpha(X_n, Y_{n+1})$.

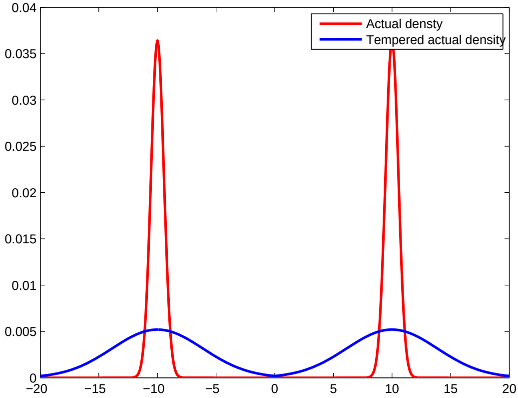


Figure: Actual density and a tempered version ($T = 50$)

- It seems easier to sample a tempered version $\pi^{1/T}$, $T > 1$ of the target distribution.
- Idea : Sample a tempered version of the target distribution as an auxiliary process and allow the process of interest to “jump” on one of the auxiliary states after and acceptation/rejection step.
- Problem : The acceptation probability could be really low.

Equi-Energy Sampler :

- Sample X_0 under any initial distribution μ.
- We know n values $Y_1, \ldots, Y_n$ of an auxiliary process. Knowing the current state X_n :
  - with probability $1 - \epsilon$, sample X_{n+1} with a symmetric random walk Metropolis-Hastings algorithm
  - with probability ϵ, choose an auxiliay value Y_i such that $\pi(Y_i)$ is "close" to $\pi(X_n)$, and set $X_{n+1} = Y_i$ or $X_{n+1} = X_n$ after an acceptation/rejection step

Fix a number of rings S. Consider a sequence of real number $\xi_0 = 0 < \xi_1 < \cdots < \xi_S = +\infty$.
Two energies $\pi(x)$ and $\pi(y)$ are said to be close if there exists l, $1 \leq l \leq S$ such that $\xi_{l-1} \leq \pi(x), \pi(y) < \xi_l$.
On the choice of the ξ_i :

- Original equi-energy sampler : fixed by user
- Problem : crucial choice
- Adaptive equi-energy sampler : quantiles estimators
  - empirical quantiles
  - stochastic approximation estimator

Empirical quantiles associated to a distribution θ :

- Cumulative distribution function : $F_\theta(x) = \int \mathbf{1}_{\{\pi(y) \leq x\}} \theta(dy)$.
- Quantile function : $F_\theta^{-1}(p) = \inf\{x \geq 0, F_\theta(x) \geq p\}$.
- For any $\{p_l, 1 \leq l \leq S\}$ (for example $p_l = \frac{l}{S}$), the ring boudaries are defined by $\hat{\xi}_{\theta,l} = F_\theta^{-1}(p_l)$.
- Rings : $A_{\theta,l} =]\hat{\xi}_{\theta,l-1}; \hat{\xi}_{\theta,l}]$.

For the adaptive EES : $\theta_n = n^{-1} \sum_{k=1}^{n} \delta_{Y_k}$.

- Selection function : $g_\theta(x,y) = \sum_{l=1}^{S} h_{\theta,l}(x) h_{\theta,l}(y)$,
- with : $h_{\theta,l}(x) = \left(1 - \frac{d(\pi(x), A_{\theta,l})}{r}\right)_+$.
- Kernel for the EE move : $K_\theta(x,A) =$ $\int_A \alpha_\theta(x,y) \frac{g_\theta(x,y)\theta(dy)}{\int g_\theta(x,z)\theta(dz)} + \mathbf{1}_A(x) \int \{1 - \alpha_\theta(x,y)\} \frac{g_\theta(x,y)\theta(dy)}{\int g_\theta(x,z)\theta(dz)}$,
- with : $\alpha_\theta(x,y) = 1 \wedge \left(\frac{\pi(y)}{\pi(x)} \frac{\pi^{1-\beta}(x) \int g_\theta(x,z)\theta(dz)}{\pi^{1-\beta}(y) \int g_\theta(y,z)\theta(dz)} \right)$.
- Kernel for the AEE sampler : $P_\theta(x,.) = (1-\epsilon) P(x,.) + \epsilon K_\theta(x,.)$.

