Monte Carlo methods and Optimization: Intertwinings

Gersende Fort

CNRS

Institut de Mathématiques de Toulouse, France

Monte Carlo Methods, Sydney, July 2019.

Intertwined, why ?

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

• 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

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

or

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

• 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

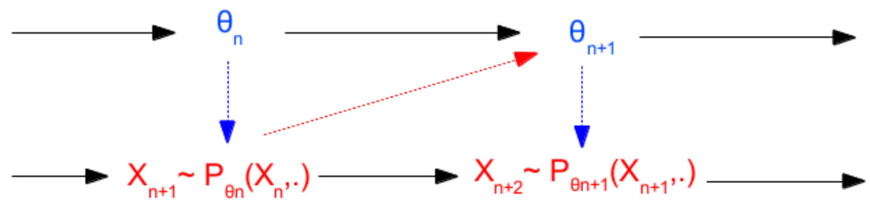
$$\nabla f(\theta) = \int_{\mathsf{X}} H(\theta, x) \, \mathrm{d}\pi_\theta(x)$$

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

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

Part I:

Theory of controlled (or adaptive) Markov chains

Example 1/ Adapted Markov chain Monte Carlo samplers

- Hastings-Metropolis algorithm, with Gaussian proposal and target $\mathrm{d}\pi$ on $\mathsf{X} \subseteq \mathbb{R}^d$

$$
\begin{array}{ll}
\text{Proposal:} & Y_{t+1} \sim \mathcal{N}_d(X_t, \theta) \\
\text{Accept-Reject} & X_{t+1} = \begin{cases} Y_{t+1} & \text{with probability } \alpha(X_t, Y_{t+1}) \\ X_t & \text{otherwise} \end{cases}
\end{array}
$$

summarized: $X_{t+1} \sim P_\theta(X_t, \cdot)$

- "Optimal" choice of the covariance matrix θ

$$\theta_{\mathrm{opt}} = \frac{(2.38)^2}{d} \, \mathrm{Cov}_\pi(X) = \frac{(2.38)^2}{d} \, \Gamma_{\mathrm{opt}}$$

Example 1 (to follow)/ Adapted Markov chain Monte Carlo samplers

- The algorithm

Sample $X_{t+1} \sim P_{\theta_t}(X_t, \cdot)$

SA scheme: $\Gamma_{t+1} =$ empirical cov matrix of $X_{1:t+1}$ computed from Γ_t, X_{t+1}

$\theta_{t+1} = (2.38)^2 d^{-1} \Gamma_{t+1}$

- In this example, a family of transition kernels $\{P_\theta, \theta \in \Theta\}$ and

$\forall \theta, P_\theta$ invariant w.r.t.π

- Convergence results: (Saksman-Vihola, 2010; F.-Moulines-Priouret, 2012)

- $\lim_t \theta_t = \theta_{\mathrm{opt}}$
- the distribution of $(X_t)_t$ converges to π (conditions on the tails of π)
- strong LLN, CLT for the samples $\{X_t\}_t$

Example 2/ Adapted Importance sampling by Wang-Landau approaches

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

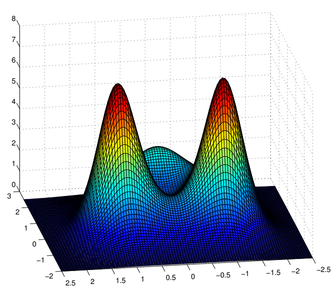
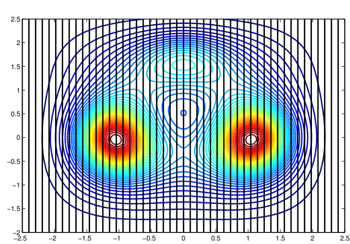
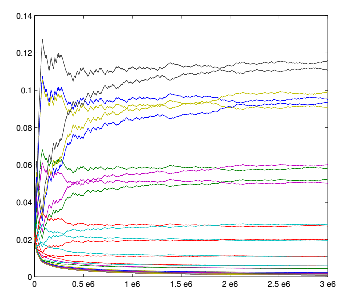
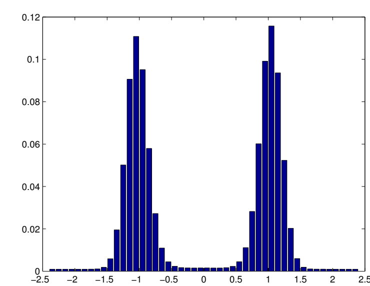
- A family of proposal mecanisms: Given a partition $\mathsf{X}_1, \cdots, \mathsf{X}_I$ of X,

