Algorithmes Gradient-Proximaux pour l'inférence statistique

Gersende Fort

Institut de Mathématiques de Toulouse, CNRS
Toulouse, France

Based on joint works with

- Yves Atchadé (Univ. Michigan, USA)
- Jean-François Aujol (IMB, Bordeaux, France)
- Eric Moulines (Ecole Polytechnique, France)
- Adeline Samson et Edouard Ollier (Univ. Grenoble Alpes, France).
- Charles Dossal (IMT).
- Laurent Risser (IMT)

↪ On Perturbed Proximal-Gradient algorithms (JMLR, 2017)
↪ Stochastic Proximal Gradient Algorithms for Penalized Mixed Models (arXiv 1704.08891v2)
↪ Acceleration for perturbed Proximal Gradient algorithms (work in progress)
↪ Algorithmes Gradient Proximaux Stochastiques (GRETSI, 2017)

Outline

Motivation 1: Pharmacokinetic (1/2)

- N patients.
- At time 0: dose D of a drug.
- For patient i, evolution of the concentration at times $t_{ij}, 1 \leq j \leq J_i$: observations $\{Y_{ij}, 1 \leq j \leq J_i\}$.

Model:

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

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

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

Motivation 1: Pharmacokinetic (1/2)

- N patients.
- At time 0: dose D of a drug.
- For patient i, evolution of the concentration at times $t_{ij}, 1 \leq j \leq J_i$: observations $\{Y_{ij}, 1 \leq j \leq J_i\}$.

Model:

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

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

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

Example of model $\mathcal{F}$: monocompartimental, with digestive absorption

$$\mathcal{F}(t, [\ln \mathsf{Cl}, \ln \mathsf{V}, \ln \mathsf{A}]) = \mathcal{C}(\mathsf{Cl},\mathsf{V},\mathsf{A},\mathsf{D}) \left(\exp(-\frac{\mathsf{Cl}}{\mathsf{V}}t) - \exp(-\mathsf{A}t)\right)$$

For each patient i,

$$\begin{bmatrix} \ln Cl \\ \ln V \\ \ln A \end{bmatrix}_i = \begin{bmatrix} \beta_{0,Cl} \\ \beta_{0,V} \\ \beta_{0,A} \end{bmatrix} + \begin{bmatrix} \beta_{1,Cl} Z^i_{1,Cl} + \cdots + \beta_{K,Cl} Z^i_{K,Cl} \\ \text{idem, with covariates } Z^i_{k,V} \text{ and coefficients } \beta_{k,V} \\ \text{idem, with covariates } Z^i_{k,A} \text{ and coefficients } \beta_{k,A} \end{bmatrix} + \begin{bmatrix} d_{\mathsf{Cl},i} \\ d_{\mathsf{V},i} \\ d_{\mathsf{A},i} \end{bmatrix}$$

Motivation 1: Pharmacokinetic (1/2)

- N patients.
- At time 0: dose D of a drug.
- For patient i, evolution of the concentration at times $t_{ij}, 1 \leq j \leq J_i$: observations $\{Y_{ij}, 1 \leq j \leq J_i\}$.

Model:

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

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

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

Statistical analysis:

- estimation of $\theta = (\beta, \sigma^2, \Omega)$, under sparsity constraints on β
- selection of the covariates based on $\hat{\beta}$.

$\hookrightarrow$ Penalized Maximum Likelihood

Motivation : Pharmacokinetic (2/2)

Model:

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

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

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

Likelihoods:

- Complete likelihood: the distribution of $\{Y_{ij}, X_i; 1 \leq i \leq N, 1 \leq j \leq J_i\}$ has an explicit expression.

$$\left(\prod_{i=1}^{N}\prod_{j=1}^{J_i}\mathcal{N}(f(t_{ij}, X_j), \sigma^2)[Y_{ij}]\right) \quad \left(\prod_{i=1}^{N}\mathcal{N}_L(Z_i\beta, \Omega)[X_i]\right)$$

- Likelihood: the distribution of $\{Y_{ij}; 1 \leq i \leq N, 1 \leq j \leq J_i\}$ is **not explicit**.
- ML: here, the **likelihood is not concave**.