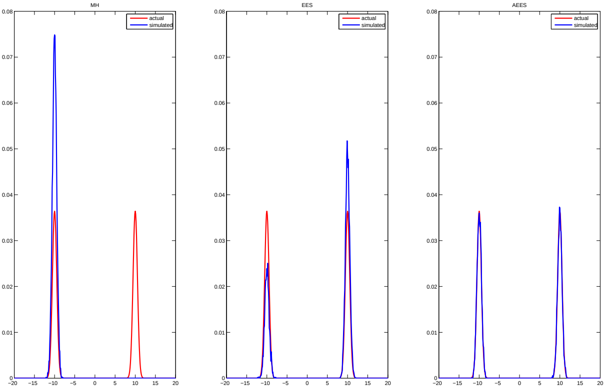


Figure: Equi-Energy Samplers for a mixture of Gaussian distributions

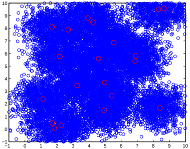


Figure: EES for a mixture of Gaussian distributions, T=60

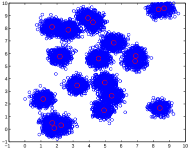


Figure: T=7

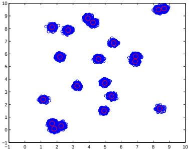


Figure: T=1

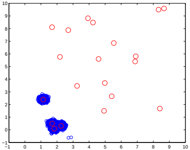


Figure: Metropolis-Hastings

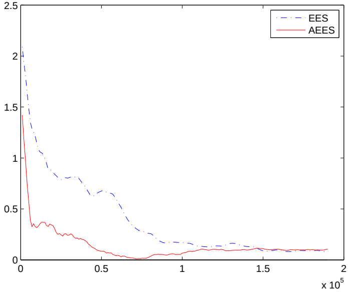


Figure: L1 error for EES and AEES

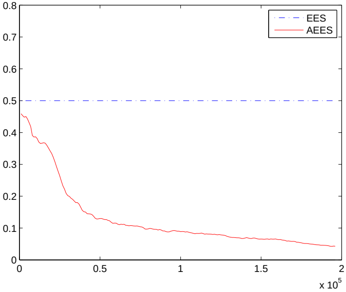
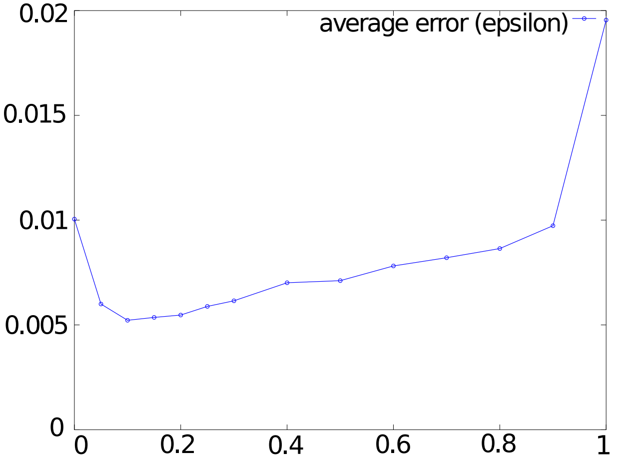


Figure: extreme case

Many parameters to choose :

- proposal distribution (could be adaptive)
- number of energy rings
- temperature of the processes
- proportion of equi-energy moves


average error (epsilon)
0.02
0.015
0.01
0.005
0
0
0.2
0.4
0.6
0.8
1

Notations :

- L : length of the DNA sequence
- S : DNA sequence. $S = (s_1, s_2, \ldots, s_L)$ with $s_i \in \{1, 2, 3, 4\}$ (1 corresponding to A, 2 to C, 3 to G and 4 to T)
- w : length of a motif
- A : array giving the position of the motifs. $A = (a_1, \ldots, a_L)$, where a_i is equal to $j \in \{0, \ldots, w\}$ if the ith element of the sequence is the jth element of a motif
- p_0 : probability for a sub-sequence of length w to be a motif

Distribution :

- Background sequence : Markov chain associated with the transition matrix denoted by θ_0
- Motif : multinomial distribution of parameter $\theta = (\theta_1, \dots, \theta_w)$

