When Monte Carlo and Optimization met in a Markovian dance

Gersende Fort

CNRS

Institut de Mathématiques de Toulouse, France

INSTITUT de MATHEMATIQUES de TOULOUSE

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

A dance, 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**

To make optimization methods tractable

- Intractable objective function

$$\theta \text{ s.t. } h(\theta) = 0 \qquad \text{when } h \text{ is not explicit } h(\theta) = \int_{\mathsf{X}} H(\theta, x)\, \mathrm{d}\pi_\theta(x)$$

or

$$\mathrm{argmin}_{\theta \in \Theta} \int_{\mathsf{X}} C(\theta, x)\, \mathrm{d}\pi_\theta(x)$$

- Intractable auxiliary quantities

Ex-1 Gradient-based methods

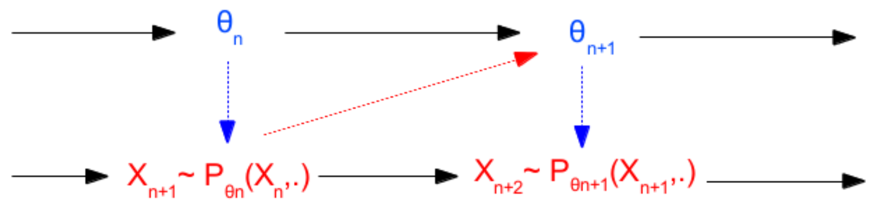
$$\nabla f(\theta) = \int_{\mathsf{X}} H(\theta, x)\ \mathrm{d}\pi_\theta(x)$$

Ex-2 Majorize-Minimization methods

$$\text{at iteration } t, \qquad f(\theta) \leq F_t(\theta) = \int_{\mathsf{X}} H_t(\theta, x)\ \mathrm{d}\pi_{t,\theta}(x)$$

- Strategies: Use Monte Carlo techniques to approximate the unknown quantities

In this talk, Markov !



- 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

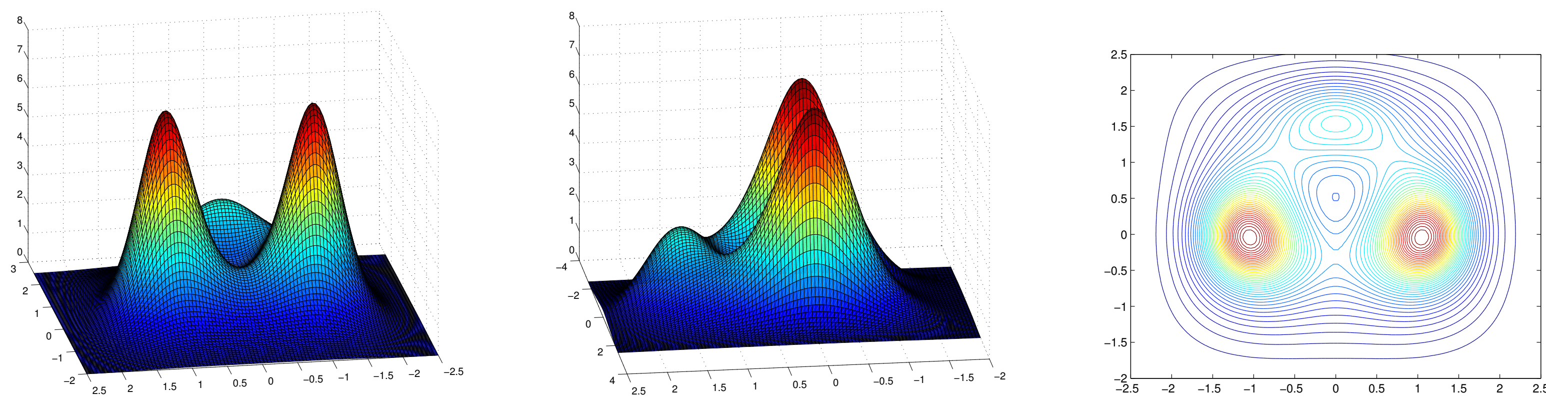
- (Talk) Application to a Computational Machine Learning pbm: penalized Maximum Likelihood through Stochastic Proximal-Gradient based methods