$$\mathrm{d}\pi_\theta(x) \propto \sum_{i=1}^{I} 1_{\mathsf{X}_i}(x) \, \frac{\mathrm{d}\pi(x)}{\theta(i)}, \qquad \theta = (\theta(1), \cdots, \theta(I)) \text{ a weight vector}$$

- Optimal proposal: $\mathrm{d}\pi_{\theta_\star}$ with $\theta_\star(i) = \int_{\mathsf{X}_i} \mathrm{d}\pi(u)$,

- $\theta_\star$, unique limiting value of a Stochastic Approximation scheme

$$\text{with mean field } \int_{\mathsf{X}} H(\theta, X) \, \mathrm{d}\pi_\theta(x) \qquad \text{and } H_i(\theta, x) = \theta(i) \left(1_{\mathsf{X}_i}(x) - \sum_{j=1}^{I} \theta(j) 1_{\mathsf{X}_j}(x) \right).$$

Example 2 (to follow)/ Adapted Importance sampling by Wang-Landau approaches

- The algorithm

Sample: $X_{t+1} \sim P_{\theta_t}(X_t, \cdot)$, where $\pi_\theta P_\theta = \pi_\theta$

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

- In this example, a family of transition kernels $\{P_\theta, \theta \in \Theta\}$ such that

$\forall \theta, P_\theta$ invariant w.r.t. π_θ

- Convergence results: (F.-Jourdain-Lelievre-Stoltz-2015,2017,2018)

- θ_t converges to $\theta_\star$ a.s.;
- the distribution of X_t converges to $\mathrm{d}\pi_{\theta_\star}$;
- θ_t is an estimate of the importance ratio $[\mathrm{d}\pi/\mathrm{d}\pi_{\theta_\star}](x)$, constant along each X_i.

Is a "theory" required ?

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

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

such that

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

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. $[t_1, t_0]$

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

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

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

- Questions:

- convergence in distribution of X_t ?
- limit theorems

- Hereafter:

- focus on the convergence in distribution
- $\theta \in \Theta \subseteq \mathbb{R}^p$

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

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

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

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

- 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

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

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

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

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

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 $\epsilon > 0$,

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

- 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 (when $\pi_\theta = \pi$ for any θ)

Under these conditions, for any bounded function f,

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

In the literature

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

- 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 $\{f(X_t)\}_t$

- In the case $\theta \in \mathbb{R}^p$ but also in more general situations: θ may be a distribution

case of "interacting" MCMC. (Del Moral-Doucet, 2010)

- Results in the case each kernel P_θ has its own invariant distribution π_θ:

$$\lim_t \mathbb{E}\left[f(X_t)\right] = \lim_t \int f(x)\; \mathrm{d}\pi_{\theta_t}(x) \qquad \text{(RHS, assumed constant a.s.)}$$

As a conclusion of this part I

• A family of ergodic kernels; to adapt the parameters θ_t, a strategy based on the past of the algorithm

• The easiest situation:
- uniform-in-θ ergodicity conditions

• 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.

Part II.
Stochastic Approximation with Markovian dynamics

Stochastic Approximation (SA) methods

- Designed to solve on $\Theta \subseteq \mathbb{R}^p$: $h(\theta) = 0$ when h is not explicit but

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

- Algorithm:

- Choose: a deterministic positive (decreasing) sequence $\{\gamma_t\}_t$ s.t. $\sum_t \gamma_t = +\infty$
- Initialisation: $\theta_0 = \theta_{\text{init}} \in \Theta, X_0 = x_{\text{init}}$
- Until convergence:

$$X_{t+1} \sim P_{\theta_t}(X_t, \cdot) \qquad \theta_{t+1} = \theta_t + \gamma_{t+1}\, H(\theta_t, X_{t+1})$$

where P_θ inv. wrt π_θ.