Knowing $a_1, \dots, a_{k-1}, s_1, \dots, s_{k-1}, \theta$ and p_0, we have :

- If $a_{k-1} \in \{1, \dots, w-1\}$, $a_k = a_{k-1} + 1$, otherwise, a_k follows a Bernouilli distribution of parameter p_0
- If $a_k = 0$, s_k follows the distribution $\theta_0(s_{k-1}, .)$, otherwise, s_k follows the distribution $\theta_{a_k}(.)$

Conditionnal distribution of A given S :

$$P(A|S) \propto \frac{\Gamma(N_1 + a)\Gamma(N_0 + b)}{\Gamma(N_1 + N_0 + a + b)} \prod_{i=1}^{w} \frac{\prod_{j=1}^{4} \Gamma(c_{i,j} + \beta_{i,j})}{\Gamma(\sum_{j=1}^{4} c_{i,j} + \beta_{i,j})}$$
$$\prod_{k=2}^{L} (\delta_{a_{k-1}+1}(a_k))^{\mathbf{1}_{a_{k-1} \in \{1,\dots,w-1\}}} \prod_{k=2}^{L} \theta_0^{1-\bar{A}_k}(s_{k-1}, s_k) \xi_{a_1}(s_1)$$

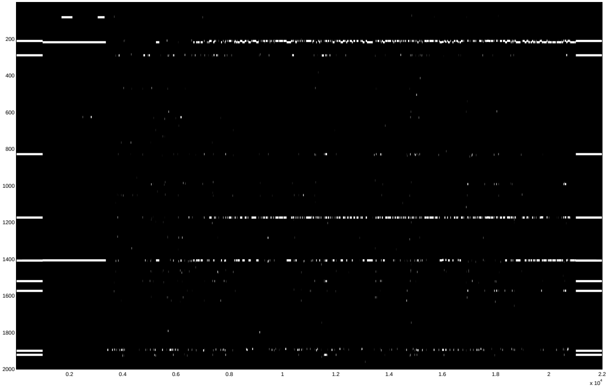


Figure: Location of the motifs retrieved by AEES at each iteration

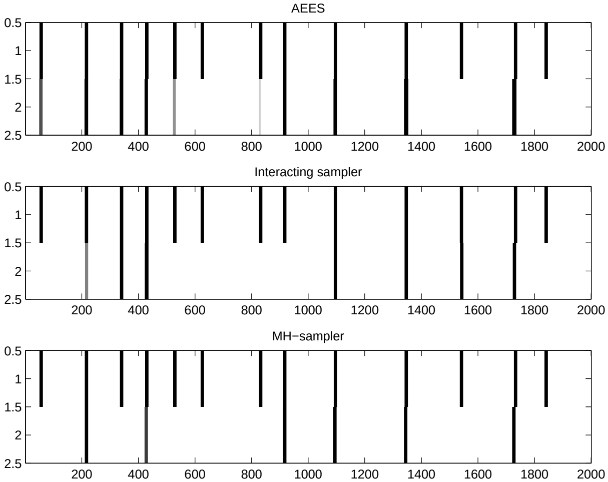


Figure: Average location of the motifs - comparison of 3 algorithms

- If $g = 1$ and $\theta = \pi^{1-\beta}$, K_θ is a Metropolis-hastings kernel with π as stationary distribution.
- If θ_n converges toward $\pi^{1-\beta}$, we expect P_{θ_n} to converge toward $P_{\pi^{1-\beta}}$ and (X_n) to converge toward π, invariant distribution of $P_{\pi^{1-\beta}}$

A few notations :

- V-norm of a function f :$|f[_V = \sup_{x\in \mathbf{X}} \frac{|f(x)|}{V(x)}$
- V-norm of a signed measure μ : $\|\mu\|_V = \sup_{f,|f|_V \leq 1} |\mu(f)|$
- We define the V-variation between P_θ and $P_{\theta'}$ by $D_V(\theta,\theta') = \sup_{x\in \mathbf{X}} \left(\frac{\|P_\theta(x,.) - P_{\theta'}(x,.)\|_V}{V(x)} \right)$
- Set $\mathcal{L}_V$: $\mathcal{L}_V = \{f : \mathbf{X} \to \mathbb{R}, \|f\|_V < +\infty\}$
- Target density : π
- Temperature of the auxiliary process $T = \frac{1}{1-\beta}$

The adaptive EE sampler generates a bivariate process (X_n, θ_n) $(\mathcal{F}_n)$-adapted for the filtration $(\mathcal{F}_n) = \sigma(Y_1, \dots, Y_n, X_1, \dots, X_n)$, and such that :

$$\mathbb{E}\left[f(X_{n+1}) | \mathcal{F}_n\right] = P_{\theta_n} f(X_n)$$

Condition on π :

(a) π is the density of a probability distribution on the measurable Polish space $(\mathbf{X}, \mathcal{X})$ and $\sup_{\mathbf{X}} \pi < \infty$.