General case: Latent variable models

- The log-likelihood of the observations Y is of the form (dependende upon Y is omitted)

$$\theta \mapsto \log L(\theta) \qquad L(\theta) = \int_{\mathsf{X}} p_\theta(x)\, \mu(\mathrm{d}x),$$

  where μ is a σ-finite positive measure on a set X.
- x collects the missing/latent data. previous example: $x \leftarrow (X_1, \cdots, X_N)$, $\mu \leftarrow$ lebesgue on $\mathbb{R}^{LN}$

In these models,

- the complete likelihood $p_\theta(x)$ can be evaluated explicitly,
- the likelihood has no closed expression.
- The exact integral could be replaced by a Monte Carlo approximation ; known to be inefficient.
  Numerical methods based on the a posteriori distribution of the missing data are preferred (see e.g. Expectation-Maximization approaches).

$\hookrightarrow$ What about the gradient of the (log)-likelihood ?

Latent variable model: Gradient of the likelihood

$$\log L(\theta) = \log \int p_\theta(x)\, \mu(\mathrm{d}x)$$

Under regularity conditions, $\theta \mapsto \log L(\theta)$ is C^1 and

$$\begin{aligned}
\nabla \log L(\theta) &= \frac{\int \partial_\theta p_\theta(x)\, \mu(\mathrm{d}x)}{\int p_\theta(z)\, \mu(\mathrm{d}z)} \\
&= \int \partial_\theta \log p_\theta(x) \quad \underbrace{\frac{p_\theta(x)\, \mu(\mathrm{d}x)}{\int p_\theta(z)\, \mu(\mathrm{d}z)}}_{\text{the a posteriori distribution}}
\end{aligned}$$

Latent variable model: Gradient of the likelihood

$$\log L(\theta) = \log \int p_\theta(x)\, \mu(\mathrm{d}x)$$

Under regularity conditions, $\theta \mapsto \log L(\theta)$ is C^1 and

$$\begin{aligned}\nabla \log L(\theta) &= \frac{\int \partial_\theta p_\theta(x)\, \mu(\mathrm{d}x)}{\int p_\theta(z)\, \mu(\mathrm{d}z)} \\ &= \int \partial_\theta \log p_\theta(x) \quad \underbrace{\frac{p_\theta(x)\, \mu(\mathrm{d}x)}{\int p_\theta(z)\, \mu(\mathrm{d}z)}}_{\text{the a posteriori distribution}}\end{aligned}$$

The gradient of the log-likelihood

$$\nabla_\theta \left\{\log L(\theta)\right\} = \int H_\theta(x)\ \pi_\theta(\mathrm{d}x)$$

is an untractable expectation w.r.t. the conditional distribution of the latent variable given the observations Y *(known up to a constant)*
For all (x, θ), $H_\theta(x)$ *can be evaluated.*

Motivation 2: relationships in a graph (1/2)

- p nodes in a graph (e.g. p senators from the US congress)
- each node takes values in $\{-1, 1\}$ (e.g. each node codes for no/yes in a vote)
- N pictures of the graph (e.g. N votes)

Model: Each observation $Y^{(i)} \in \{-1, 1\}^p$; i.i.d. observations with distribution

$$\pi_\theta(y) \propto \exp\left(\sum_{i=1}^{p} \theta_i y_i + \sum_{i=1}^{p-1} \sum_{j=i+1}^{p} \theta_{ij} y_i y_j\right)$$

Statistical Analysis:

- estimation of θ, under penalty (sparse graph, regularization $N << p^2/2$)
- classification of the nodes

$\hookrightarrow$ Penalized Maximum Likelihood

Motivation 2: relationships in a graph (2/2)

Model: Each observation $Y^{(n)} \in \{-1,1\}^p$; i.i.d. observations with distribution

$$\pi_\theta(y) = \frac{1}{Z_\theta} \exp\left(\sum_{i=1}^{p} \theta_i y_i + \sum_{i=1}^{p-1} \sum_{j=i+1}^{p} \theta_{ij} y_i y_j \right)$$

Log-Likelihood: $\mathbf{Y} \overset{\text{def}}{=} (Y^{(1)}, \cdots, Y^{(N)})$

$$\begin{aligned} \ell(\theta) &= \sum_{i=1}^{p} \theta_i \left(\sum_{n=1}^{N} Y_i^{(n)} \right) + \sum_{i=1}^{p-1} \sum_{j=i+1}^{p} \theta_{ij} \left(\sum_{n=1}^{N} Y_i^{(n)} Y_j^{(n)} \right) - N \log Z_\theta \\ &= \langle \Theta, S(\mathbf{Y}) \rangle - N \log Z_\theta \\ &= \langle \Psi(\theta), S(\mathbf{Y}) \rangle + \Phi(\theta) \end{aligned}$$

- Likelihood : **not** explicit

$$Z_\theta \overset{\text{def}}{=} \sum_{y \in \{-1,1\}^p} \exp\left(\sum_{i=1}^{p} \theta_i y_i + \sum_{i=1}^{p-1} \sum_{j=i+1}^{p} \theta_{ij} y_i y_j \right)$$

- ML: here, the likelihood is concave.

