When Markov Chains control Monte Carlo sampling

Gersende Fort

CNRS

Institut de Mathématiques de Toulouse, France

Processus Markoviens, semi-markoviens et leurs applications - Montpellier, 2023.

Joint works with

- Yves Atchade, Univ. Michigan, USA
- Benjamin Jourdain, ENPC, France
- Estelle Kuhn, INRAE, France
- Tony Lelievre, ENPC, France
- Eric Moulines, Ecole Polytechnique, France
- Pierre Priouret, Univ. Paris 6, France
- Amandine Schreck, Telecom ParisTech, France
- Gabriel Stoltz, ENPC, France
- Pierre Vandekerkhove, Univ. Marne-la-Vallée, France

Convergence of adaptive and interacting Markov chain Monte Carlo algorithms. Fort,Moulines,Priouret (2011, Ann Stat).

A Central limit theorem for adaptive and interacting Markov Chains. Fort, Moulines, Priouret, Vandekerkhove (2014, Bernoulli).

Self-Healing Umbrella Sampling: Convergence and Efficiency. Fort, Jourdain, Lelièvre and Stoltz (2017, Stat & Comput)

Adaptive Equi-Enery sampler: Convergence and Illustration. Schreck, Fort, Moulines (2013, TOMACS).

A control, why ?

To improve Monte Carlo methods targetting: $\mathrm{d}\pi = \pi \, \mathrm{d}\mu$

•The "naive" MC sampler depends on design parameters in $\mathbb{R}^p$ or in infinite dimension θ

•Theoretical studies caracterize an optimal choice of theses parameters $\theta_\star$ by

$$\theta_\star \in \Theta \text{ s.t. } \int H(\theta, x) \; \mathrm{d}\pi(x) = 0$$

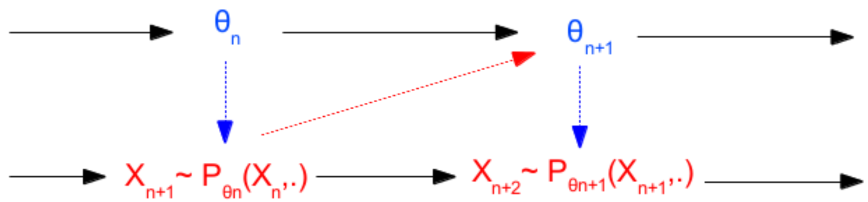
or

$$\theta_\star \in \mathrm{argmin}_{\theta \in \Theta} \int C(\theta, x) \; \mathrm{d}\pi(x) = 0.$$

• Strategies:

- Strategy 1: a preliminary "machinery" for the approximation of $\theta_\star$; **then** run the MC sampler with $\theta \leftarrow \theta_\star$
- Strategy 2: learn θ and sample **concomitantly**

In this talk, Monte Carlo sampling !



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

- from the Monte Carlo point of view:
  which conditions on the updating scheme for convergence of the sampler ?
  Case: Markov chain Monte Carlo sampler

- from the optimization point of view:
  which conditions on the Monte Carlo approximation for convergence of the stochastic optimization ?
  Case: Stochastic Approximation methods with Markovian inputs

Outline

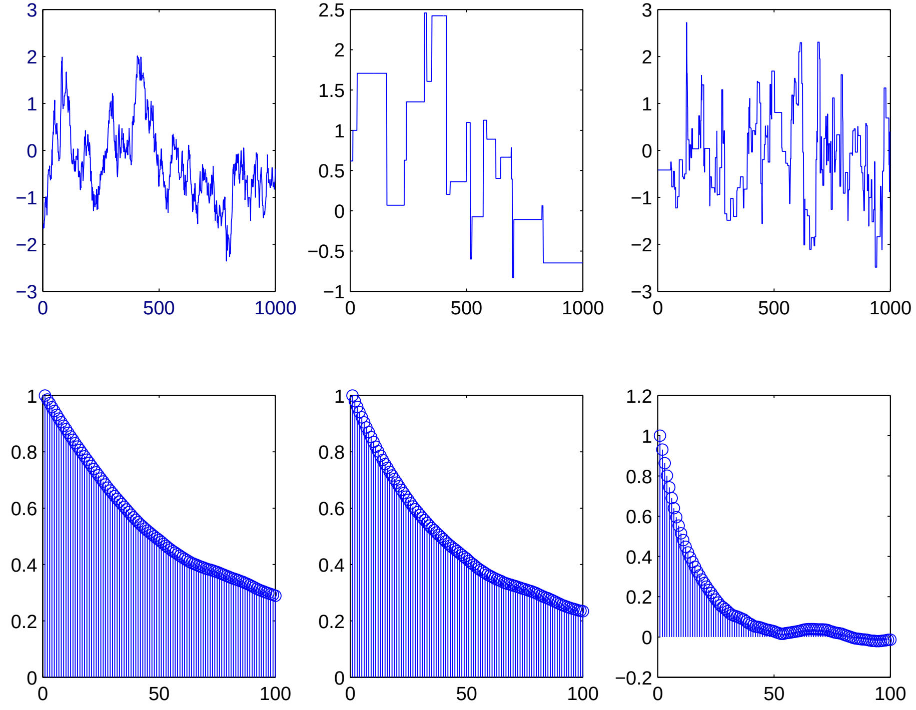
- Part I. Motivating examples: adaptive and interacting Markov chain Monte Carlo samplers.

- Part II. Ergodicity and limit theorems.