Beware! a **biased** approximation

$$\mathbb{E}\left[H(\theta_t, X_{t+1})|\mathcal{F}_t\right] - h(\theta_t) = \int_{\mathsf{X}} \Big(P_{\theta_t}(X_t, \mathrm{d}x) - \mathrm{d}\pi_{\theta_t}(x)\Big)\, H(\theta_t, x)$$

Convergence analysis for SA: the successive steps

1- The sequence $\{\theta_t\}_t$ is stable i.e. (w.p.1) there exists a compact subset $\mathcal{K}$ of Θ such that $\theta_t \in \mathcal{K}$ for any t.
2- Convergence of $\{\theta_t\}_t$ to $\mathcal{L}$ (or to a connected component of $\mathcal{L}$; or to a point $\theta_\star \in \mathcal{L}$).

- Required: there exists a non-negative Lyapunov function V:

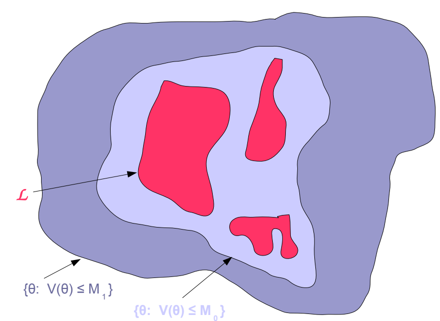
$$V(\theta_{t+1}) \leq V(\theta_t) - \gamma_{t+1}\, \phi^2(\theta_t) + \gamma_{t+1}\, \underbrace{W_{t+1}}_{\text{signed}}.$$

whose level sets are compact subsets of Θ, and ϕ is s.t. that

$$\inf_{\text{compact} \subset \Theta \setminus \mathcal{L}} \phi^2 > 0 \quad \text{with } \mathcal{L} := \{\phi^2 = 0\} \subset \{V \leq M_\star\}.$$

Control of the "noise":

$$\sup_t \left| \sum_{k=1}^{t} \gamma_{k+1} \left(H(\theta_k, X_{k+1}) - h(\theta_k) \right) \right|$$



Stability: a crucial point - Different strategies

- Stable by definition:

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

quite unlikely

- Force the stability by a projection on a compact subset $\mathcal{K}$

$$\theta_{t+1} = \Pi_{\mathcal{K}}\left(\theta_t + \gamma_{t+1} H(\theta_t, X_{t+1})\right)$$

Limiting points: in $\mathcal{L} \cap \mathcal{K}$. *How to choose* $\mathcal{K}$ *?*

- Use the Chen's technique: projection on growing compact subsets.

(Chen-Zhu, 1986)

Self-stabilized Stochastic Approximation (the Chen's technique)

Choose compact subsets $\{\mathcal{K}_i\}_{i \geq 0}$ s.t. $\bigcup_i \mathcal{K}_i = \Theta$ and $\mathcal{K}_i \subset \mathcal{K}_{i+1}$.

- (Start - Block 1):

$\theta_0 = \theta_{\text{init}} \in \mathcal{K}_0$ and $X_0 = x_{\text{init}}$ and repeat for $t \geq 0$

$$X_{t+1} \sim P_{\theta_t}(X_t, \cdot) \qquad \theta_{t+1} = \theta_t + \gamma_{t+1} H(\theta_t, X_{t+1})$$

until $\theta_{t+1} \notin \mathcal{K}_0$. Set $T_1 = t + 1$.

- $\cdots$

- (Stop & re-start, Block $q+1$)

$\theta_{T_q} = \theta_{\text{init}}, \quad X_{T_q} = x_{\text{init}}$ and repeat for $t \geq 0$

$$X_{T_q+t+1} \sim P_{\theta_{T_q+t}}(X_{T_q+t}, \cdot) \qquad \theta_{T_q+t+1} = \theta_{T_q+t} + \gamma_{q+t+1} H(\theta_{T_q+t}, X_{T_q+t+1})$$

until $\theta_{T_q+t+1} \notin \mathcal{K}_q$. Set $T_{q+1} = T_q + t + 1$.

