When Monte Carlo and Optimization met in a Markovian dance

Gersende Fort

CNRS

Institut de Mathématiques de Toulouse, France

ICTS "Advances in Applied Probability", Bengaluru, August 2019.

Intertwined, why ?

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 $\mathsf{X} \times \Theta$-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 $(\mathsf{X}, \mathcal{X})$
- a conditional distribution satisfying

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

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 $\theta \in \Theta \subseteq \mathbb{R}^p$; then few words on a more general situation.

Assumptions (1/3) Invariant distribution

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

Assumptions (2/3) (Generalized) Containment condition

- Uniform-in-θ ergodicity condition

$$\sup_{\theta \in \Theta} \|P_\theta^r(x;\cdot) - \pi_\theta\|_{\mathsf{TV}} \leq C \rho^r$$

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

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

<comment>

- A generalized condition: for any $\epsilon > 0$, there exists a non-decreasing sequence r_ϵ s.t. $\lim_t r_\epsilon(t)/t = 0$ and

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

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

$$\|P_\theta^r(x;\cdot) - \pi_\theta\|_{\mathsf{TV}} \leq C_\theta \, \rho_\theta^r$$

$$t^{-\tau} \ \|\theta_t\| < \infty \ \text{ a.s.} \qquad \limsup_t t^{-\tilde{\tau}} \ \left(C_{\theta_t} \vee (1-\rho_{\theta_t})^{-1}\right) < \infty \text{ a.s.}$$

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

- When uniform-in-θ ergodic condition, check

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

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

- Otherwise: 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 $\quad D(\theta, \theta') \leq C\,\|\theta - \theta'\|$
- 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,

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

<proof>

- Example: Adaptive Hastings Metropolis by Haario et al., 2001

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. $\lim_t \int f \, \mathrm{d}\pi_{\theta_t} = \alpha$

then (f bounded)

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

- 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.

- Example: Adaptive IS by Wang-Landau approaches <see Lecture 1>

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 ? A solution when f is bounded continous.

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 :

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

the steps are:

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

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

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

$\hookrightarrow$ Tool: Polish space 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 X nor the parameter space Θ have to be finite / countable / compact
- each kernel may have its own invariant distribution, explicitly known or not
- (when $\pi_\theta = \pi$) without requiring convergence of the sequence $\{\theta_t\}_t$ as a preliminary step for the proof
- without assuming the stability of the sequence $\{\theta_t\}_t$ as a preliminary step for the proof.

• Based on strenghtened "containment" and "diminishing adaptation" conditions,

- strong Law of Large Numbers for $\{f(X_t)\}_t$ and $\{f(\theta_t, X_t)\}_t$ <See lecture 3 for similar techniques>
- Central Limit Theorem for $\{f(X_t)\}_t$ (see below)

What can be said for more general Θ ? (1/5)

• 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, 2014)

Example: the Adaptive Equi-Energy sampler (2/5) extension of the EE sampler by Kou-Zhou-Wong, 2006

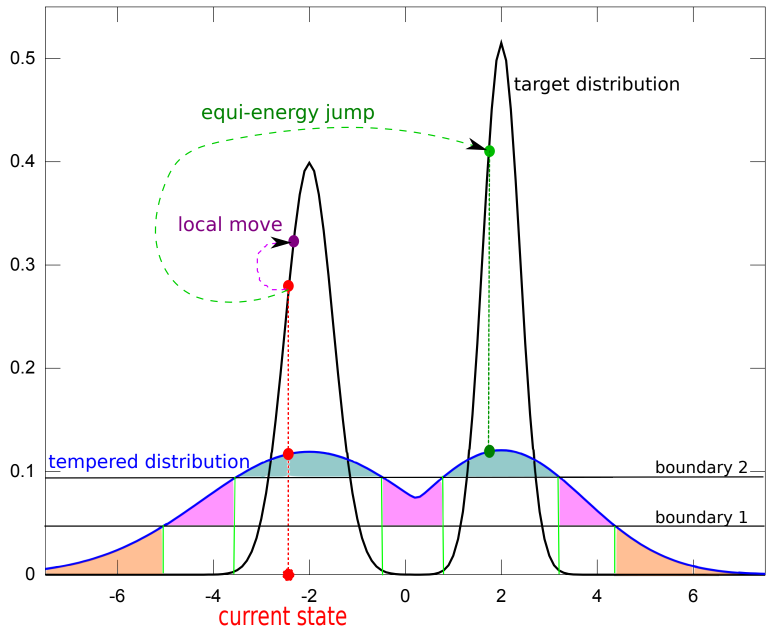
- Both **interacting** and **tempering** and **adaptive** algorithm.

- 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.