Part I: Motivating examples

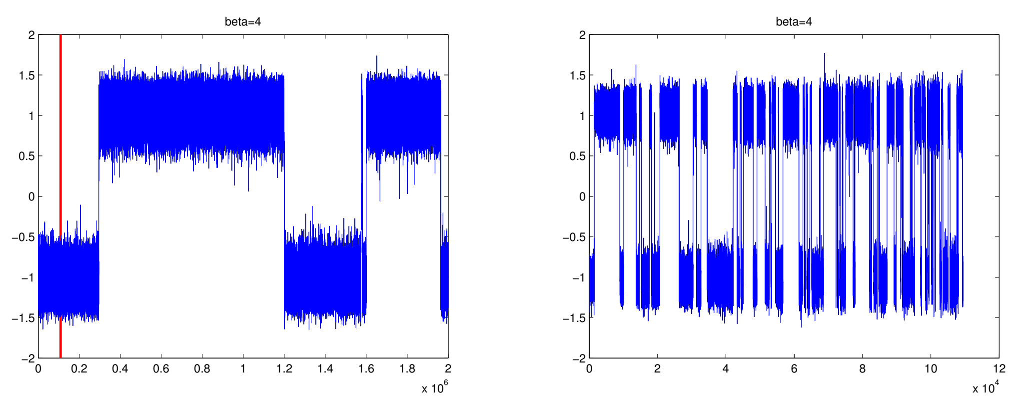
1st Ex. Adaptive Importance sampling by Wang-Landau approaches (1/6)

The problem

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



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



1st Ex. (2/6)

The strategy for choosing the proposal mecanism

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

- given a partition $\mathsf{X}_1, \cdots, \mathsf{X}_I$ of X,
- for any weight vector $\theta = (\theta(1), \cdots, \theta(I))$

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

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

- Unfortunately, $\theta_\star$ unavailable.

1st Ex. (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

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

- When approximating an expectation, set

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

1st Ex. (4/6)

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

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

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

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

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

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

- This update scheme is a normalized counter of the number of visits to X_i

<proof>

1st Ex. (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 $\theta_\star$

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

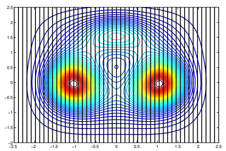
where $X_{t+1} \sim P_{\theta_t}(X_t, \cdot)$ and P_θ inv. wrt $\mathrm{d}\pi_\theta$.

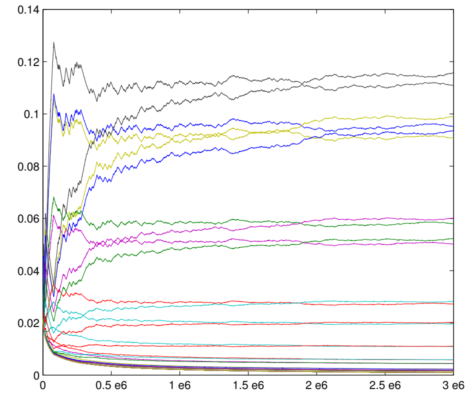
- Expected:

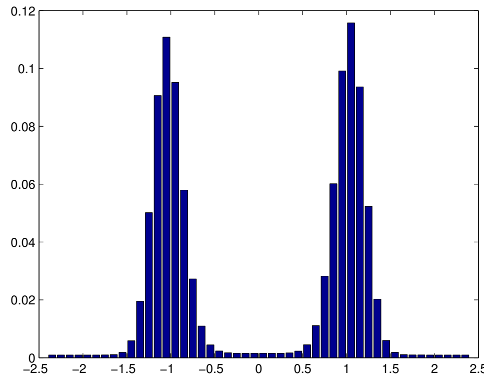
- the convergence of θ_t to $\theta_\star$: SA scheme, fed with adaptive (controlled) MCMC sampler,
- the convergence of the distribution of X_t to $\mathrm{d}\,\pi_{\theta_\star}$