Part I: Motivating examples

1st Ex. Adaptive Hastings-Metropolis (1/3)

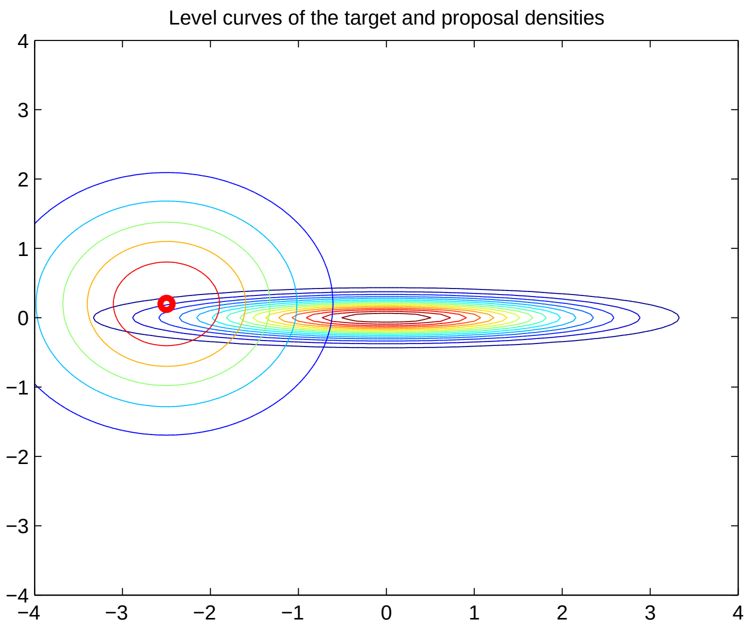
Symmetric Random Walk : proposal $\quad X_{t+1/2} \sim X_t + \mathcal{N}(0, \Gamma)$



(d=1) Different values of Γ : [top] a path of the Markov chain. [bottom] auto-correlation function

1st Ex. Adaptive Hastings-Metropolis (2/3)

Pioneering works: Gelman, Roberts and Gilks (1996)



$(d = 2)$ The level curves of the target density, and of the proposal distribution $\mathcal{N}(X_t, Id)$.

The optimal rule

$$\Gamma = \frac{2.38^2}{d}\Sigma_\pi$$

But Σ_π is unknown

1st Ex. Adaptive Hastings-Metropolis (3/3)

Pioneering works: Haario, Saksman and Tamminen (1999).

Iterate

- Sample X_{t+1} by a HM kernel with proposal $\mathcal{N}(X_t, \widehat{\Gamma}_t)$: $X_{t+1} \sim P_{\widehat{\Gamma}_t}(X_t, \cdot)$.

- Update the unknonw prameter

$$\begin{aligned}
\widehat{\Gamma}_{t+1} &= \frac{1}{t+1} \sum_{\ell=1}^{t+1} (X_\ell - \widehat{\mu}_\ell)(X_\ell - \widehat{\mu}_\ell)^\top \\
&= \widehat{\Gamma}_t + \frac{1}{t+1} \left\{ (X_{t+1} - \widehat{\mu}_{t+1})(X_{t+1} - \widehat{\mu}_{t+1})^\top - \widehat{\Gamma}_t \right\}
\end{aligned}$$

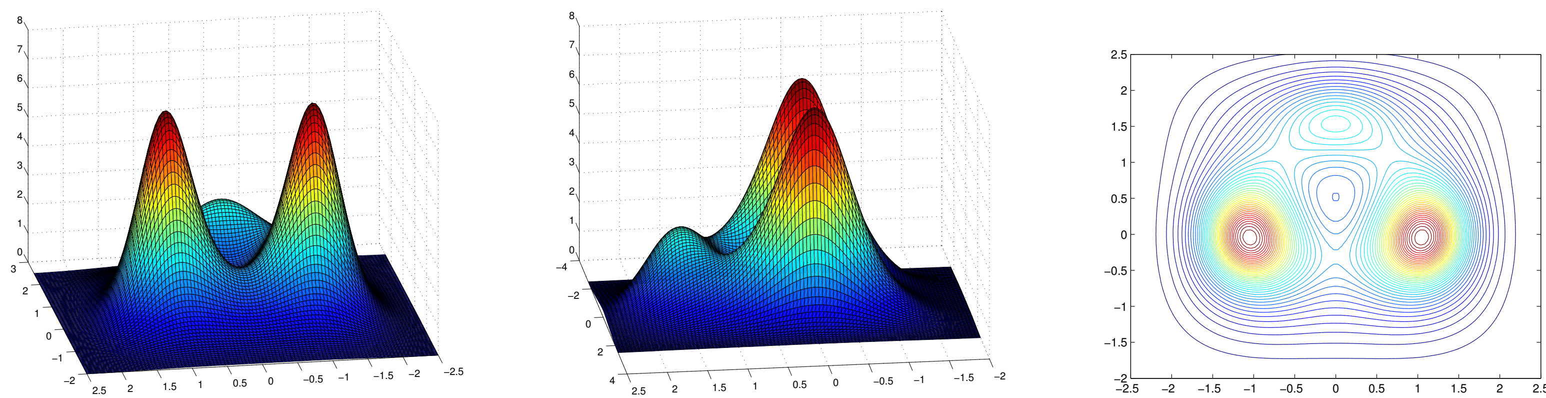
by one step of a *Stochastic Approximation* algorithm designed to solve

$$\mathbb{E}_\pi \left[(X - \mathbb{E}_\pi[X])(X - \mathbb{E}_\pi[X])^\top \right] - \Gamma = 0.$$

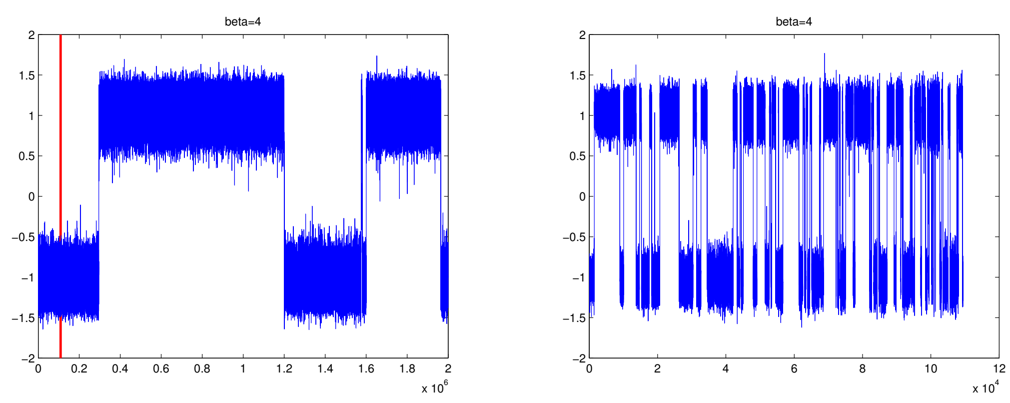
2nd Ex. Adaptive Importance Sampling (1/6)

The problem

- A highly multimodal target density $\mathrm{d}\pi$ on $\mathcal{X} \subseteq \mathbb{R}^d$.