- $\cdots$

When does self-stabilization SA "work" ? (1/3)

- If the number of "stop & re-start" is finite, it works !

then there exists L s.t.
(a) $\{\theta_t\}_t$ is in the compact set $\mathcal{K}_L$
(b) for any $t \geq 0$

$$X_{T_L+t+1} \sim P_{\theta_{T_L+t}}(X_{T_L+t}, \cdot) \qquad \theta_{T_L+t+1} = \theta_{T_L+t} + \gamma_{L+t+1} H(\theta_{T_L+t}, X_{T_L+t+1})$$

- If it is not: as if with $\rho_{t+1} \leftarrow \gamma_{L+t+1}$ for arbitrarily large L:

$$\theta_0 = \theta_{\mathrm{init}}, X_0 = x_{\mathrm{init}}, \quad X_{t+1} \sim P_{\theta_t}(X_t, \cdot), \qquad \theta_{t+1} = \theta_t + \rho_{t+1} H(\theta_t, X_{t+1})$$

if it is not finite (2/3)

•Lemma. Assume that h is continuous and there exists a C^1 non-negative function V s.t.

- the level sets $\{V \leq M\}$ are compact subset of Θ;
- the set $\mathcal{L} = \{\langle \nabla V ; h \rangle = 0\}$ is compact;
- and on $\mathcal{L}^c$, $\langle \nabla V ; h \rangle < 0$.

Let $\theta_{\text{init}} \in \mathcal{K}_{\prime}$. Let M_0 be s.t. $\mathcal{K}_0 \cup \mathcal{L} \subset \{V \leq M_0\}$.

There exist $\delta, \lambda > 0$ such that

$$\left[\sup_{1 \leq k \leq t} \rho_k \leq \lambda, \sup_{1 \leq k \leq t} | \sum_{j=1}^{k} \rho_j \left(H(\theta_j, X_{j+1}) - h(\theta_j) \right) | \leq \delta \right] \Longrightarrow \theta_{1:t} \in \{V \leq M_0 + 1\}.$$

if it is not finite (3/3)

- Prove for any **compact subset** $\mathcal{K}$

$$\lim_{L\to\infty} \mathbb{P}_{(x_{\text{init}},\theta_{\text{init}}),\gamma_{L+\bullet}} \left(\sup_{k\geq 1} 1_{\theta_{1:k}\in\mathcal{K}} \left| \sum_{j=1}^{k} \gamma_{L+j} \left(H(\theta_j, X_{j+1}) - h(\theta_j) \right) \right| > \delta \right) = 0.$$

- Apply the B-T inequality

$$\mathbb{E}_{(x_{\text{init}},\theta_{\text{init}})} \left[\sup_{k\geq 1} \ 1_{\theta_{1:k}\in\mathcal{K}} \ \left| \sum_{j=1}^{k} \rho_j \left(H(\theta_j, X_{j+1}) - h(\theta_j) \right) \right| \right]$$

- Use the decomposition below and use properties on **controlled** Markov chains since $X_{j+1} \sim P_{\theta_j}(X_j, \cdot)$.

The Poisson equation: $\widehat{H}_\theta$ s.t. $\widehat{H}_\theta(x) - P_\theta \widehat{H}_\theta(x) = H(\theta, x) - h(\theta)$.

$$\sum_{j=1}^{k} \rho_j \left(H(\theta_j, X_{j+1}) - h(\theta_j) \right) = \sum_{j=1}^{k} \rho_j \left(\widehat{H}_{\theta_j}(X_{j+1}) - P_{\theta_j} \widehat{H}_{\theta_j}(X_j) \right)$$
$$+ \sum_{j=1}^{k} \rho_j \left(P_{\theta_j} \widehat{H}_{\theta_j}(X_j) - P_{\theta_{j+1}} \widehat{H}_{\theta_{j+1}}(X_{j+1}) \right) + \sum_{j=1}^{k} \rho_j \left(P_{\theta_{j+1}} \widehat{H}_{\theta_{j+1}}(X_{j+1}) - P_{\theta_j} \widehat{H}_{\theta_j}(X_{j+1}) \right)$$

In the literature, SA with Markovian dynamics

(F,2015; F.-Moulines-Schreck-Vihola,2016; Morral-Bianchi-F.,2017; Crepey-F.-Gobet-Stazhinski,2018)

• In the case $\theta \in \mathbb{R}^p$,
- Sufficient conditions for the convergence
- Central Limit Theorems (along a converging path) for both the sequence $\{\theta_t\}_t$ and the averaged sequence

$$\bar{\theta}_t = \frac{1}{t} \sum_{k=1}^{t} \theta_k$$

- Distributed SA

• Some results in the infinite dimensional framework for θ; with i.i.d. dynamics.

