An adaptive version of the equi-energy sampler

Amandine Schreck, Gersende Fort and Eric Moulines

Télécom ParisTech

April 10th 2014

- Goal : sample a target distribution Π (known up to a multiplicative constant)
- Problem : for multimodal distributions, some algorithms remain **trapped** in one of the modes

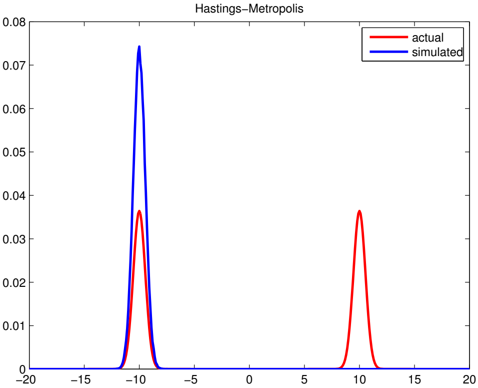

Figure : Random walk Metropolis-Hastings for a mixture of Gaussian distributions

1. The algorithm
   - Why interact ?
   - The adaptive equi-energy sampler
2. On the convergence of AEE
   - Control of each kernel
   - Make the control uniform
   - Convergence results
   - On the choice of the energy levels
3. Illustrations
   - Toy example
   - Motif sampling
   - Shape of proteins

The algorithm

Base : Metropolis Hastings algorithm (with dominating measure $\mathrm{d}\mu$)

- Goal : sample a distribution $\Pi = \pi \mathrm{d}\mu$ known **up to a multiplicative constant**.
- Tool : a transition kernel q such that for any x, it is **possible to sample from** $q(x, \cdot)\mathrm{d}\mu$.
- An iteration starting from X^t :
  - Sample Y^{t+1} according to $q(X^t, \cdot)\mathrm{d}\mu$.
  - Compute the acceptance probability

$$\alpha(X^t, Y^{t+1}) = \min\left(1, \frac{\pi(Y^{t+1})q(Y^{t+1}, X^t)}{\pi(X^t)q(X^t, Y^{t+1})}\right) .$$

  - Set $X^{t+1} = Y^{t+1}$ with probability $\alpha(X^t, Y^{t+1})$ and $X^{t+1} = X^t$ with probability $1 - \alpha(X^t, Y^{t+1})$.

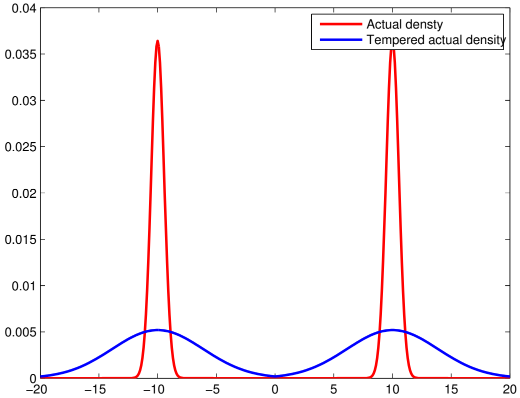

Figure : Actual density and a tempered version ($T = 50$)

- It seems easier to sample a tempered version $\theta_\star$ with density $\pi^{1/T}$, $T > 1$ 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 sampled **auxiliary** states after and acceptance/rejection step.
- **Problem** : The acceptance probability could be really low.

Equi-Energy Sampler [Kou, Zhou and Wong, 2006] :

- 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 acceptance/rejection step

What "close" is

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
- Our **adaptive equi-energy sampler** : quantile estimators
  - empirical quantiles
  - stochastic approximation estimators

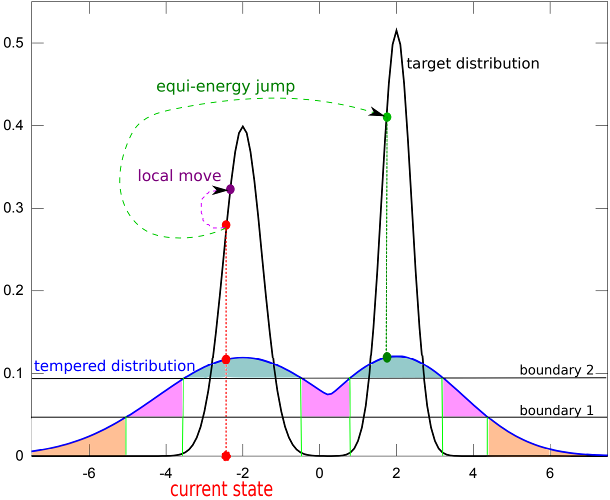

Figure : the equi-energy samplers

Method 1 : Empirical quantiles associated to a distribution θ :

- Cumulative distribution function : $F_\theta(x) = \int \mathbf{1}_{\{\pi(y) \leq x\}} \theta(dy)$.
- Quantile function : $F_\theta^{-1}(p) = \inf\{x \geq 0, F_\theta(x) \geq p\}$.
- $\boldsymbol{\xi_{\theta,l} = F_\theta^{-1}(l/S)}$.