1st Ex. (6/6)

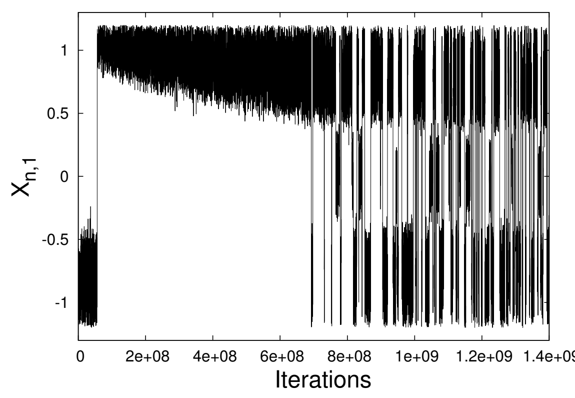
Does it work ? Plot: convergence of θ_t and first exit times from one mode

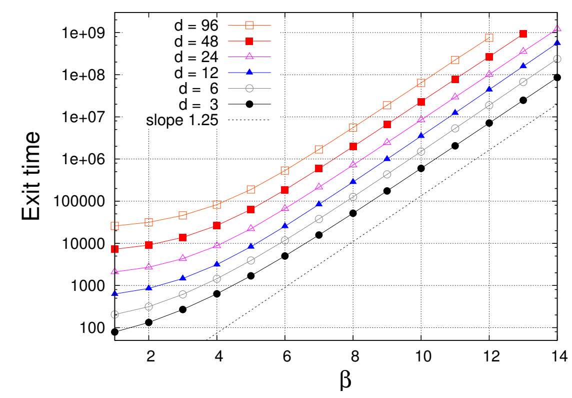






▶ see F, Kuhn, Jourdain, Lelièvre, Stoltz (2014); F, Jourdain, Lelièvre, Stoltz (2015,2017,2018) for studies of these Wang-Landau bases algorithms; including self-tuned SA update rules (γ_t is random).





Conclusion of the 1st example

- Iterative sampler

- Each iteration combines : (i) a sampling step $X_{t+1} \sim P_{\theta_t}(X_t, \cdot)$; and (ii) an optimization step 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

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

2nd Example: penalized ML in latent variable models (1/6)

- An example from Pharmacokinetic:

- N patients.
- At time 0: dose D of a drug.
- For patient #i, observations $Y_{i1}, \cdots, Y_{iJ_i}$ giving the evolution of the concentration at times $t_{i1}, \cdots, t_{iJ_i}$.

- The model:

$$Y_{ij} = \mathcal{F}\left(t_{ij}, X_i\right) + \epsilon_{ij} \qquad \epsilon_{ij} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$$

where $X_i \in \mathbb{R}^L$ is modeled as

$$X_i = Z_i \beta + d_i \in \mathbb{R}^L \qquad d_i \overset{i.i.d.}{\sim} \mathcal{N}_L(0, \Omega) \text{ and independent of } \epsilon_{\bullet}$$

and Z_i known matrix s.t. each row of X_i has in intercept (fixed effect) and covariates.

- Statistical analysis: (i) estimation of $\theta = (\beta, \sigma^2, \Omega)$, under sparsity constraints on β; (ii) selection of the covariates based on $\widehat{\beta}$.