General Case: Discrete graphical models

N independent observations of an undirected graph with p nodes.
Each node takes values in a finite alphabet X.

- N i.i.d. observations $Y^{(i)}$ in X^p with distribution

$$y = (y_1, \cdots, y_p) \mapsto \pi_\theta(y) \stackrel{\text{def}}{=} \frac{1}{Z_\theta} \exp\left(\sum_{k=1}^{p} \theta_{kk} B(y_k, y_k) + \sum_{1 \le j < k \le p} \theta_{kj} B(y_k, y_j) \right)$$
$$= \frac{1}{Z_\theta} \exp\left(\left\langle \theta, \bar{B}(y) \right\rangle \right)$$

  where $\bar{B}$ is a symmetric function.
- θ is a symmetric $p \times p$ matrix.
- the normalizing constant (partition function) Z_θ can not be computed - sum over $|\mathsf{X}|^p$ terms.

Markov random field: Likelihood

- Likelihood of the form (scalar product between matrices = Frobenius inner product)

$$\frac{1}{N}\log L(\theta) = \left\langle \theta, \frac{1}{N}\sum_{i=1}^{N} \bar{B}(Y_i) \right\rangle - \log Z_\theta$$

The likelihood is untractable.

Markov random field: Gradient of the likelihood

- Gradient of the form

$$\nabla_\theta \left(\frac{1}{N} \log L(\theta) \right) = \frac{1}{N} \sum_{i=1}^{N} \bar{B}(Y_i) - \int_{\mathsf{X}^p} \bar{B}(y)\, \pi_\theta(y)\, \mu(\mathrm{d}y)$$

with

$$\pi_\theta(y) \stackrel{\text{def}}{=} \frac{1}{Z_\theta} \exp\left(\left\langle \theta, \bar{B}(y) \right\rangle \right).$$

The gradient of the (log)-likelihood is untractable

Markov random field: Gradient of the likelihood

- Gradient of the form

$$\nabla_\theta \left(\frac{1}{N} \log L(\theta) \right) = \frac{1}{N} \sum_{i=1}^{N} \bar{B}(Y_i) - \int_{\mathsf{X}^p} \bar{B}(y)\, \pi_\theta(y)\, \mu(\mathrm{d}y)$$

with

$$\pi_\theta(y) \stackrel{\text{def}}{=} \frac{1}{Z_\theta} \exp\left(\left\langle \theta, \bar{B}(y) \right\rangle \right).$$

The gradient of the (log)-likelihood is untractable

The gradient of the log-likelihood

$$\nabla_\theta \left\{ \log L(\theta) \right\} = \int H_\theta(x)\ \pi_\theta(\mathrm{d}x)$$

is an untractable expectation w.r.t. the Gibbs distribution (known up to a constant)
For all (x, θ), $H_\theta(x)$ can be evaluated.

Conclusion, part I

Problem of minimization:

$$\text{argmin}_{\theta \in \Theta} F(\theta) \qquad \text{with } F(\theta) = f(\theta) + g(\theta)$$

when

- $\theta \in \Theta \subseteq \mathbb{R}^d$
- the function g non-smooth nonnegative function (explicit), possibly convex

Conclusion, part I

Problem of minimization:

$$\text{argmin}_{\theta\in\Theta} F(\theta) \qquad \text{with } F(\theta) = f(\theta) + g(\theta)$$

when

- $\theta \in \Theta \subseteq \mathbb{R}^d$
- the function g non-smooth nonnegative function (explicit), possibly convex
- the function f is
  - · not necessarily convex,
  - · C^1 and ∇f is L-Lipschitz

$$\exists L > 0,\ \forall \theta, \theta' \qquad \|\nabla f(\theta) - \nabla f(\theta')\| \leq L\|\theta - \theta'\|.$$

  - · with an untractable gradient of the form

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

Approximation of the gradient

$$\nabla_\theta f(\theta) = -\nabla_\theta \left\{\log L(\theta)\right\} = \int_{\mathsf{X}} H_\theta(x)\ \pi_\theta(\mathrm{d}x)$$

1. Quadrature techniques: poor behavior w.r.t. the dimension of X
2. use i.i.d. samples from π_θ (or an auxiliary distribution) to define a Monte Carlo approximation: not possible or not efficient in general.
3. use m samples from a non stationary Markov chain $\{X_{j,\theta}, j \geq 0\}$ with unique stationary distribution π_θ, and define a Monte Carlo approximation. MCMC samplers provide such a chain.