- Two samplers with different behaviors (plot: the x-path of a chain in $\mathbb{R}^2$)



2nd Ex. Adaptive IS by Wang Landau approaches (2/6)

The strategy for choosing the proposal mechanism

- A family of proposal mechanisms obtained by biasing locally the target:

\- given a partition $\mathcal{X}_1, \cdots, \mathcal{X}_I$ of $\mathcal{X}$,

\- for any weight vector $\theta = (\theta(1), \cdots, \theta(I))$

$$\mathrm{d}\pi_\theta(x) = \frac{1}{\sum_{i=1}^I \frac{\theta_\star(i)}{\theta(i)}} \sum_{i=1}^I 1_{\mathcal{X}_i}(x) \, \frac{\mathrm{d}\pi(x)}{\theta(i)}, \qquad \text{with } \theta_\star(i) := \int_{\mathcal{X}_i} \mathrm{d}\pi(u).$$

- Optimal proposal: $\mathrm{d}\pi_{\theta_\star}$ <proof>

- Unfortunately, $\theta_\star$ unavailable.

2nd Ex. Adaptive IS by Wang Landau approaches (3/6)

If $\pi_{\theta_\star}$ were available

- The algorithm would be:

- Sample $X_1, \cdots, X_n, \cdots$ i.i.d. with distribution $\mathrm{d}\pi_{\theta_\star}$ (or a MCMC with target $\mathrm{d}\pi_{\theta_\star}$)
- Compute the importance ratio

$$\frac{\mathrm{d}\pi}{\mathrm{d}\pi_{\theta_\star}}(X_k) = I \sum_{i=1}^{I} 1_{\mathcal{X}_i}(X_k)\ \theta_\star(i)$$

- When approximating an expectation, set

$$\int \phi\, \mathrm{d}\pi \approx \frac{I}{T} \sum_{t=1}^{T} \left(\sum_{i=1}^{I} 1_{\mathcal{X}_i}(X_t)\ \theta_\star(i) \right)\ \phi(X_t).$$

2nd Ex. Adaptive IS by Wang Landau approaches (4/6)

$\theta_\star$ and therefore $\mathrm{d}\pi_{\theta_\star}$ are unknown, so ?

- $\theta_\star \in \mathbb{R}^I$ collects $\int_{\mathcal{X}_i} \mathrm{d}\pi$ for all $i \in \{1, \cdots, I\}$,

- $\theta_\star$ the unique root of $\theta \mapsto \int_{\mathcal{X}} H(\theta, x)\ \mathrm{d}\pi_\theta(x) \in \mathbb{R}^I$ where for all $i \in \{1, \cdots, I\}$

$$H_i(\theta, x) := \theta(i) 1_{\mathcal{X}(i)}(x) - \theta(i) \sum_{j=1}^{I} 1_{\mathcal{X}_j}(x)\theta(j).$$

thus suggesting the use of a Stochastic Approximation procedure: $\theta_\star \approx \lim_t \theta_t$

$$\theta_{t+1} = \theta_t + \gamma_{t+1} H(\theta_t, X_{t+1}) \qquad X_{t+1} \sim \mathrm{d}\,\pi_{\theta_t}$$

- This update scheme is a normalized counter of the number of visits to $\mathcal{X}_i$

2nd Ex. Adaptive IS by Wang Landau approaches (5/6)

The algorithm: Wang-Landau based procedures

- Initialisation: a weight vector θ_0

Repeat for $t = 1, \cdots, T$
- sample a point $X_{t+1} \sim \mathrm{d}\pi_{\theta_t}$
- update the estimate of $\theta_\star$

$$\theta_{t+1} = \theta_t + \gamma_{t+1} H(\theta_t, X_{t+1}) \qquad .$$

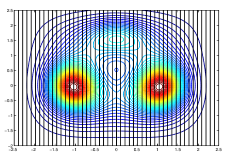
where $X_{t+1} \sim P_{\theta_t}(X_t, \cdot)$ and P_θ inv. wrt $\mathrm{d}\pi_\theta$.

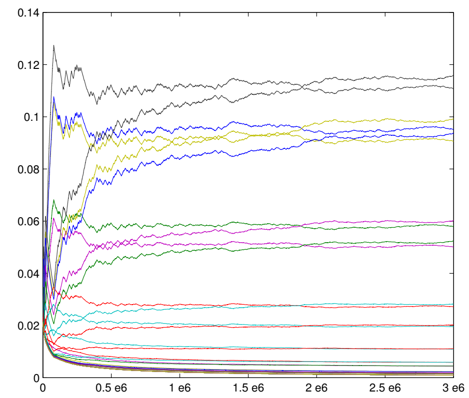
- Expected:

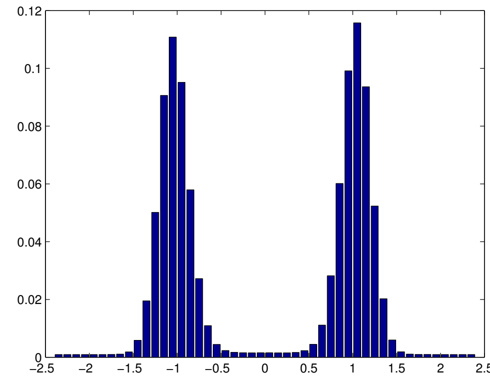
- the convergence of θ_t to $\theta_\star$: SA scheme, fed with adaptive (controlled) MCMC sampler,
- the convergence of the distribution of X_t to $\mathrm{d}\,\pi_{\theta_\star}$