2nd Ex. (2/6)

Penalized Maximum Likelihood

- The likelihood of $Y := \{Y_{ij}, 1 \leq i \leq N, 1 \leq j \leq J_i\}$ is not explicit:

- The distribution of $Y_{i,j}$ given X_i is simple; the distribution of X_i is simple.
- The joint distribution has an explicit expression - It is an example of latent variable model:

$$\log L(Y; \theta) = \log \int p(Y, x_{1:N}; \theta) \; \mathrm{d}\nu(x_{1:N})$$

- Sparsity constraints on the parameter θ: through a penalty term $g(\theta)$

- The penalized ML is of the form

$$\mathrm{argmin}_{\Theta} \left(-\log L(Y; \theta) + g(\theta) \right)$$

with an intractable objective function.

2nd Ex. (3/6)

What about first-order methods for solving the optimization ?

- On the likelihood term:

\- Usually regular enough so that the Gradient exists and <proof>

$$\nabla_\theta \log L(Y;\theta) = \int \frac{\partial_\theta \; p(Y,x;\theta)}{p(Y,x;\theta)} \; \frac{p(Y,x;\theta)\, \mathrm{d}\mu(x)}{\int p(Y,z;\theta)\, \mathrm{d}\mu(z)}$$

$$= \int \partial_\theta \left(\log p(Y,x;\theta)\right) \underbrace{\mathrm{d}\pi_\theta(x)}$$

the a posteriori distribution of x given Y
the dep upon Y is omitted

\- the a posteriori distribution is known up to a normalizing constant.

- On the penalty term

\- May be non smooth, but: convex and lower semi-continuous

\- Hence a Proximal operator (implicit gradient) is associated - <See the talk, on tuesday afternoon>.

2nd Ex. (4/6)

What about EM-like methods for solving the optimization ?

- Expectation-Maximization introduced to solve `below:  modified for a minimizati`

$$\text{argmin}_{\theta\in\Theta}\left(\log\int_{\mathsf{X}} p(x;\theta)\mathrm{d}\mu(x) - g(\theta)\right)$$

where the first part is untractable; by iterating two steps
- Expectation step

$$Q(\theta,\theta_t) := \int \log p(x;\theta)\ \frac{p(x;\theta_t)\ \mathrm{d}\mu(x)}{\int p(z;\theta_t)\ \mathrm{d}\mu(z)} = \int \log p(x;\theta)\ \mathrm{d}\pi_{\theta_t}(x)$$

- Minimization step

$$\theta_{t+1} := \text{argmin}_{\theta}\left(-Q(\theta,\theta_t) + g(\theta)\right).$$

- $\theta \mapsto Q(\theta,\theta_t)$ is an integral which is untractable; $\mathrm{d}\pi_\theta$ is known up to a normalizing constant.

see F,Moulines (2003); F,Ollier,Samson (2018)

2nd Ex. (5/6)

- Both in EM-like approaches and in gradient-based approaches,
- faced with untractable auxiliary quantities of the form

