Adaptive Equi-Energy samplers

Gersende FORT

LTCI
CNRS & Telecom ParisTech
Paris, France

Based on joint works with

- A. Schreck, A. Garivier and E. Moulines - from Telecom ParisTech, France.
- P. Priouret - from Univ. Paris VI, France.

A talk in two steps:

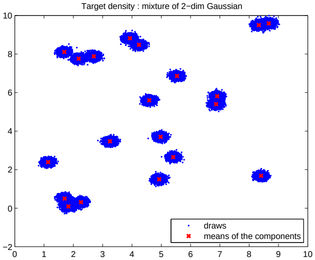
1. Equi-Energy sampler by Kou, Zhou, Wong (2006), an example of interacting MCMC sampler.
2. Adaptive Equi-Energy sampler: algorithm and convergence results.

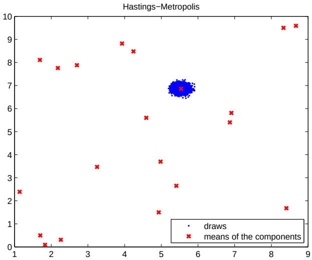
Algorithms designed to:

obtain samples from a chain with target $\propto \pi$ when π is multimodal.

MCMC with a multimodal target density fails:

Ex. target density : $\pi = \sum_{i=1}^{20} \mathcal{N}_2(\mu_i, \Sigma_i)$





[left] i.i.d. samples under π [right] Symmetric Random Walk HM

The **Equi-Energy sampler** Kou et al (2006) is an example of Interacting Tempering algorithm:

- "*interacting*": many chains are run in parallel and interactions between these chains are allowed.
- "*tempering*": each parallel chain designed so that its target is $\propto \pi^{1/T_k}$, $(T_1 > T_2 > \cdots > T_K = 1)$

Equi-Energy sampler Kou et al (2006)

- Will define $X^{(t)} = \{X_n^{(t)}, n \geq 0\}$ with
  - $X^{(1)}$ - high temperature, target : π^{1/T_1}
  - $\cdots$
  - $X^{(K)}$ - coolest one, target : $\pi^{1/T_K} = \pi$

Equi-Energy sampler Kou et al (2006)

- Will define $X^{(t)} = \{X_n^{(t)}, n \geq 0\}$ with
  - $X^{(1)}$ - high temperature, target : π^{1/T_1}
  - $\cdots$
  - $X^{(K)}$ - coolest one, target : $\pi^{1/T_K} = \pi$

- `Algorithm`: given

  points of the previous level $X_{1:n}^{(k-1)}$

  and the current point $X_{n-1}^{(k)}$

  define $X_n^{(k)}$ as follows:
  - with probability $1-\epsilon$, (MCMC step)

    $$X_n^{(k)} \sim P^{(k)}(X_{n-1}^{(k)}, \cdot) \qquad \text{with } P^{(k)} \text{ s.t.} \qquad \pi^{1/T_k} P^{(k)} = \pi^{1/T_k}$$

  - with probability ϵ, (Interaction step): interaction with $X_{1:n}^{(k-1)}$

Equi-energy jumps: the interaction step

- Interaction $\equiv$ choose a point among $X_{1:n}^{(k-1)}$ as the next value for $X^{(k)}$.
- The interaction with the previous level will favor jumps to other modes since the previous process has a high-tempered target distribution
- But $X^{(k-1)}$ and $X^{(k)}$ do not have the same target distribution: introduction of an acceptance rejection step The acceptance-rejection ratio is given by

$$\alpha(x,y) = 1 \wedge \frac{\pi^{1/T_k}(y)}{\pi^{1/T_k}(x)} \frac{\pi^{1/T_{k-1}}(x)}{\pi^{1/T_{k-1}}(y)} = 1 \wedge \left(\frac{\pi(y)}{\pi(x)}\right)^{\frac{1}{T_k} - \frac{1}{T_{k-1}}}$$

Equi-energy jumps: the interaction step

- Interaction $\equiv$ choose a point among $X_{1:n}^{(k-1)}$ as the next value for $X^{(k)}$.
- The interaction with the previous level will favor jumps to other modes since the previous process has a high-tempered target distribution
- But $X^{(k-1)}$ and $X^{(k)}$ do not have the same target distribution: introduction of an acceptance rejection step
- To make this ratio close to 1, select a point among $X_{1:n}^{(k-1)}$, with density close to $\pi(X_{n-1}^{(k)})$

  or equivalently, with an energy $\mathcal{H}$ close enough to the energy of $X_{n-1}^{(k)}$