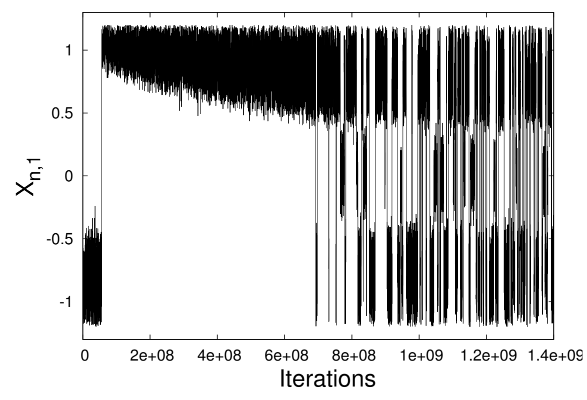
2nd Ex. Adaptive IS by Wang Landau approaches (6/6)

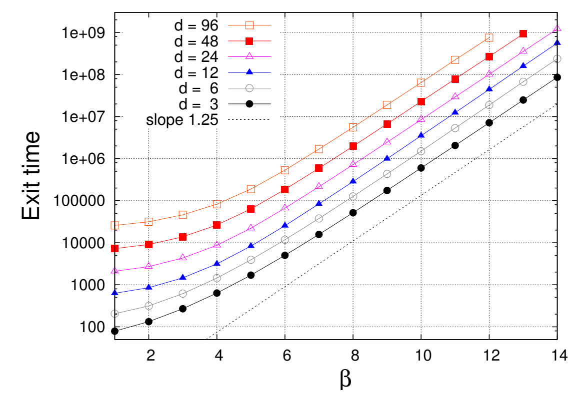
Does it work ? Plot: convergence of θ_t and first exit times from one mode











- see F, Kuhn, Jourdain, Lelièvre, Stoltz (2014); F, Jourdain, Lelièvre, Stoltz (2015,2017,2018) for studies of these Wang-Landau bases algorithms; including self-tuned SA update rules (γ_t is random).

Conclusion of the 2nd example

- Iterative sampler

- Each iteration combines : (i) a sampling step $X_{t+1} \sim P_{\theta_t}(X_t, \cdot)$; and (ii) a step from an optimization algo. to update the knownledge of some optimal parameter.

- The points $\{X_1, \cdots, X_t, \cdots\}$ can be seen as the output of a controlled Markov chain

$$\mathbb{E}\left[f(X_{t+1})|\mathcal{F}_t\right] = P_{\theta_t}(X_t, \cdot) \qquad \mathcal{F}_t := \sigma(X_{0:t}, \theta_0)$$

where P_θ has $\mathrm{d}\pi_\theta$ as its unique invariant distribution.

- The convergence of the parameter θ_t is the convergence of a SA scheme with "controlled Markovian" dynamics

$$\theta_{t+1} = \theta_t + \gamma_{t+1} \, H(\theta_t, X_{t+1})$$

3rd Ex. the Adaptive Equi-Energy sampler (1/4) extend the EE sampler by Kou-Zhou-Wong, 2006

- We discussed the case when $\theta \in \mathbb{R}^p$. But there are more general situations: θ may be a distribution case of "interacting" MCMC. (Del Moral-Doucet, 2010; F.-Moulines-Priouret, 2012; Schreck-F.-Moulines, 2013; F.-Moulines-Priouret-Vandekerkhove, 2016

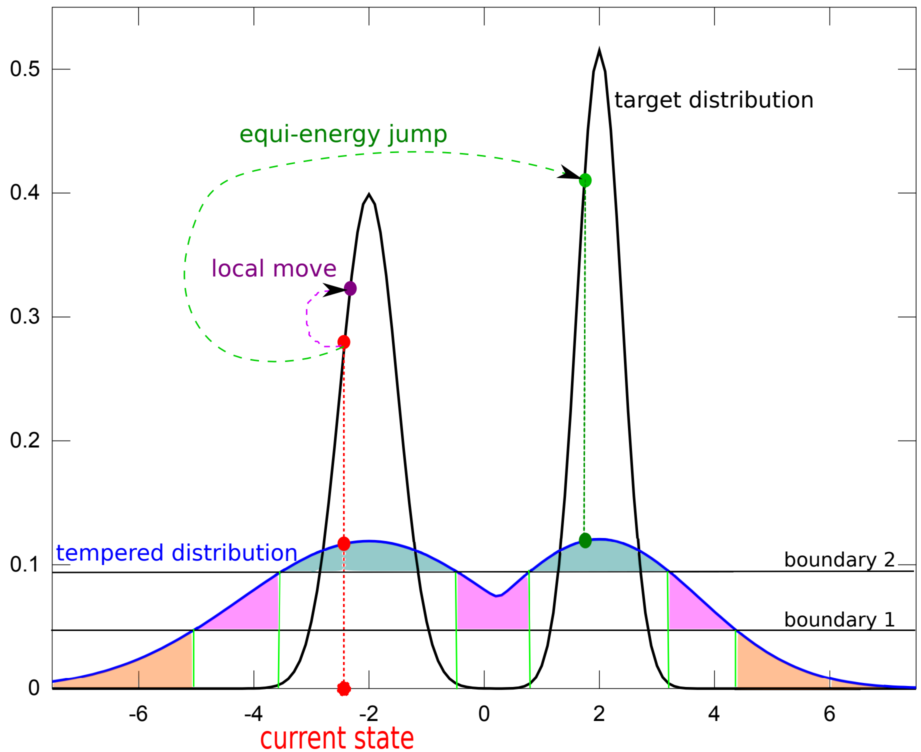
- Both **interacting** and **tempering** and **adaptive** algorithm.

    - Interacting: run K chains in parallel, s.t. chain $\#k$ is built by using the points of chain $\#(k-1)$. Except the chain $\#1$.