$$\int_{\mathsf{X}} H(\theta, x)\ \mathrm{d}\pi_{\theta_t}(x) \quad (1)$$

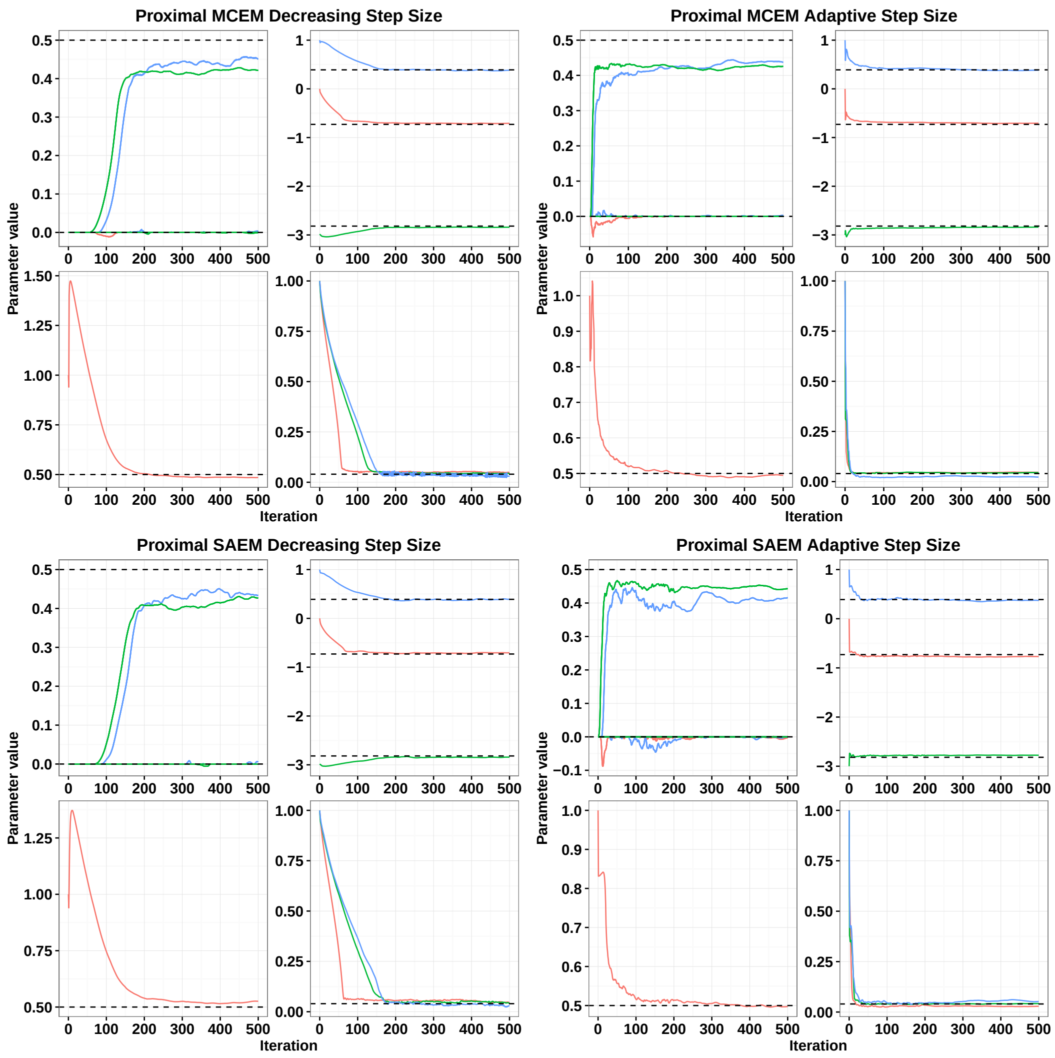
at itreration t of the optimization algorithm.
- untractable integral; $\mathrm{d}\pi_\theta$ is often known up to a normalizing constant.

- What kind of stochastic approximation of the integral (1) at iteration t ?
- Quadrature techniques: poor behavior w.r.t. the dimension of X
- I.i.d. samples from π_{θ_t} to define a Monte Carlo approximation: not possible, in general.
- use T samples from a MCMC sampler $\{X_{j,t+1}, j \geq 0\}$ with unique inv. dist. $\mathrm{d}\pi_{\theta_t}$.

2nd Ex. (6/6)

Does it work ?

see F,Moulines (2003) for EM-like approaches; see Atchadé,F,Moulines (2017) and F,Ollier,Samso (2018) for gradient-based approaches; see F,Ollier,Samson (2018) for the parallel between EM-like and Gradient-based techniques



Conclusion of the 2nd example

- Iterative optimization technique

- Each iteration combines : (i) an update of the parameter; (ii) a sampling step $X_{j+1,t+1} \sim P_{\theta_t}(X_{j,t+1}, \cdot)$ to approximate auxiliary quantities.

