Parallel tempering and Interacting MCMC algorithms

Gersende FORT / Eric MOULINES

Telecom Paris Tech
CNRS - LTCI

Part II: Adaptive Equi-Energy samplers

Joint work with

Amandine Schreck

Aurélien Garivier and Eric Moulines

from LTCI, Telecom ParisTech & CNRS, France.

From Parallel Tempering to Interacting Tempering

- The **Equi Energy sampler** Kou et al (2006) is an example of **Interacting Tempering** algorithm.
- The idea is to replace an **instantaneous swap** by an **interaction** with the whole past of a **neighboring** process on the **temperature ladder**.

From Parallel Tempering to Interacting Tempering

- The **Equi Energy sampler** Kou et al (2006) is an example of **Interacting Tempering** algorithm.
- The idea is to replace an **instantaneous swap** by an **interaction** with the whole past of a **neighboring** process on the **temperature ladder**.

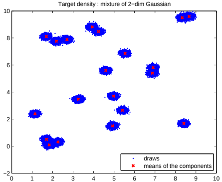
Equi-Energy sampler **Kou et al (2006)**

- Will define $X^{(t)} = \{X_n^{(t)}, n \geq 0\}$ with $X^{(1)}$ (hot temperature), $\cdots$, $X^{(K)}$ target process.
- `Algorithm:` given the previous level $X_{1:n-1}^{(k-1)}$ and the current point $X_{n-1}^{(k)}$, define $X_n^{(k)}$ as follows:
  - `(MCMC step / local moves)` with probability ϵ,

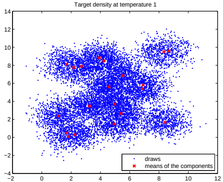
$$X_n^{(k)} \sim P^{(k)}(X_{n-1}^{(k)}, \cdot) \qquad \text{with } P^{(k)} \text{ s.t.} \qquad \pi^{(k)} P^{(k)} = \pi^{(k)}$$

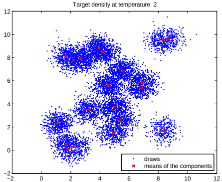
  - `(Interaction step / global moves)` otherwise,
    - (i) selection of a point $X_\bullet^{(k-1)}$ among the set $\{X_{1:n-1}^{(k-1)}\}$ with **the same energy level** as $X_{n-1}^{(k)}$
    - (ii) acceptance-rejection ratio.

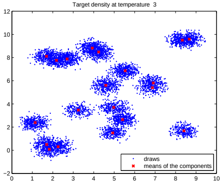
Numerical application: on the interest of EE

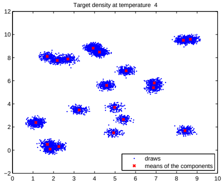


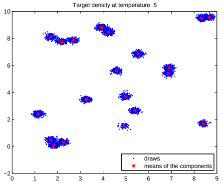
- target density : $\pi = \sum_{i=1}^{20} \mathcal{N}_2(\mu_i, \Sigma_i)$
- K processes with target distribution π^{1/T_k} $(T_K = 1)$

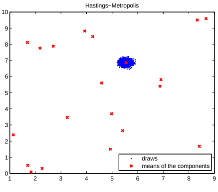












"Design parameters" of the EE sampler

1. How to choose the **probability of interaction** ϵ ?
2. How many **temperatures**, and which ones ?
3. How many **energy levels**, and which ones ?

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) these problems are still open.

"Design parameters" of the EE sampler

1. How to choose the **probability of interaction** ϵ ?
2. How many **temperatures**, and which ones ?
3. How many **energy levels**, and which ones ?

- In the original EE: energy rings = strata in the range of the energy $\mathcal{H}$ of the target π

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

Choose H_i s.t. $\min \mathcal{H} < H_1 < \cdots < H_L.$

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

- *Our contribution*: tune adaptively the boundaries of the strata

Num. Appl.: fixed boundaries vs adapted boundaries

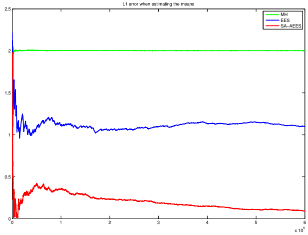
- Target distribution on $\mathbb{R}^6$

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

- 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 $u^T \propto [1, 1, \cdots, 1] \qquad v^T \propto [1, 1, 1, -1, -1, -1]$