    - Tempering: given $\beta_1 < \cdots < \beta_K = 1$, chain $\#k$ is designed to target $\mathrm{d}\pi^{\beta_k}$.

    - Adaptive: the mecanism of interaction is learnt on the fly.

3rd Ex. the Adaptive Equi-Energy sampler (2/4)

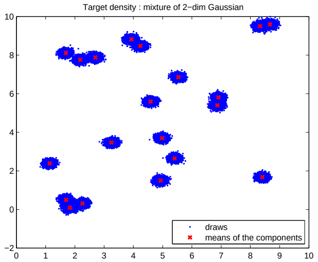


The equi-energy jump: (i) adaptive definition of the equi-energy rings as an estimation of the quantiles of $-\log \pi^{\beta_k}(Z)$ with $Z \sim \pi^{\beta_k}$; (ii) acceptance-rejection ratio;

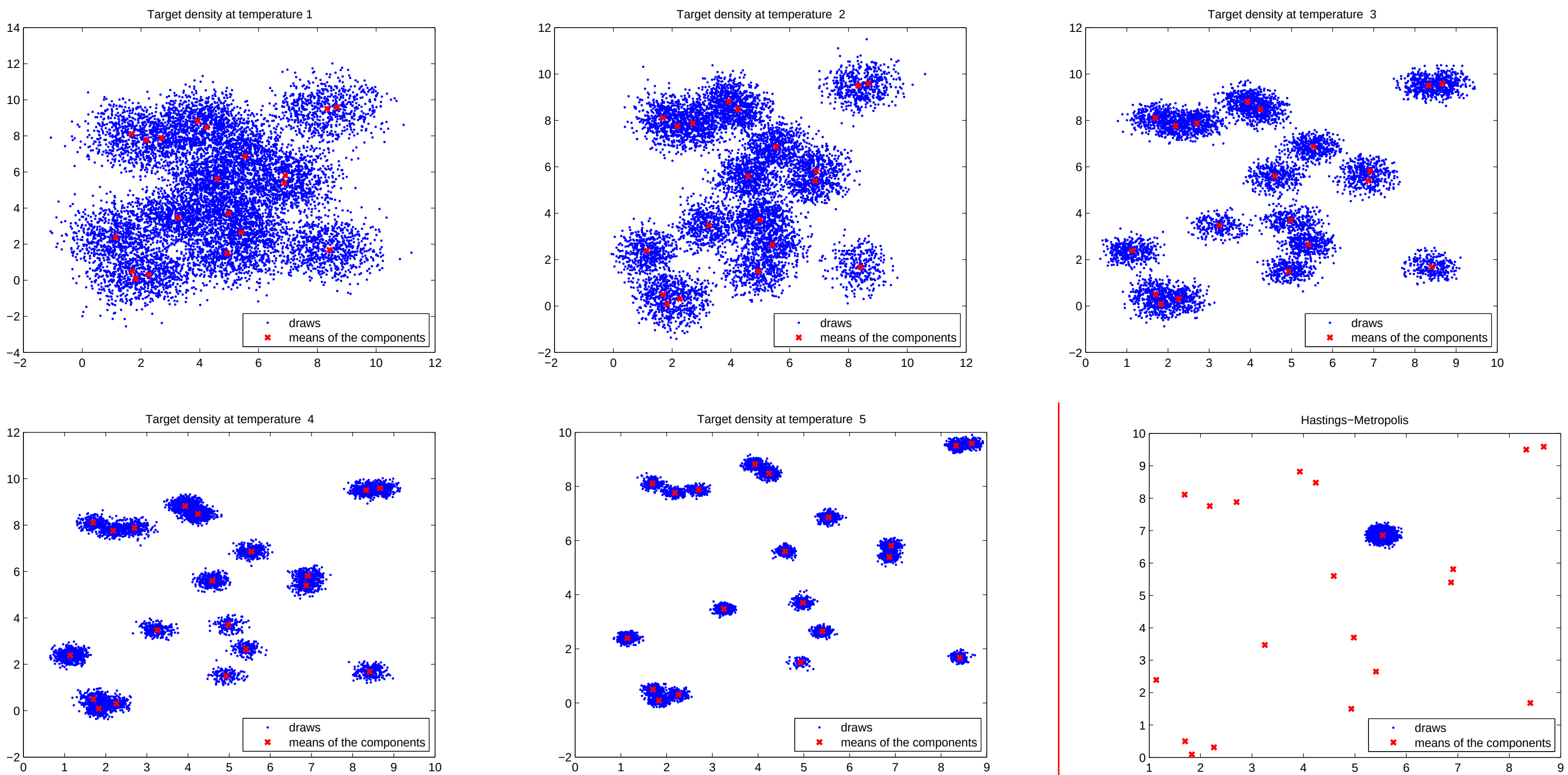
From chain $X^{(k)}$ to $X^{(k+1)}$:

$$P_{\theta_{k,t}}(X_t^{(k+1)}, \cdot) = (1-\epsilon) \underbrace{Q(X_t^{(k+1)}, \cdot)}_{\text{MCMC with target } \pi^{\beta_{k+1}}} + \epsilon \underbrace{\tilde{Q}_{\theta_{k,t}}(X_t^{(k+1)}, \cdot)}_{\substack{\text{kernel depending on} \\ \text{the empirical distribution } \theta_{k,t} \\ \text{of the auxiliary process } X_{1:t}^{(k)}}}$$

3rd Ex. the Adaptive Equi-Energy sampler (3/4)



- target density : $\pi = \sum_{i=1}^{20} \mathcal{N}_2(\mu_i, \Sigma_i)$
- 5 parallel processes with target distribution π^{β_k} $(\beta_5 = 1)$



3rd Ex. the Adaptive Equi-Energy sampler (4/4)