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

Method 2 : Stochastic approximation procedure with a stepsize sequence $(\gamma_n)_{n\in\mathbb{N}}$:

$$\xi_{n,l} = \xi_{n-1,l} + \gamma_n \left(l/S - 1_{\{\pi(Y_n)\leq\xi_{n-1,l}\}} \right) .$$

Theoretical results

On the convergence of AEE

The adaptation can destroy the convergence results. But for the adaptation made with empirical quantiles,

- under classical conditions on the target density, the local moves and the auxiliary process,
- under **conditions on the energy levels**,
- introducing the solution of the **Poisson equation** to obtain a martingale term,

we show the ergodicity and a strong LLN for AEE and EE.

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

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

Conditions 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 $\mathbf{X}$.

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 : $\|\mu\|_V = \sup_{f,|f|_V \leq 1} |\mu(f)|$
- V-variation between P_θ and $P_{\theta'}$ by $D_V(\theta, \theta') = \sup_{x \in \mathbf{X}} \left(\frac{\|P_\theta(x,.) - P_{\theta'}(x,.)\|_V}{V(x)} \right)$

Control of each kernel

Conditions on the proposal distribution P :

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

(b) **Drift inequality :** 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 sets :** for all $p \in (0, \sup_{\mathbf{X}} \pi)$, the sets $\{\pi \geq p\}$ are 1-small for P.

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

where g_θ is the selection function.

Extension of previous assumptions :

(a) For all θ, P_θ is **phi-irreducible** and for all $p \in (0, \sup_{\mathsf{X}} \pi)$, the sets $\{\pi \geq p\}$ are **1-small** for P_θ.

(b) **Drift inequality :** there exist $\tilde{\lambda} \in (0,1)$, $\tilde{b} < +\infty$ such that for all $m \geq 1$ and $\theta \in \Theta_m$,

$$P_\theta W \leq \tilde{\lambda} W + \tilde{b}\; m\; \theta(W) .$$

Corollary :

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

$$\|P_{\theta_n}^k(x,.) - \Pi_{\theta_n}\|_W \leq C_{\theta_n}\rho_{\theta_n}^k W(x)$$

Make the control uniform

Conditions on the auxiliary process :

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

(b) **Law of large numbers :** $\theta_\star(W) < +\infty$, and for all continuous function f such that $|f| \leq CW$, $\theta_n(f) \to \theta_\star(f)$ a.s.

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

Control on the measure of the rings :

If the energy levels converge, there exists $m_\star$ such that

$$\mathbb{P}\left(\bigcup_{q\geq 1}\bigcap_{n\geq q}\{\theta_n \in \Theta_{m_\star}\}\right) = 1\,.$$

$\rightarrow$ similar constants for all P_{θ_n} in geometric drift and geometric ergodicity (containment condition).

Control on the measure of the fluctuations between kernels :

There exists a constant C such that, on the set $\bigcap_{j\geq q}\{\theta_n \in \Theta_{m_\star}\}$,

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

$\rightarrow$ with some conditions on $\sup_l \left|\xi_{\theta_k,l} - \xi_{\theta_{k-1},l}\right|$: **diminishing adaptation**.

Conditions required on the energy boundaries :

(a) **Convergence :** 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) **Stability :** there exists $\Gamma > 0$ such that for any $k \in \{1, \ldots, K-1\}$, for any $l \in \{1, \ldots, S-1\}$, and any $\gamma < \Gamma$,

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

Convergence results

Convergence of the invariant 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 $\|P^k_\theta(x,.) - \Pi_\theta\|_V \leq C_\theta \rho^k_\theta V(x)$ $\mathbb{P}$-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 $\|P^k_\theta(x,.) - \Pi_\theta\|_V \leq C_\theta \rho^k_\theta V(x) \quad \mathbb{P}\text{-ps}$
- Terme 3 : convergence of the invariant distributions

Strong law of large numbers :

$$
\begin{aligned}
&\frac{1}{n}\sum_{k=1}^{n}[f(X_k)-\Pi(f)] \\
&=\frac{1}{n}\sum_{k=1}^{n}\left[f(X_k)-\Pi_{\theta_{k-1}}(f)\right]+\frac{1}{n}\sum_{k=1}^{n}\left[\Pi_{\theta_{k-1}}(f)-\Pi(f)\right]
\end{aligned}
$$