Behavior along one path: HM EE A-EE

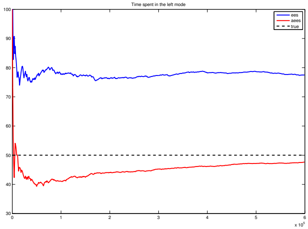


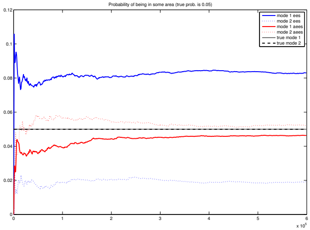
[Top] Error when estimating the means

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

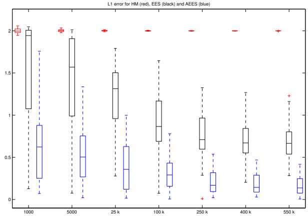
[Bottom L] Time spent in one of the mode where the path is initialized.

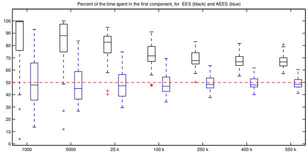
[Bottom R] Probability of being in some ellipsoids, for the first mode (line) and the second one (dashed line)





Behavior on 50 ind. run HM EE A-EE



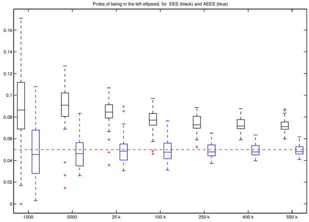


[Top] Error when estimating the means

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

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

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



Adaptive tuning of the boundaries of the energy rings

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

Algorithm

- Level 1 (Hot level)
  - Sample $X^{(1)}$ 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
  - Sample $X^{(2)}$ (MCMC step and interaction step) 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.

On the convergence of such adaptive schemes

Convergence result: we prove ergodicity and a strong law of large numbers for A-EE.

Our approach for the proof is by induction:

- we assume the process $X^{(k-1)}$ "converges".
- we prove that the process $X^{(k)}$ has the same convergence properties.
- Repeat from level 1 to K.

Tools for the proof:

- the conditional distribution $\mathcal{L}(X_n^{(k)} | \mathrm{past}_{n-1}^{(1:k)})$ is $P_{\theta_{n-1}}^{(k)}(X_{n-1}^{(k)}, \cdot)$

$$P_{\theta_n}^{(k)}(x, dy) = \epsilon P^{(k)}(x, dy) + (1-\epsilon) K_{\theta_n}^{(k)}(x, dy)$$

$$K_{\theta_n}^{(k)}(x, A) = \int_A \alpha_{\theta_n}^{(k)}(x, y) \, \frac{g_{\theta_n}(x, y)\theta_n(dy)}{\int g_{\theta_n}(x, z)\theta_n(dz)} + \delta_x(A) \int \{1 - \alpha_{\theta_n}^{(k)}(x, y)\} \, \frac{g_{\theta_n}(x, y)\theta_n(dy)}{\int g_{\theta_n}(x, z)\theta_n(dz)}$$

$$\theta_n = \frac{1}{n} \sum_{j=1}^{n} \delta_{X_j^{(k-1)}} \qquad \alpha_{\theta_n}^{(k)}(x, y) = 1 \wedge \frac{\pi^{1/T_k - 1/T_{k-1}}(y)}{\pi^{1/T_k - 1/T_{k-1}}(x)} \, \frac{\int g_{\theta_n}(x, z)\theta_n(dz)}{\int g_{\theta_n}(y, z)\theta_n(dz)}$$

$$g_{\theta_n}(x, y) = \text{"}x \text{ and } y \text{ are in the same energy ring with boundaries defined by } H_{n,\bullet}^{(k-1)}\text{"}$$

$$\stackrel{(ex.)}{=} \begin{cases} 0 & if \text{ if } x, y \text{ are in the same energy level} \\ 1 & if \text{ otherwise} \end{cases}$$

On the convergence of such adaptive schemes

Convergence result: we prove ergodicity and a strong law of large numbers for A-EE.

Our approach for the proof is by induction:

- we assume the process $X^{(k-1)}$ "converges".
- we prove that the process $X^{(k)}$ has the same convergence properties.
- Repeat from level 1 to K.