Approximation of the gradient

$$\nabla_\theta f(\theta) = -\nabla_\theta \left\{\log L(\theta)\right\} = \int_{\mathsf{X}} H_\theta(x)\ \pi_\theta(\mathrm{d}x)$$

1. Quadrature techniques: poor behavior w.r.t. the dimension of X
2. use i.i.d. samples from π_θ (or an auxiliary distribution) to define a Monte Carlo approximation: not possible or not efficient in general.
3. use m samples from a non stationary Markov chain $\{X_{j,\theta}, j \geq 0\}$ with unique stationary distribution π_θ, and define a Monte Carlo approximation. MCMC samplers provide such a chain.

Stochastic approximation of the gradient

A biased approximation, since for MCMC samples $X_{j,\theta}$

$$\mathbb{E}\left[h(X_{j,\theta})\right] \neq \int h(x)\, \pi_\theta(\mathrm{d}x).$$

If the Markov chain is ergodic, the bias vanishes when $j \to \infty$.

Conclusion, part I

Problem of minimization:

$$\text{argmin}_{\theta\in\Theta} F(\theta) \qquad \text{with } F(\theta) = f(\theta) + g(\theta)$$

when

- $\theta \in \Theta \subseteq \mathbb{R}^d$
- the function g non-smooth nonnegative function (explicit), possibly convex
- the function f is
  - not necessarily convex,
  - C^1 and ∇f is L-Lipschitz

$$\exists L > 0,\ \forall \theta, \theta' \qquad \|\nabla f(\theta) - \nabla f(\theta')\| \leq L\|\theta - \theta'\|.$$

  - with an untractable gradient of the form

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

which can be approximated by **biased** Monte Carlo techniques

Outline

The Proximal-Gradient algorithm (1/3)

$$\operatorname{argmin}_{\theta\in\Theta} F(\theta) \qquad \text{with } F(\theta) = \underbrace{f(\theta)}_{\text{smooth}} + \underbrace{g(\theta)}_{\text{non smooth,convex}}$$

The Proximal Gradient algorithm

Given a stepsize sequence $\{\gamma_n, n \geq 0\}$, *iterative algorithm:*

$$\theta_{n+1} = \operatorname{Prox}_{\gamma_{n+1},g}\left(\theta_n - \gamma_{n+1}\nabla f(\theta_n)\right)$$

where

$$\operatorname{Prox}_{\gamma,g}(\tau) \stackrel{\text{def}}{=} \operatorname{argmin}_{\theta\in\Theta}\left(g(\theta) + \frac{1}{2\gamma}\|\theta-\tau\|^2\right)$$

Proximal map: Moreau(1962)
Proximal Gradient algorithm: Beck-Teboulle(2010); Combettes-Pesquet(2011); Parikh-Boyd(2013)

Forward-Backward algorithm: Chen-Rockafeller(1997); Tseng (1998)

The Proximal-Gradient algorithm (2/3)

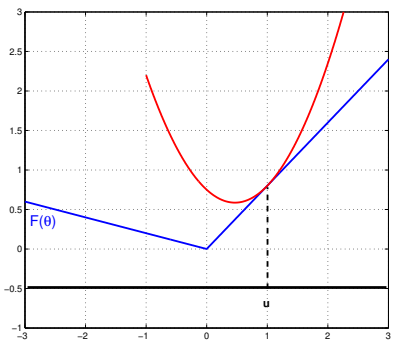
$$\text{argmin}_{\theta\in\Theta} F(\theta) \qquad \text{with } F(\theta) = \underbrace{f(\theta)}_{\text{smooth}} + \underbrace{g(\theta)}_{\text{non smooth,convex}}$$