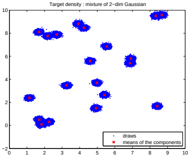
$$\mathcal{H}(x) = -\log \pi(x).$$

In the Kou et al. algorithm, **before running the algorithm**, **fix** a partition of the energy space

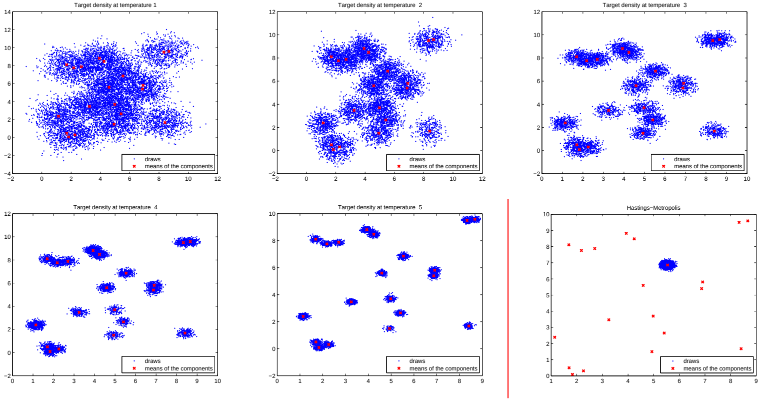
$$\text{Energy Ring } \#i \quad = \quad \{x, \mathcal{H}(x) \in [H_{i-1}, H_i]\}$$

and choose a point in the same energy level.

EE on an example



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



Design parameters

Before running the algorithm, fix

- the number of parallel processes and the set of temperatures .
- the number of strata in the energy space and the boundaries of the strata
- the probability of interaction ϵ.

Despite many convergence analysis (on EE with no selection)

- ergodicity: $\lim_n \mathbb{E}[h(X_n^{(K)})] = \pi(h)$
- law of large numbers: $\frac{1}{n} \sum_{j=1}^{n} h(X_j^{(K)}) \to \pi(h)$ in $\mathbb{P}$ or a.s.
- CLT: $\sqrt{n}^{-1} \sum_{j=1}^{n} \{h(X_j^{(K)}) - \pi(h)\} \to_{\mathcal{D}} \mathcal{N}(0,\sigma^2)$

see e.g. Kou, Zhou, Wong (2006); Atchadé (2010); Andrieu, Jasra, Doucet, Del Moral (2011); Fort, Moulines, Priouret (2012); Fort, Moulines, Priouret, Vandekerkhove (2012) how to fix these design parameters is an open question.

Our contribution: tune adaptively the boundaries of the strata

II. Adaptive Equi-Energy sampler

Algorithm and Convergence results

Adaptive tuning of the boundaries of the energy rings

$\hookrightarrow$ How to define the boundaries $H_1, \cdots, H_L$ of the energy rings?

Our approach: the energy rings used in the mecanism "$X^{(k-1)} \to X^{(k)}$" are defined by using the samples of level $X^{(k-1)}$.

Algorithm

- Level 1 (Hot level)
  - Draw $X^{(1)}$ with target π^{1/T_1} (MCMC).
  - at *each time* n, update the boundaries $H^{(1)}_{n,1}, \cdots, H^{(1)}_{n,L}$ computed from $X^{(1)}_{1:n}$
- Level 2
  - Draw $X^{(2)}$ with target π^{1/T_2} . For the interaction step, use the boundaries $H^{(1)}_{\bullet}$.
  - at *each time* n, update the boundaries $H^{(2)}_{n,1}, \cdots, H^{(2)}_{n,L}$ computed from $X^{(2)}_{1:n}$
- Repeat until Level K.

Example of adaptive boundaries (1/2)

- We are able to prove convergence properties for A-EE when, w.p.1

  the proportion of points of the auxiliary process in each stratum, is uniformly for all large n lower bounded.