Example: the Adaptive Equi-Energy sampler (3/5)

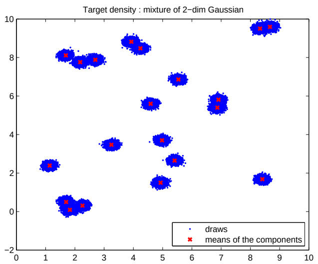


The equi-energy jump: (i) adaptive definition of the equi-energy rings as an estimation of the quantiles of $-\log \pi^{\beta_k}(Z)$ with $Z \sim \pi^{\beta_k}$; (ii) acceptance-rejection ratio;

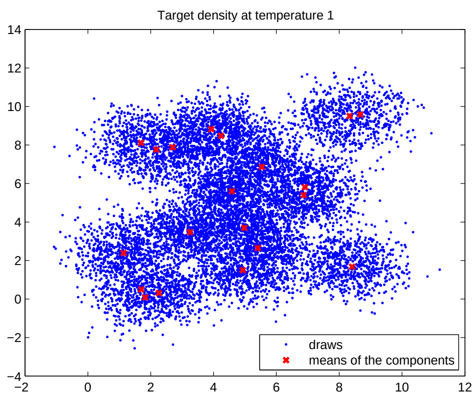
From chain $X^{(k)}$ to $X^{(k+1)}$:

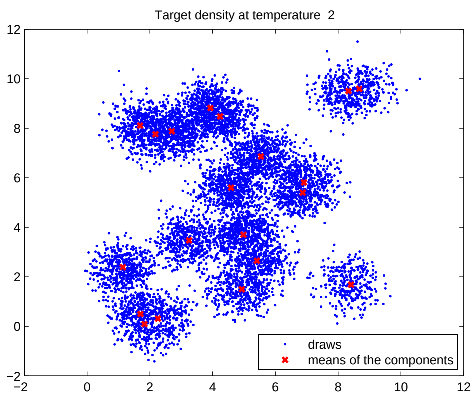
$$P_{\theta_{k,t}}(X_t^{(k+1)}, \cdot) = (1-\epsilon) \underbrace{Q(X_t^{(k+1)}, \cdot)}_{\text{MCMC with target } \pi^{\beta_{k+1}}} + \epsilon \underbrace{\tilde{Q}_{\theta_{k,t}}(X_t^{(k+1)}, \cdot)}_{\substack{\text{kernel depending on} \\ \text{the empirical distribution } \theta_{k,t} \\ \text{of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

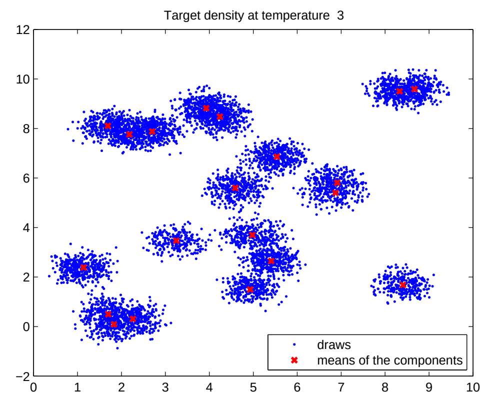
Example: the Adaptive Equi-Energy sampler (4/5)

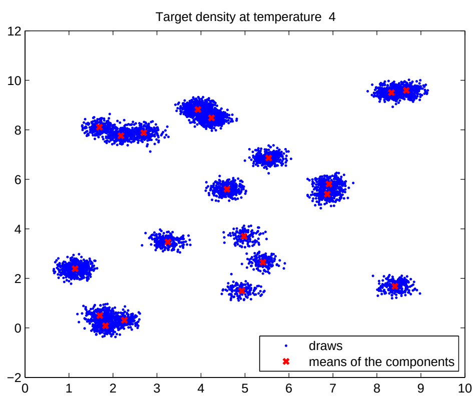


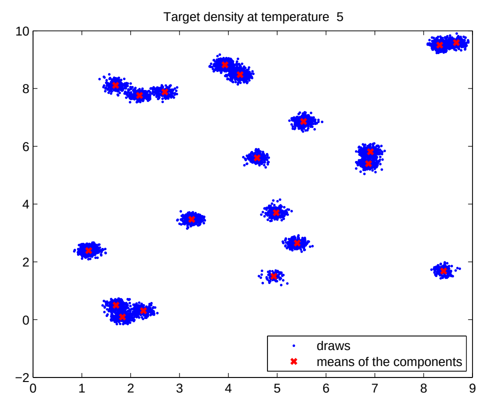
- target density : $\pi = \sum_{i=1}^{20} \mathcal{N}_2(\mu_i, \Sigma_i)$
- 5 parallel processes with target distribution π^{β_k} $(\beta_5 = 1)$

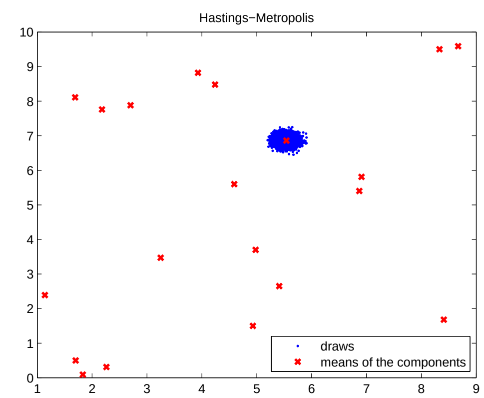












Example: the Adaptive Equi-Energy sampler (5/5)

- 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

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 ... <see lecture 3>)*

• 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.