Part III:
Stochastic Proximal-Gradient algorithms

Penalized Maximum Likelihood inference

- An intractable log-likelihood of the observations $Y_{1:n}$

- Ex: Latent variable models

$$\ell(Y_{1:n};\theta) = \log \int p(Y_{1:n}, x; \theta)\ \mathrm{d}\nu(x)$$

- A sparsity condition on θ through a **non smooth and convex** penalty

- Ex-1: $g(\theta) = \lambda \|\theta\|_1$

- Solve

$$\mathrm{argmin}_\theta \left(\underbrace{f(\theta)}_{\text{smooth, intractable}} + \underbrace{g(\theta)}_{\text{non smooth, convex, tractable}} \right)$$

Monte Carlo approximations for gradient-based optimization methods

- In this "latent variable model" example, as in many examples:

$$\nabla f(\theta) = \int H(\theta, x) \; \mathrm{d}\pi_\theta(x)$$

where π_θ: (the a posteriori) distribution known up to a normalization constant

(dependance upon $Y_{1:n}$ omitted)

$\hookrightarrow$ intractable integral.

- If the gradient were available: iterative algorithm

$$u_{t+1} = \mathrm{Prox}_{\gamma_{t+1}\, g}\left(u_t - \gamma_{t+1} \nabla f(u_t)\right) \qquad \mathrm{Prox}_{\gamma\, g}(\tau) = \mathrm{argmin}_u \left(g(u) + \frac{1}{2\gamma}\|u-\tau\|^2\right)$$

- Since it is not: iterative algorithm

$$\theta_{t+1} = \mathrm{Prox}_{\gamma_{t+1}\, g}\left(\theta_t - \gamma_{t+1} \; \frac{1}{m_{t+1}} \sum_{k=1}^{m_{t+1}} H(\theta_t, X_{t+1,k})\right) \qquad X_{t+1,k} \sim P_{\theta_t}(X_{t+1,k-1}, \cdot)$$

Questions

- Does the stochastic version inherit the same asymptotic behavior as the (exact) Gradient-Proximal algorithm ? i.e. convergence of $\{\theta_t\}_t$

- How to choose the stepsize sequence $\{\gamma_t\}_t$?

- How to choose the number of Monte Carlo samples m_t ? Is the "SA regime" (i.e. $m_t = 1$) possible ?

- What about the rate of convergence ?

- Is the rate improved by Nesterov-based acceleration ? is it improved by Averaging techniques ?

Assumptions

- On the non-smooth part: $g : \mathbb{R}^p \to [0, \infty]$, is not identically $+\infty$, convex and lower semi-continuous.

- On the smooth part: $f : \mathbb{R}^p \to \mathbb{R}$ is **convex**, C^1 on $\mathbb{R}^p$ and there exists L such that for any θ, θ'

$$\|\nabla f(\theta) - \nabla f(\theta')\| \leq L \ \|\theta - \theta'\|$$

- On the solution set: $\mathcal{L} := \mathrm{argmin}_\theta (f + g) = \{\theta = \mathrm{Prox}_{\gamma\, g}(\theta - \gamma \nabla f(\theta))\}$ is a non empty subset of $\Theta = \{g < \infty\}$.