Control of term 2 : convergence of the invariant distributions + Cesaro

For $\frac{1}{n}\sum_{k=1}^{n}\left[f(X_k) - \Pi_{\theta_{k-1}}(f)\right]$:

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.

$$\frac{1}{n}\sum_{k=1}^{n}\left[f(X_k) - \Pi_{\theta_{k-1}}(f)\right] = T_{1,n} + T_{2,n} + T_{3,n}$$

$$T_{1,n} = \text{remainder term}$$

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

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

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 **martingale**. 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 kernel fluctuation. Controled with results of **diminishing adaptation**.

On the choice of the energy levels

Empirical quantiles :

We show with a **Hoeffding inequality** on non stationary Markov chains that the conditions are satisfied.

Stochastic approximation :

- $\left|\xi_{n+1,l} - \xi_{n,l}\right| \leq \gamma_{n+1} \rightarrow$ **stability assumption satisfied** for well-chosen stepsize sequence.
- Problem for convergence : stochastic approximation scheme

$$\xi_{n+1} = \xi_n + \gamma_{n+1} \; H(\xi_n, Y_{n+1}) \; ,$$

with **discontinuous field**

$$H(\xi, y) = q - \mathbf{1}_{\{\pi(y) \leq \xi\}} \; .$$

If the auxiliary process is a Markov chain, we show that

$$\xi_n \underset{n\to\infty}{\longrightarrow} \xi_\star \quad \text{where} \quad \mathbb{E}_{\theta_\star}[H(\xi_\star, Y)] = 0 \; .$$

Regularity condition used on H : There exists $\alpha \in (0, 1]$, and for all compact set $\mathcal{K} \subseteq \mathbb{R}$, there exists $C > 0$ such that for all $\delta > 0$,

$$\sup_{\xi \in \mathcal{K}} \int \sup_{\{\xi', |\xi'-\xi| \leq \delta\}} \left| H(\xi', x) - H(\xi, x) \right| \pi^{1/T}(x) \mathrm{d}x \leq C\delta^\alpha \; .$$

$\to$ Need to be extended to chains with external randomness.

Illustrations

Toy example

Toy example

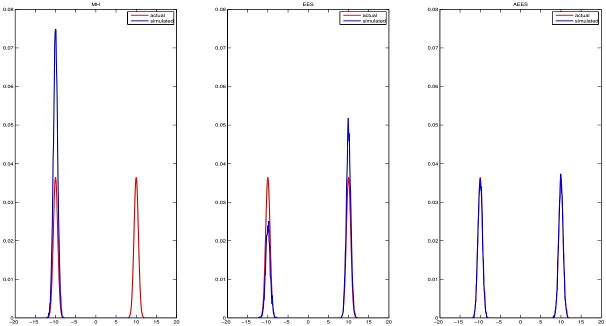

Figure : [left] Metropolis-Hastings [center] equi-energy sampler [right] adaptive equi-energy sampler for a mixture of Gaussian distributions

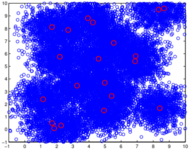

Figure : EES for a mixture of Gaussian distributions, T=60

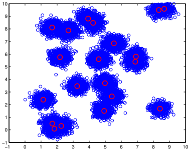

Figure : T=7

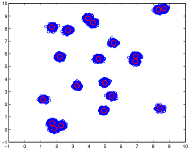

Figure : T=1

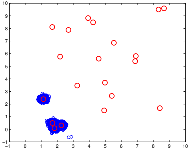

Figure : Metropolis-Hastings

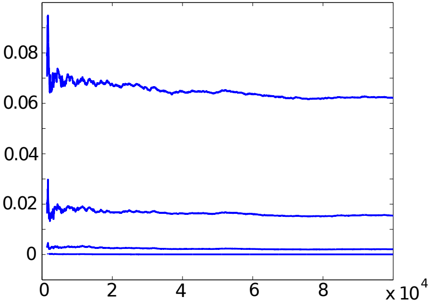

Figure : Evolution of the ring bounds

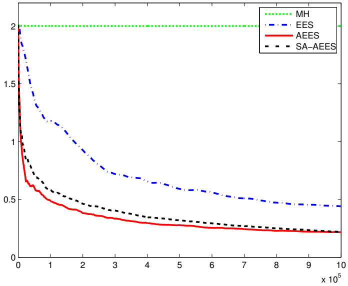

Figure : 1-norm of the empirical mean for EES and AEES in 6 dimensions

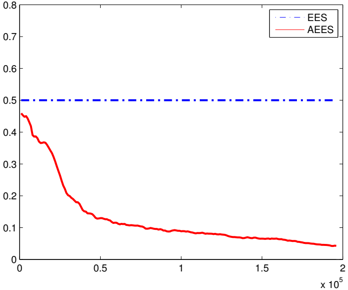

Figure : extreme case (ring bounds badly defined)