(b) π is continuous on $\mathbf{X}$.

Condition on the proposal distribution P :

(a) P is a ϕ-irreducible transition kernel which is Feller on $(\mathbf{X}, \mathcal{X})$ and such that $\pi P = \pi$.

(b) (drift) There exist $\lambda \in (0,1)$, $b < +\infty$ and $\tau \in (0, 1-\beta)$ such that $PW \leq \lambda W + b$ with

$$W(x) = \left(\frac{\pi(x)}{\sup_{\mathbf{X}} \pi} \right)^{-\tau} \quad . \tag{1}$$

(c) (small) For all $p \in (0, \sup_{\mathbf{X}} \pi)$, the sets $\{\pi \geq p\}$ are 1-small for P.

Condition on the auxiliary process

(a) $\theta_\star(W) < +\infty$, and for all continuous function f in $\mathcal{L}_W$, $\theta_n(f) \to \theta_\star(f)$ a.s.

(b) $\sup_n \mathbb{E}\left[W\left(Y_n\right)\right] < \infty$.

where $\theta_\star$ is the density proportionnal to $\pi^{1/T}$.

With these conditions, we prove the "convergence" of our adaptation (only for a 2-level algorithm for the moment) :

(a) For any $l \in \{1, \ldots, S-1\}$, $\lim_n \left|\xi_{\theta_n,l} - \xi_{\theta_\star,l}\right| = 0$, w.p.1

(b) There exists $\Gamma > 0$ such that for any $k \in \{1, \ldots, K-1\}$, for any $l \in \{1, \ldots, S-1\}$, and any $\gamma < \Gamma$,

$$\limsup_n \; n^\gamma \; \left|\xi_{\theta_{n+1},l} - \xi_{\theta_n,l}\right| < \infty \quad , \mathbb{P} - \text{a.s.}$$

We also prove that :

- For all $n \in \mathbb{N}$, the kernel P_{θ_n} admits a finite stationnary distribution π_{θ_n}
- For all $n \in \mathbb{N}$, there exist some random variables C_{θ_n} and ρ_{θ_n} such that for all $x \in \mathbf{X}$:

$$\|P_{\theta_n}^k(x,.) - \pi_{\theta_n}\|_V \leq C_{\theta_n}\rho_{\theta_n}^k V(x)$$

Finally, this allow to control the V-variation between P_θ and $P_{\theta'}$: on the set $\bigcap_j \{\theta_j \in \Theta_m\}$, where

$$\Theta_m = \left\{ \theta \in \Theta : \frac{1}{m} \leq \inf_x \int g_\theta(x, y)\theta(\mathrm{d}y) \right\} \ ,$$

there exists a constant C_m such that

$$D_V(\theta_k, \theta_{k-1}) \\ \leq C_m \left(\sup_l \left| \xi_{\theta_k, l} - \xi_{\theta_{k-1}, l} \right| + \|\theta_k - \theta_{k-1}\|_{\mathrm{TV}} \right) (\|\theta_k\|_V + \|\theta_{k-1}\|_V) \\ + C_m \ \|\theta_k - \theta_{k-1}\|_V \ .$$

Convergence of the stationnary distributions :

$$\begin{aligned}\left|\pi_{\theta_n(x)}(f) - \pi_{\theta_\star(w)}(f)\right| \leq & \left|\pi_{\theta_n(w)}(f) - P^k_{\theta_n(w)}f(x)\right| \\ & + \left|P^k_{\theta_n(w)}f(x) - P^k_{\theta_\star(w)}f(x)\right| \\ & + \left|P^k_{\theta_\star(w)}f(x) - \pi_{\theta_\star}(f)\right|\end{aligned}$$