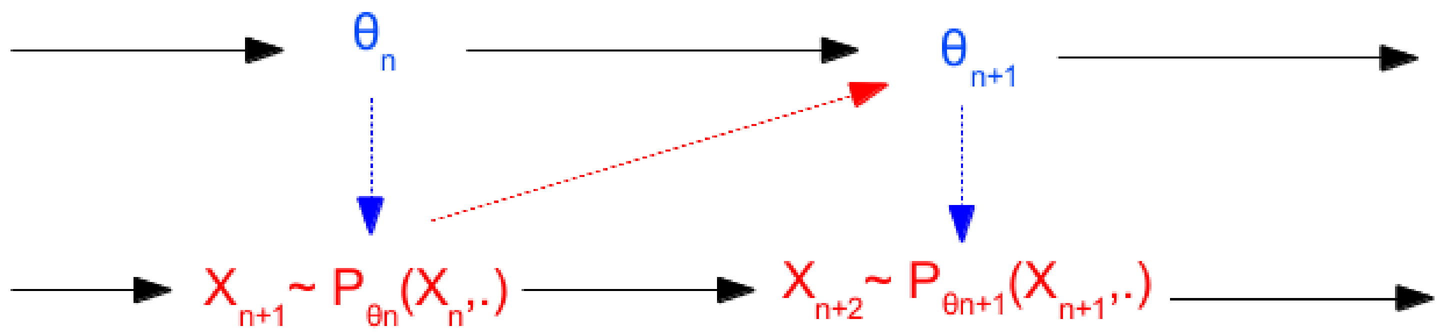
- In this example, θ is homogeneous to an empirical distribution (random probability measure).

- For this adaptive sampler Schreck-F.-Moulines (2013): convergence in distribution, law of large numbers.

- For the non-adaptive sampler: convergence anaysis in Kou-Zhou-Wong, 2006; Atchadé, 2010; Andrieu-Jasra-Doucet-Del Moral, 2011; F.-Moulines-Priouret, 2012; F.-Moulines-Priouret-Vandekerkhove, 2014

- General results when θ is not necessarily in $\mathbb{R}^p$: convergence in distribution, law of large numbers, CLT in F.-Moulines-Priouret, 2012; F.-Moulines-Priouret-Vandekerkhove, 2014

Conclusion of this first part (1/3): is a theory required ?



Conclusion of this first part (2/3): Yes !

YES ! convergence can be lost by the adaption mechanism

Even in a simple case when

$\forall\theta \in \Theta,$ $\quad$ P_θ invariant wrt $\mathrm{d}\pi$,

one can define a simple adaption mechanism

$$X_{t+1}|\mathrm{past}_{1:t} \sim P_{\theta_t}(X_t, \cdot) \qquad \theta_t \in \sigma(X_{1:t})$$

such that

$$\lim_t \mathbb{E}\left[f(X_t)\right] \neq \int f \; \mathrm{d}\pi.$$

Conclusion of the first part (3/3): look !

Fix $t_0, t_1 \in (0,1)$ s.t. $t_0 + t_1 = 1$.

A $\{0,1\}$-valued chain $\{X_t\}_t$ defined by $X_{t+1} \sim P_{X_t}(X_t, \cdot)$ where the transition matrices are

$$P_0 = \begin{bmatrix} t_0 & (1-t_0) \\ (1-t_0) & t_0 \end{bmatrix} \qquad P_1 = \begin{bmatrix} t_1 & (1-t_1) \\ (1-t_1) & t_1 \end{bmatrix}$$

Then

- P_0 and P_1 are invariant w.r.t $[1/2, 1/2]$
- **But** $\{X_t\}$ is a Markov chain invariant w.r.t. $[t_1, t_0]$.

$\{X_t\}$ does not have the same invariant distribution as the one of P_0 anf P_1.

Part II: Convergence of Adaptive/Controlled Markov chains

Convergence results

- The framework:

- a filtration $\{\mathcal{F}_t, t \geq 0\}$ on $(\Omega, \mathcal{A}, \mathbb{P})$
- a $\mathcal{F}_t$-adapted $\mathcal{X} \times \Theta$-valued process $\{(X_t, \theta_t), t \geq 0\}$ defined on $(\Omega, \mathcal{A})$
- a family of transition kernels $\{P_\theta, \theta \in \Theta\}$ on a general state space $(\mathcal{X}, \mathcal{X})$
- a conditional distribution satisfying

$$\mathbb{E}\left[f(X_{t+1})|\mathcal{F}_t\right] = \int P_{\theta_t}(X_t, \mathrm{d}x) f(x) \qquad f \text{ bounded continuous}$$

and a convergence (in some sense) of the kernels $\{P_{\theta_t}, t \geq 0\}$

BEWARE: the chain $\{X_t\}_t$ is NOT a Markov chain

- Questions:

- convergence in distribution of X_t ?
- limit theorems (SLLN, CLT)

- Hereafter:

- focus on the convergence in distribution; then few words on CLT.
- focus first on $\theta \in \Theta \subset \mathbb{R}^p$; then few words on a more general situation.

Assumptions (1/3) Invariant distribution

$\forall \theta \in \Theta, \quad \exists \pi_\theta$ s.t. the kernel P_θ invariant wrt π_θ

Assumptions (2/3) (Generalized) Containment condition

- Uniform-in-θ ergodicity condition

$$\sup_{\theta\in\Theta} \|P_\theta^r(x;\cdot) - \pi_\theta\|_{\mathsf{TV}} \le C\rho^r \qquad\qquad \rho \in (0,1).$$

In practice: a drift and a minorization condition $\to$ explicit control of ergodicity

$$P_\theta V \le \lambda_\theta V + b_\theta, \qquad\qquad P_\theta(x,\cdot) \ge \delta_\theta \nu_\theta(\cdot) \text{ for } x \in \{V \le 2b_\theta(1-\lambda_\theta)^{-1} - 1\}$$

- [Weakened Cond.] for any $\epsilon > 0$, there exists a non-decreasing sequence r_ϵ s.t. $\lim_t r_\epsilon(t)/t = 0$ and

$$\limsup_t \mathbb{E}\left[\|P^{r_\epsilon(t)}_{\theta_{t-r_\epsilon(t)}}(X_{t-r_\epsilon(t)};\cdot) - \pi_{\theta_{t-r_\epsilon(t)}}\|_{\mathsf{TV}}\right] \le \epsilon$$

- Controlled rate of growth-in-θ *here,* $r_\epsilon(t) = t^\bullet$

$\|P_\theta^r(x;\cdot) - \pi_\theta\|_{\mathsf{TV}} \le C_\theta\, \rho_\theta^r$ $\qquad t^{-\tau}\ \|\theta_t\| < \infty$ a.s. $\qquad \limsup_t t^{-\tilde{\tau}}\ \left(C_{\theta_t} \vee (1-\rho_{\theta_t})^{-1}\right) < \infty$ a.s.