- On the stepsize: $\sum_t \gamma_t = \infty$

- On the perturbation $\eta_{t+1} := m_{t+1}^{-1} \sum_{j=1}^{m_{t+1}} H(\theta_t, X_{t+1,j}) - h(\theta_t)$: the series

$$\sum_t \gamma_t \eta_t, \qquad \sum_t \gamma_t^2 \|\eta_t\|^2, \qquad \sum_t \gamma_t \left\langle T_{\gamma_t}(\theta_{t-1}); \eta_t \right\rangle$$

converge

Results (Atchade-F-Moulines, 2017)

$$\theta_{t+1} = \mathrm{Prox}_{\gamma_{t+1}\, g}\left(\theta_t - \gamma_{t+1}\ \frac{1}{m_{t+1}} \sum_{k=1}^{m_{t+1}} H(\theta_t, X_{t+1,k})\right)$$

- Convergence of the iterates $\{\theta_t\}_t$: there exists $\theta_\star \in \mathcal{L}$ s.t. $\lim_t \theta_t = \theta_\star$.

- For non-negative weights $\{a_{k,t}\}_k$ s.t. $\sum_{k=1}^t a_{k,t} = 1$, an explicit upper bound of

$$(f+g)\left(\bar{\theta}_t\right) - \min(f+g) \leq \sum_{k=1}^{t} a_{k,t}\ (f+g)(\theta_k) - \min(f+g) \leq \cdots$$

where

$$\bar{\theta}_t = \sum_{k=1}^{t} a_{k,t}\, \theta_k$$

Rates of convergence on the functional $(f+g)(\theta_t) - \min(f+g)$

- Rate of the exact algorithm: $O(1/t)$

- Stochastic version with increasing batch size

- After t iterations, the same rate by choosing

$$\gamma_t = \gamma \qquad m_t = t \qquad \bar{\theta}_t = t^{-1} \sum_{k=1}^{t} \theta_k$$

- BUT the total Monte Carlo cost is $O(t^2)$: complexity $O(1/\sqrt{t})$.

- Stochastic version with fixed batch size

- After t iteratons, a rate $O(1/\sqrt{t})$ by choosing

$$\gamma_t = t^{-1/2} \qquad m_t = m \qquad \bar{\theta}_t = t^{-1} \sum_{k=1}^{t} \theta_k$$

- the total Monte Carlo cost is $O(t)$: complexity $O(1/\sqrt{t})$.

Nesterov's acceleration, rate of convergence of the functional

$$u_{t+1} = \mathrm{Prox}_{\gamma_{t+1}\,g}\Big(\vartheta_t - \gamma_{t+1}\;\nabla f(\vartheta_t)\Big) \qquad\qquad \vartheta_t = u_t + \frac{\mu_{t-1}-1}{\mu_t}(u_t - u_{t-1})$$

where $\mu_t = O(t)$.

- Rate of the exact algorithm: $O(1/t^2)$

- Stochastic version with increasing batch size

- After t iterations, the same rate by choosing

$$\gamma_t = \gamma \qquad\qquad m_t = t^3 \qquad\qquad \theta_t$$

- BUT the total Monte Carlo cost is $O(t^4)$: complexity $O(1/\sqrt{t})$.

Conclusion

(F.-Risser-Atchade-Moulines,2018;F-Ollier-Samson,2019)

Given a Monte Carlo budget t:

- The (perturbed) Proximal-Gradient combined with averaging has the same complexity as the (perturbed) Nesterov-accelerated Proximal-Gradient: $O(1/\sqrt{t})$

- Nesterov-accelerated Proximal-Gradient + weighted averaging strategies: no improvement

- Nesterov-accelerated Proximal-Gradient + other relaxations $\mu_t = O(t^d)$ for some $d \in (0,1)$: no improvement

Joint works with

- Yves Atchade, Univ. Michigan, France
- Jean-François Aujol, Univ. Bordeaux, France
- Stéphane Crepey, Univ. Evry, France
- Charles Dossal, Univ. Toulouse, France
- Pierre Gach, Univ. Toulouse, France
- Emmanuel Gobet, Ecole Polytechnique, France
- Benjamin Jourdain, ENPC, France
- Tony Lelievre, ENPC, France
- Eric Moulines, Ecole Polytechnique, France
- Pierre Priouret, Univ. Paris 6, France
- Laurent Risser, Univ. Toulouse, France
- Adeline Samson, Univ. Grenoble-Alpes, France
- Amandine Schreck, Telecom ParisTech, France
- Gabriel Stoltz, ENPC, France
- Pierre Vandekerkhove, Univ. Marne-la-Vallée, France
- Matti Vihola, Univ. Jyvaskyla, Finland