- Idea: when n is large, $\mathcal{L}(X_n^{(k)}) \approx \pi^{1/T_k}$. Therefore, choose $H_i^{(k)}$ for $1 \leq i \leq L$ as the quantiles of order $i/(L+1)$ of the distribution of

$$\mathcal{H}(Z) = -\log \pi(Z) \qquad \text{when } Z \sim \pi^{1/T_k}$$

- In practice: choose $H_{n,i}^{(k)}$ for $1 \leq i \leq L$ as an estimator of the quantiles of order $i/(L+1)$ of the distribution of

$$\mathcal{H}(Z) = -\log \pi(Z) \qquad \text{when } Z \sim \pi^{1/T_k}$$

  computed from $X_{1:n}^{(k)}$.

Example of adaptive boundaries (2/2)

1) A first estimator, is based on the inversion of the empirical cdf

$$F_n^{(k)}(h) = \frac{1}{n} \sum_{j=1}^{n} 1_{\pi(X_j^{(k)}) \leq h}$$

2) A second one is based on Stochastic Approximation procedures

$$H_{n+1,\bullet}^{(k)} = H_{n,\bullet}^{(k)} + \gamma_{n+1} \; \Xi \left(X_{n+1}^{(k)}, H_{n,\bullet}^{(k)} \right)$$

Num. Appl.: fixed boundaries vs adapted boundaries

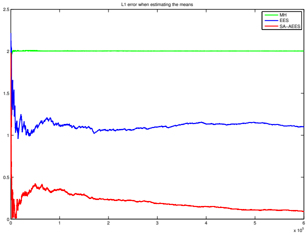
- Target distribution on $\mathbb{R}^6$

$$\pi = \frac{1}{2}\mathcal{N}_6\left(\mu, 0.3 \text{ Id}\right) + \frac{1}{2}\mathcal{N}_6\left(-\mu, 0.2 \text{ Id}\right) \qquad \mu = [2, \cdots, 2]$$

- We compare Hastings-Metropolis (HM); and the EE sampler and the Adaptive EE sampler when applied with 3 temperatures and 11 strata.

- The last plot is for the 2-d projection $\left(u^T X; v^T X\right)$ with $u^T \propto [1,1,\cdots,1]$ $v^T \propto [1,1,1,-1,-1,-1]$

Behavior along one path: HM EE A-EE

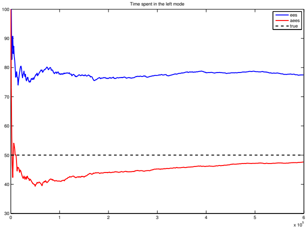


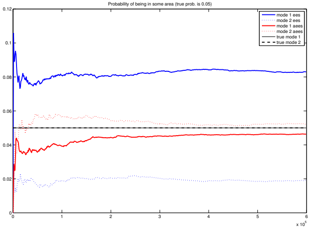
[Top] Error when estimating the means

$$\frac{1}{6}\sum_{i=1}^{6}\left|\frac{1}{n}\sum_{j=1}^{n}X_{j,i}^{(K)}-\mathbb{E}_{\pi}[X_i]\right|$$

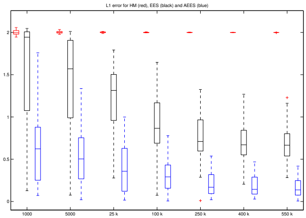
[Bottom L] Time spent in one of the mode where the path is initialized.

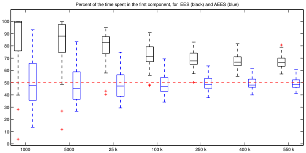
[Bottom R] Probability of being in some ellipsoids, for the first mode (line) and the second one (dashed line)





Behavior on 50 ind. run HM EE A-EE



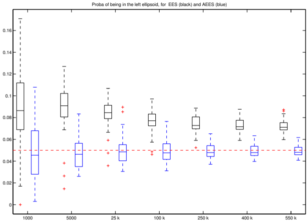


[Top] Error when estimating the means

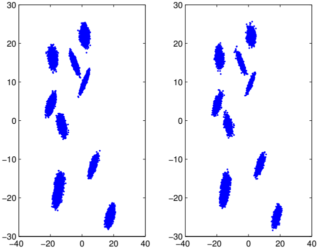
$$\frac{1}{6}\sum_{i=1}^{6}\left|\frac{1}{n}\sum_{j=1}^{n}X_{j,i}^{(K)}-\mathbb{E}_{\pi}[X_i]\right|$$

[Bottom L] Time spent in one of the mode where the path is initialized.