Control :

- Terms 1 and 3 : controled with $\|P^k_\theta(x,.) - \pi_\theta\|_V \leq C_\theta \rho^k_\theta V(x) \quad \mathbb{P}\text{-ps}$
- Term 2 : weak convergence of the kernels P_{θ_n} toward $P_{\theta_\star}$, and equi-continuity of these kernels

Ergodicity :

$$\begin{aligned} |\mathbb{E}[f(X_n)] - \pi(f)| \leq & \left|\mathbb{E}\left[f(X_n) - P^N_{\theta_{n-N}} f(X_{n-N})\right]\right| \\ & + \left|\mathbb{E}\left[P^N_{\theta_{n-N}} f(X_{n-N}) - \pi_{\theta_{n-N}}(f)\right]\right| \\ & + \left|\mathbb{E}\left[\pi_{\theta_{n-N}}(f) - \pi(f)\right]\right| \end{aligned}$$

Control :

- Term 1 : sum of some $D_V(\theta_{n+j}, \theta_{n+j-1})$
- Term 2 : controled with $\|P^k_\theta(x,.) - \pi_\theta\|_V \leq C_\theta \rho^k_\theta V(x)$ $\mathbb{P}$-ps
- Terme 3 : convergence of the stationnary distributions

Strong law of large numbers : The idea is to introduce the solution $\hat{f}_\theta$ of the Poisson equation

$$\hat{f}_\theta - P_\theta \hat{f}_\theta = f - \pi_\theta(f)$$

to isolate a martingale term.

$$\frac{1}{n}\sum_{k=0}^{n-1} f(X_k) - L = T_{1,n} + T_{2,n} + T_{3,n} + T_{4,n} + T_{5,n}$$

$$T_{1,n} = 1/n(f(X_0) - L)$$

$$T_{2,n} = \frac{1}{n}\sum_{k=1}^{n-1}\{\hat{f}_{\theta_{k-1}}(X_k) - P_{\theta_{k-1}}\hat{f}_{\theta_{k-1}}(X_{k-1})\}$$

$$T_{3,n} = \frac{1}{n}\sum_{k=1}^{n-1}\{P_{\theta_k}\hat{f}_{\theta_k}(X_k) - P_{\theta_{k-1}}\hat{f}_{\theta_{k-1}}(X_k)\}$$

$$T_{4,n} = \frac{1}{n}P_{\theta_0}\hat{f}_{\theta_0}(X_0) - \frac{1}{n}P_{\theta_{n-1}}\hat{f}_{\theta_{n-1}}(X_{n-1})$$

$$T_{5,n} = \frac{1}{n}\sum_{k=0}^{n-2}\{\pi_{\theta_{k-1}}(f) - L\}$$

Term $T_{2,n}$:

$$T_{2,n} = \frac{1}{n}\sum_{k=1}^{n-1}\{\hat{f}_{\theta_{k-1}}(X_k) - P_{\theta_{k-1}}\hat{f}_{\theta_{k-1}}(X_{k-1})\}$$

$T_{2,n}$ is a sum of martingale increments. We control it by showing that there exists $\alpha > 1$ such that $\sum_{k=1}^{\infty} k^{-\alpha}\mathbb{E}\left[\left|\{\hat{f}_{\theta_{k-1}}(X_k) - P_{\theta_{k-1}}\hat{f}_{\theta_{k-1}}(X_{k-1})\right|^{\alpha}\middle|\mathcal{F}_{k-1}\right] < \infty$ as

Term $T_{3,n}$:

$$T_{3,n} = \frac{1}{n} \sum_{k=1}^{n-1} \{ P_{\theta_k} \hat{f}_{\theta_k}(X_k) - P_{\theta_{k-1}} \hat{f}_{\theta_{k-1}}(X_k) \}$$

is caused by the adaptation. To control it, we show that $n^{-1} \sum_{k=1}^{n} D_V(\theta_k, \theta_{k-1}) V(X_k) \to 0$ almost surely.

In practice :

- Far more efficient than Metropolis-Hastings (mix better)
- Does not require the user to choose the rings

But :

- A lot of parameters to choose
- Quite high computational cost

To go further :

- Extend results of convergence for the empirical quantiles
- Central limit theorem ?
- Adaptive proposal