Tools for the proof:

- the conditional distribution $\mathcal{L}(X_n^{(k)}|\mathrm{past}_{n-1}^{(1:k)})$ is $P_{\theta_{n-1}}^{(k)}(X_{n-1}^{(k)}, \cdot)$
- containment and diminishing adaptation conditions extensions from the pioneering work by (Roberts, Rosenthal (2005)) + Poisson equation + Limit Theorems for Martingales.
- **condition on the adapted boundaries**
  - (i) There exists $\beta > 0$ s.t. $\lim_n n^\beta \left| H_{n,\bullet}^{(k)} - H_{n-1,\bullet}^{(k)} \right| = 0$ w.p.1.
  - (ii) $H_{n,\bullet}^{(k)} \to H_{\infty,\bullet}^{(k)}$ w.p.1 when $n \to \infty$.
  - (iii) assumption on the limiting boundaries:

$$\inf_x \int g_\infty^{(k)}(x,y) \pi^{1/T_k}(dy) > 0$$

Example of adaptive boundaries

Example of adaptive boundaries:
choose $\exp(-H_i^{(k)})$ for $1 \leq i \leq L$ (computed from $X^{(k)}$) as the **quantiles of order** $i/(L+1)$ **of** the distribution of

$$\pi(Z) \qquad \text{when } Z \sim \pi^{1/T_k}$$

Example of adaptive boundaries

Example of adaptive boundaries:
choose $\exp(-H_{n,i}^{(k)})$ for $1 \leq i \leq L$ (computed from $X_{1:n}^{(k)}$) as an estimator of the **quantiles of order** $i/(L+1)$ **of** the distribution of

$$\pi(Z) \qquad \text{when } Z \sim \pi^{1/T_k}$$

Example of adaptive boundaries

Example of adaptive boundaries:
choose $\exp(-H_{n,i}^{(k)})$ for $1 \leq i \leq L$ (computed from $X_{1:n}^{(k)}$) as an estimator of the **quantiles of order $i/(L+1)$ of** the distribution of

$$\pi(Z) \qquad \text{when } Z \sim \pi^{1/T_k}$$

Note that in EE, when using the interacting step to sample $X_n^{(k)}$

- determine the ring such that $\quad H_{i-1} \leq -\log \pi(X_{n-1}^{(k)}) \leq H_i$
- choose (at random) one point among $X_1^{(k-1)}, \cdots, X_{n-1}^{(k-1)}$ such that

$$\exp(-H_i) \leq \pi(X_\bullet^{(k-1)}) \leq \exp(-H_{i-1})$$

  and accept / reject.
- When convergence: $\quad \mathcal{L}(X_n^{(k-1)}) \to \pi^{1/T_{k-1}}$ when $n \to \infty$

Quantile estimators

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

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

   (+) easy implementation
   (-) time consuming

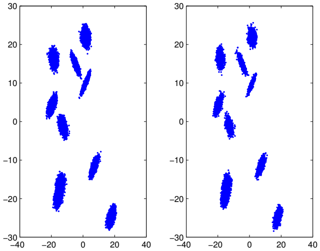
2) A second one is based on Stochastic Approximation procedures

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

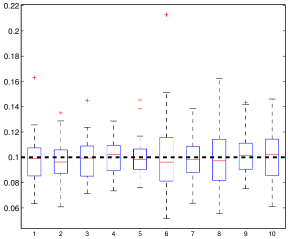
   (+) running time

   (-) implementation of SA algorithm (choice of the step-size, initialization)

Num. Appl.: Adaptive EE



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



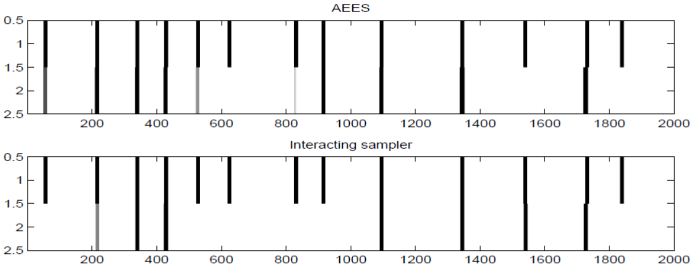
Frequency of the visit to each component of the mixture. Boxplot with 50 ind. run

Num. Appl.: Motif discovery in DNA sequence

Same model as in the talk of Dawn, yesterday:

- a background sequence, with a **Markovian** transition (known)
- motifs, of known length, with independent multinomial transition (unknown)

Here is the result for A-EE and EE



Conclusion

- EE depends on many design parameters that all play a role on the efficiency of the sampler. We propose an **adaptive** procedure **to tune** on the fly the **energy rings**.
- Convergence results are established * when the quantiles are estimated by inversion of the cdf.
- Work in progress: convergence when the quantiles are estimated by a Stochastic Approximation procedure.
  Challenging: convergence of SA algorithms when the draws are **not** Markovian (thanks to M. Vihola).
- First convergence results on EE **with selection** of the auxiliary point during the interaction step.

*Submitted, available at `http://perso.telecom-paristech.fr/ schreck`