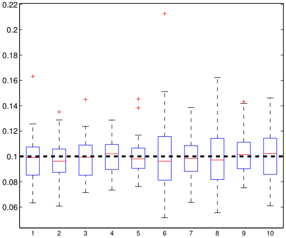
[Bottom R] Probability of being in some ellipsoid, for the first mode



Num. Appl.: Adaptive EE



[left] True density (mixture of Gaussian, same weights);
[right] Adaptive EE



Frequency of the visit to each component of the mixture. Boxplot with 50 ind. run

Convergence results for A-EE (1/3)

- It is known that adaptation can destroy convergence !
- The successive processes $X^{(k)}$ are NOT markovian except the first one $X^{(1)}$ but are controlled Markov chains.
- Sufficient conditions for convergence of such processes are Roberts & Rosenthal (2007); Atchadé, F. (2011, 2012), Atchadé, F., Moulines, Priouret (2011), F., Moulines, Priouret (2012)
  - (geometric) ergodicity of each transition kernel $P_\theta^{(k)}$.
  - containment condition
  - diminishing adaptation

Convergence results for A-EE (2/3)

We prove ergodicity and law of large numbers under the assumptions:

1. π is continuous, positive on the measurable Polish space (X,$\mathcal{B}$(X)) and $\int \pi^s(x)dx < \infty$ for any $s \in (0,1)$.

2. a set of conditions on each MCMC kernel $P^{(k)}$, implying $P^{(k)}$ is W_k-*Geometrically ergodic*

Convergence results for A-EE (2/3)

We prove ergodicity and law of large numbers under the assumptions:

1. π is continuous, positive on the measurable Polish space $(\mathsf{X},\mathcal{B}(\mathsf{X}))$ and $\int \pi^s(x)dx < \infty$ for any $s \in (0,1)$.

2. a set of conditions on each MCMC kernel $P^{(k)}$, implying $P^{(k)}$ is W_k*-Geometrically ergodic*

3. Condition on the adapted boundaries: for any $k \in \{1,\ldots,K-1\}$,
   (a) There exists $\gamma_\star > 0$ such that for any $\ell \in \{1,\cdots,S-1\}$ and any $\gamma \in (0,\gamma_\star)$,
   $$\limsup_n \; n^\gamma \; \left| H^{(k)}_{n+1,\ell} - H^{(k)}_{n,\ell} \right| < \infty \qquad \text{w.p.1}$$
   (b) for any $\ell \in \{1,\cdots,S-1\}$,
   $$\lim_{n\to\infty} \left| H^{(k)}_{n,\ell} - H^{(k)}_{\infty,\ell} \right| = 0 \qquad \text{w.p.1}$$
   (c)
   $$\inf_{\ell\in\{1,\cdots,S-1\}} \int \mathbb{1}_{\{x, H^{(k)}_{\infty,\ell-1} \leq \mathcal{H}(x) \leq H^{(k)}_{\infty,\ell}\}} \; \pi^{1/T_k}(dx) > 0.$$

Convergence results for A-EE (3/3)

Theorem: Under

- the previous assumptions,
- and $\mathbb{E}[W_k(X_0^{(k)})] < \infty$ for all $k \in \{1, \cdots, K\}$,

it holds for any $k \in \{1, \cdots, K\}$

- (Ergodicity) and for all bounded continuous function $f : \mathsf{X} \to \mathbb{R}$,

$$\lim_{n\to\infty} \mathbb{E}[f(X_n^{(k)})] = \frac{\pi^{1/T_k}(f)}{\pi^{1/T_k}(1)}.$$

- (strong LLN) Let $a \in (0, \frac{1+\gamma_\star}{2} \wedge 1)$. For all continuous function f s.t. $\sup_x |f|/W_k^a < \infty$,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{m=1}^{n} f(X_m^{(k)}) = \frac{\pi^{1/T_k}(f)}{\pi^{1/T_k}(1)} \qquad \mathbb{P}-\text{a.s.}.$$

III. Conclusion

Conclusion

1. A new adaptive interacting algorithm: methodology and convergence results. First results on convergence of Equi-Energy sampler with selection mecanism.

2. Convergence results for interacting Monte Carlo algorithms.

3. Application to a non trivial example (motif discovery in DNA sequences)
   not shown here, but available in the paper

Paper available:
Adaptive Equi-Energy Sampler : Convergence and Illustration, A. Schreck, G.F. and E. Moulines. *ACM : Transactions on Modeling and Computer Simulation*, 2012.