Assumptions (3/3) (Generalized) Diminishing adaptation condition

- Uniform-in-θ ergodic condition,

$$\lim_t \mathbb{E}\left[D(\theta_t, \theta_{t-1})\right] = 0$$

where $D(\theta, \theta') = \sup_x \|P_\theta(x, \cdot) - P_{\theta'}(x, \cdot)\|_{\mathrm{TV}}$.

- [Weakened cond.] For any $\epsilon > 0$,

$$\lim_t \mathbb{E}\left[\sum_{j=1}^{r_\epsilon(t)-1} D(\theta_{t-r_\epsilon(t)+j}, \theta_{t-r_\epsilon(t)})\right] = 0$$

In practice
- Prove a Lipschitz property $D(\theta, \theta') \leq C\,\|\theta - \theta'\|$
- Use the definition of θ_t as a function of $(X_\ell)_{\ell \leq t}$ and possibly other "external" sampled points
- Require controls of the form $\mathbb{E}\left[W(X_\ell)\right]$, solved e.g. by drift inequalities

$$\mathbb{E}\left[W(X_\ell) | \mathcal{F}_{\ell-1}\right] = P_{\theta_{\ell-1}} W(X_{\ell-1}) \leq \lambda_{\theta_{\ell-1}} W(X_{\ell-1}) + b_{\theta_{\ell-1}}$$

Convergence in distribution (1/3)

When $\pi_\theta = \pi$ for any θ

Under these conditions, for any bounded function f,

$$\lim_t \mathbb{E}\left[f(X_t)\right] = \int f(x)\ \mathrm{d}\pi(x)$$

Sketch of proof :

$$\mathbb{E}\left[f(X_t)\right] - \int f(x)\ \mathrm{d}\pi(x) = \mathbb{E}\left[f(X_t)\right] - \mathbb{E}\left[P^r_{\theta_{t-r}} f(X_{t-r})\right] + \mathbb{E}\left[P^r_{\theta_{t-r}} f(X_{t-r})\right] - \int f(x)\ \mathrm{d}\pi(x)$$

$$|\qquad \cdot \qquad| \leq \left|\mathbb{E}\left[f(X_t)\right] - \mathbb{E}\left[P^r_{\theta_{t-r}} f(X_{t-r})\right]\right| + \|f\|_\infty \, \mathbb{E}\left[\sup_\theta \|P^r_\theta(X_{t-r}, \cdot) - \pi\|_{\mathsf{TV}}\right]$$

$$\left|\mathbb{E}\left[f(X_t)\right] - \mathbb{E}\left[P^r_{\theta_{t-r}} f(X_{t-r})\right]\right| \leq \sum_{j=1}^{r-1} \mathbb{E}\left[D\left(\theta_{t-r+j}, \theta_{t-r}\right)\right]$$

Convergence in distribution (2/3)

When each kernel P_θ has its own invariant distribution π_θ, with an explicit expression

- Under these three conditions, and

- there exists a constant α s.t. $\lim_t \int f \, \mathrm{d}\pi_{\theta_t} = \alpha$ a.s.

then

$$\lim_t \mathbb{E}\left[f(X_t)\right] = \alpha.$$

- Corollary: if $\{\pi_{\theta_t}\}_t$ converges weakly to π a.s., then $\alpha = \int f \, \mathrm{d}\pi$ for any bounded continuous function f.

Convergence in distribution (3/3)

When π_θ exists but its expression is unknown

It is the most technical case: how to prove the convergence of $\int f \, \mathrm{d}\pi_{\theta_t}$ when only properties on the kernels P_{θ_t} are available ?

We write

$$\int f \mathrm{d}\pi_{\theta_t} - \int f \, \mathrm{d}\pi_{\theta_\star} = \left(\int f \, \mathrm{d}\pi_{\theta_t} - \int f(y) \, P^k_{\theta_t}(x, \mathrm{d}y) \right) \\ + \left(\int P^k_{\theta_t}(x, \mathrm{d}y) f(y) - \int P^k_{\theta_\star}(x, \mathrm{d}y) f(y) \right) + \left(\int P^k_{\theta_\star}(x, \mathrm{d}y) f(y) - \int f \, \mathrm{d}\pi_{\theta_\star} \right)$$

and control the blue terms by a condition on the ergodicity of the transition kernels. For the red one,

$$P^k_{\theta_t} f(x) - P^k_{\theta_\star} f(x) = \int \left(P_{\theta_t}(x, \mathrm{d}y) - P_{\theta_\star}(x, \mathrm{d}y) \right) P^{k-1}_{\theta_\star} f(y) \\ + \int P_{\theta_t}(x, \mathrm{d}y) \left(P^{k-1}_{\theta_t} f(y) - P^{k-1}_{\theta_\star} f(y) \right)$$

Convergence in distribution (3/3) (to follow)

Starting from :

$$\forall x \in \mathcal{X}, A \in \mathcal{X}, \quad \exists \Omega_{x,A}, \quad \mathbb{P}(\Omega_{x,A}) = 1 \quad \forall \omega \in \Omega_{x,A} \quad \lim_t P_{\theta_t(\omega)}(x, A) = P_{\theta_\star}(x, A),$$

the steps are:

$$\forall x \in \mathcal{X}, \quad \exists \Omega_x, \qquad \mathbb{P}(\Omega_x) = 1 \qquad \forall \omega \in \Omega_x \qquad \lim_t P_{\theta_t(\omega)}(x, \cdot) \xrightarrow{w} P_{\theta_\star}(x, \cdot)$$