Motif sampling

Motif sampling

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 $\theta = (\theta_1, \dots, \theta_w)$

This gives the distribution of A knowing S, θ, θ_0 and p_0. We then put a prior on θ and p_0, and study the distribution of A knowing S and θ_0.

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

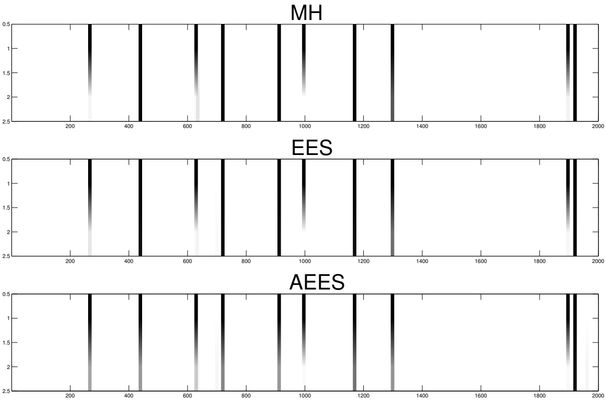

Figure : Average location of the motifs - comparison of 3 algorithms

Shape of proteins

A protein of size M is described by

- A **monomer sequence** $S = [S_1, S_2, \ldots, S_M] \in \{0, 1\}^M$, with $S_i = 0$ (resp. $S_i = 1$) if the i-th monomer is hydrophobic (resp. hydrophilic).
- A sequence of **angles** $X = [X_1, X_2, \ldots, X_{M-2}]$.

- **Energy** of a protein :

$$U(X) = \sum_{i=1}^{M-2} \frac{1}{4}(1 - \cos(X_i)) + 4 \sum_{i=1}^{M-2} \sum_{j=i+2}^{M} \left[d_{i,j}^{-12} - C(S_i, S_j) d_{i,j}^{-6} \right] ,$$

  where $C(0,0) = 1$, $C(0,1) = C(1,0) = -1/2$ and $C(1,1) = 1/2$.

- Goal : sample the distribution with density $\pi(x) \propto \exp(-U(x)/\tau)$, with $\tau > 0$ small (multimodalty).

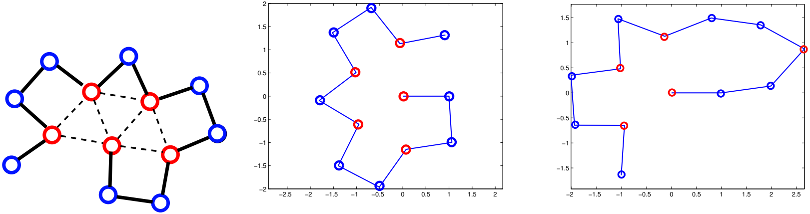

Figure : (left) actual shape, (center) global energy minimum, (right) local energy minimum

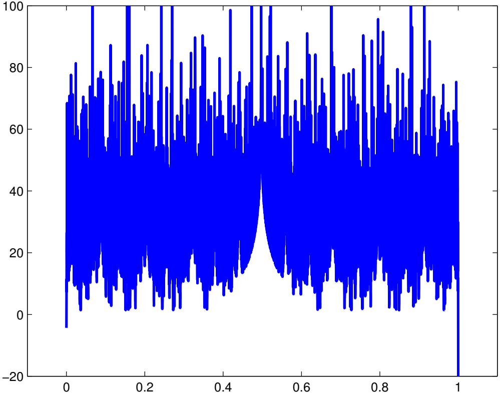

Figure : log of energy on a segment

	AEES	WL
Mean minimum energy	-3.1970	-3.1632
Associated standard deviation	0.032	0.048
Global minimum energy	-3.2925	-3.2764

Conclusion

In practice :

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

But :

- Higher computational cost than the non-adaptive algorithm
- Still a lot of parameters to choose

To go further :

- Study the effects of the design parameters
- Combine with adaptive proposal
- Central limit theorem ?

Thank you !

URL of the paper : http ://arxiv.org/abs/1207.0662

- 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 : $P_\theta(x, .) = (1 - \epsilon)P(x, .) + \epsilon K_\theta(x, .)$.