- The convergence of $\{\theta_t\}_t$ is the convergence of a `stochastically perturbed` iterative optimization algorithm. At each iteration: an exact quantity $\int H(\theta, x)\ \mathrm{d}\pi_{\theta_t}(x)$ is approximated by a Monte Carlo sum

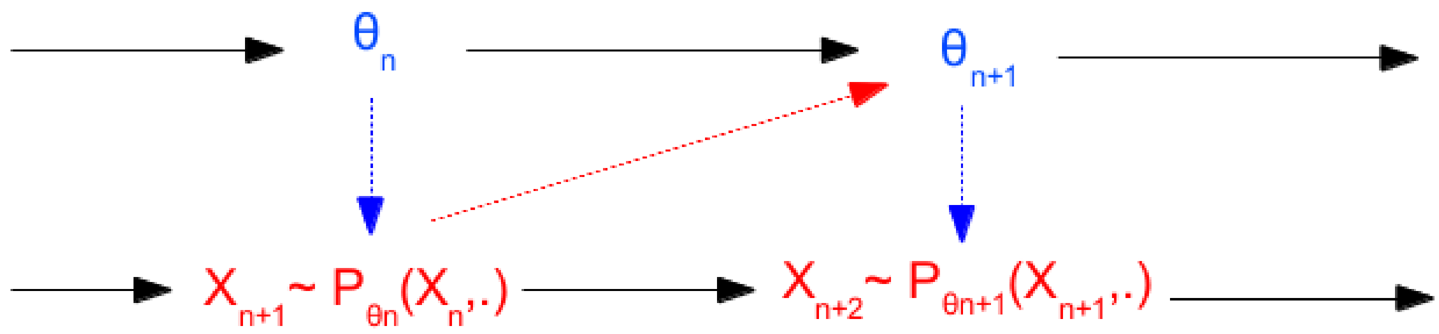
$$\int H(\theta, x)\ \mathrm{d}\pi_{\theta_t}(x) \approx \frac{1}{m_{t+1}} \sum_{j=1}^{m_{t+1}} H(\theta, X_{j,t+1})$$

- The points $\{X_{j,t+1}\}_j$ satisfy

$$\mathbb{E}\left[f(X_{j,t+1}) | \mathcal{F}_t\right] = P^j_{\theta_t}(X_{0,t+1}, \cdot) \qquad \mathcal{F}_t := \sigma(X_{:,0:t}, \theta_0), \quad X_{0,t+1} = X_{m_t,t}$$

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

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



Conclusion of this first part (2/3): is a theory required when sampling ?

YES ! convergence can be lost by the adaption mecanism

Even in a simple case when

$$\forall \theta \in \Theta, \qquad P_\theta \text{ invariant wrt } \mathrm{d}\pi,$$

one can define a simple adaption mecanism

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

such that

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

<proof> 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. $[t_1, t_0]$

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

YES ! Unfortunately ,

- a biased approximation <proof>

$$\mathbb{E}\left[\frac{1}{m_{t+1}}\sum_{j=1}^{m_{t+1}} H(\theta, X_{j,t+1})\Big|\mathcal{F}_t\right] = ? \neq \int_{\mathsf{X}} H(\theta, x)\, \mathrm{d}\pi_{\theta_t}(x)$$

- For a reduced computational cost: a bias which we would like NOT vanishing i.e. $m_t = m (= 1)$.

Ex. Stochastic Approximation with controlled Markovian dynamics

$$\begin{aligned}\theta_{t+1} &= \theta_t + \gamma_{t+1} H(\theta_t, X_{t+1}) \qquad X_{t+1} \sim P_{\theta_t}(X_t, \cdot) \\ &= \theta_t + \gamma_{t+1} \underbrace{\int H(\theta_t, x)\mathrm{d}\pi_{\theta_t}(x)}_{h(\theta_t)} + \gamma_{t+1} \underbrace{\Big(H(X_{t+1}, \theta_t) - h(\theta_t)\Big)}_{\text{non centered}}\end{aligned}$$