$\hookrightarrow$ Tool: separable metric space $\mathcal{X}$ (ex. Polish)

$$\exists \Omega', \qquad \mathbb{P}(\Omega') = 1 \qquad \forall \omega \in \Omega', x \in \mathcal{X} \qquad \lim_t P_{\theta_t(\omega)}(x, \cdot) \xrightarrow{w} P_{\theta_\star}(x, \cdot),$$

$\hookrightarrow$ Tool: Polish space $\mathcal{X}$ + equicontinuity of $\{P_\theta f - P_{\theta_\star} f, \theta \in \Theta\}$

$$\exists \Omega_\star, \qquad \mathbb{P}(\Omega_\star) = 1 \qquad \forall \omega \in \Omega_\star \qquad \lim_t P^k_{\theta_t(\omega)}(x, \cdot) \xrightarrow{w} P^k_{\theta_\star}(x, \cdot),$$

$\hookrightarrow$ Tool: Feller properties of the kernels $\{P_\theta, \theta \in \Theta\}$.

See F.-Moulines-Priouret, 2012)

In the literature

(Roberts-Rosenthal,2007; Atchadé-F.-Moulines-Priouret, 2011; F.-Moulines-Priouret,2012; F.-Moulines-Priouret-Vandekerkhove, 2012)

- Extensions of the sufficient conditions for "convergence in distribution" to the case
  - when NO uniform-in-θ ergodic behavior of the transition kernels $\{P_\theta\}_\theta$ i.e. neither the state space $\mathcal{X}$ nor the parameter space Θ have to be finite / countable / compact
  - without requiring convergence of the sequence $\{\theta_t\}_t$ as a preliminary step for the proof (when $\pi_\theta = \pi$)
  - without assuming the stability of the sequence $\{\theta_t\}_t$ as a preliminary step for the proof
  - each kernel may have its own invariant distribution, explicitly known or not.

- Based on strenghtened "containment" and "diminishing adaptation" conditions,
  - strong Law of Large Numbers for $\{f(X_t)\}_t$ and $\{f(\theta_t, X_t)\}_t$
  - Central Limit Theorem for $\{f(X_t)\}_t$ (see below)

Strong Law of Large Numbers

Under additional assumptions strenghtening the conditions between
- the diminishing adaptation condition

$$D_V(\theta_t, \theta_{t-1}) = \sup_x \frac{\|P_{\theta_t}(x, \cdot) - P_{\theta_{t-1}}(x, \cdot)\|_V}{V(x)}$$

- the rate of convergence of the kernels P_θ to stationarity
- the stability (control of growth in t) of the sequence $\{\theta_t\}_t$

- for any measurable function f such that $\sup_x |f|/V < \infty$

$$\lim_T \frac{1}{T} \sum_{t=1}^{T} f(X_t) = \lim_t \int f(x) \; \mathrm{d}\pi_{\theta_t}(x) \; a.s.$$

when the RHS exists a.s.

- Extensions: SLLN for $(x, \theta) \mapsto f(x, \theta)$.

Central Limit Theorem (1/2)

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left(f(X_t)-\int f \mathrm{d}\pi_{\theta_\star}\right)=\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left(f(X_t)-\int f\,\mathrm{d}\pi_{\theta_{t-1}}\right)+\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left(\int f\,\mathrm{d}\pi_{\theta_{t-1}}-\int f\,\mathrm{d}\pi_{\theta_\star}\right)$$

Under the assumptions
- each kernel P_θ is geometrically ergodic (drift, minorization)
- Trade off: diminishing adaptation, moment conditions, stability of $\{\theta_t\}_t$
- "containment": rate of ergodicity, moment conditions, stability of $\{\theta_t\}_t$

- CLT for the first part, with limiting variance given by

$$\sigma^2(f)=\lim_T \frac{1}{T}\sum_{t=1}^{T}F(\theta_t,X_t)$$

where <comment on the Poisson equation>

$$F(\theta,x)=P_\theta(\Lambda_\theta f)^2-(P_\theta\Lambda_\theta f)^2, \qquad \Lambda_\theta f=(I-P_\theta)^{-1}f$$

Central Limit Theorem (2/2)

For the second part:
- Restricted to algorithms satisfying <comment>

$$\mathbb{E}\left[f(X_{0:t})|\theta_{0:t-1}\right] = \int f(x_{0:t}) \; \mathrm{d}\nu(x_0) \prod_{j=1}^{t} P_{\theta_{j-1}}(x_{j-1}, \mathrm{d}x_j)$$

- Upon noting the linearization

$$\pi_\theta(f) - \pi_{\theta_\star}(f) = \pi_{\theta_\star}(P_\theta - P_{\theta_\star})\Lambda_{\theta_\star} f + \pi_\theta(P_\theta - P_{\theta_\star})\Lambda_{\theta_\star}(P_\theta - P_{\theta_\star})\Lambda_{\theta_\star} f$$

- Assuming: a CLT with variance $\gamma^2(f)$ for the first part, and a cvg in Prob to zero for the second part

- A global CLT with additive variance

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \left(f(X_t) - \int f \mathrm{d}\pi_{\theta_\star} \right) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2(f) + \gamma^2(f)\right)$$

As a conclusion of this part II

• A family of ergodic kernels $\{P_\theta\}_{\theta \in \Theta}$; to adapt the parameters θ_t, a strategy based on the past of the algorithm.

• The easiest situation:
- uniform-in-θ ergodicity conditions *(i.e. roughly: may be true if the sequence* $\{\theta_t\}_t$ *remains in a compact set ...)*

• Far more flexible but also more technical:
- an ergodic behavior depending on θ
- and the rate of growth of $t \mapsto |\theta_t|$ is controlled

• In both cases,
- the updating rule $\theta_t \longrightarrow \theta_{t+1}$ is s.t. the adaption is diminishing along iterations.