The algorithm

$$\theta_{n+1} = \text{argmin}_{\theta\in\Theta} \left(g(\theta) + \frac{1}{2\gamma_{n+1}} \|\theta - \theta_n + \gamma_{n+1}\nabla f(\theta_n)\|^2 \right)$$

can be seen as

1. A Majorize-Minimize algorithm from a quadratic majorization of f (since Lipschitz gradient) which produces a sequence $\{\theta_n, n \geq 0\}$ such that $F(\theta_{n+1}) \leq F(\theta_n)$



For all $\gamma < 1/L$,

$$F(\theta) \leq f(\theta_n) + \langle \nabla f(\theta_n), \theta - \theta_n \rangle + \frac{1}{2\gamma}\|\theta - \theta_n\|^2 + g(\theta)$$

The Proximal-Gradient algorithm (2/3)

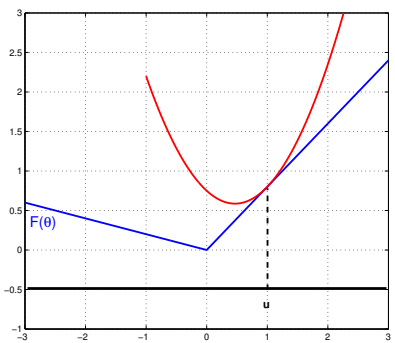
$$\text{argmin}_{\theta\in\Theta} F(\theta) \qquad \text{with } F(\theta) = \underbrace{f(\theta)}_{\text{smooth}} + \underbrace{g(\theta)}_{\text{non smooth,convex}}$$

The algorithm

$$\theta_{n+1} = \text{argmin}_{\theta\in\Theta}\left(g(\theta) + \frac{1}{2\gamma_{n+1}} \|\theta - \theta_n + \gamma_{n+1}\nabla f(\theta_n)\|^2 \right)$$

can be seen as

1. A Majorize-Minimize algorithm from a quadratic majorization of f (since Lipschitz gradient) which produces a sequence $\{\theta_n, n \geq 0\}$ such that $F(\theta_{n+1}) \leq F(\theta_n)$



For all $\gamma < 1/L$,

$$F(\theta) \leq f(\theta_n) + \langle \nabla f(\theta_n), \theta - \theta_n \rangle + \frac{1}{2\gamma}\|\theta - \theta_n\|^2 + g(\theta)$$

2. A generalization of the gradient algorithm to a composite objective function.
3. An Explicit-Implicit gradient algorithm

$$\theta_{n+1/2} = \theta_n - \gamma_{n+1}\nabla f(\theta_n) \qquad \theta_{n+1} \text{ s.t. } \theta_{n+1} = \theta_{n+1/2} - \gamma_{n+1}\,\partial g(\theta_{n+1})$$

The proximal-gradient algorithm (3/3)

$$\text{argmin}_{\theta\in\Theta} F(\theta) \qquad \text{with } F(\theta) = \underbrace{f(\theta)}_{\text{smooth}} + \underbrace{g(\theta)}_{\text{non smooth}}$$

About the Prox-step in the algorithm

$$\theta_{n+1} = \text{argmin}_{\theta\in\Theta} \left(g(\theta) + \frac{1}{2\gamma_{n+1}} \|\theta - \theta_n + \gamma_{n+1}\nabla f(\theta_n)\|^2 \right)$$

- when $g = 0$: $\text{Prox}(\tau) = \tau$
- when g is the $\{0, +\infty\}$-valued indicator fct of a closed convex set: the algorithm is the projected gradient.
- in some cases, Prox is explicit (e.g. elastic net penalty). Otherwise, numerical approximation:

$$\theta_{n+1} = \text{Prox}_{\gamma_{n+1},g}\left(\theta_n - \gamma_{n+1}\nabla f(\theta_n)\right) + \epsilon_{n+1} \qquad \text{in this talk, } \epsilon_{n+1} = 0$$

The perturbed proximal-gradient algorithm

The Perturbed Proximal Gradient algorithm

Given a stepsize sequence $\{\gamma_n, n \geq 0\}$, *iterative algorithm:*

$$\theta_{n+1} = \mathrm{Prox}_{\gamma_{n+1},g}\left(\theta_n - \gamma_{n+1}\mathbf{H_{n+1}}\right)$$

where H_{n+1} *is an approximation of* $\nabla f(\theta_n)$.

Monte Carlo-Proximal Gradient algorithm

In the case:

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

The MC-Proximal Gradient algorithm

Choose a stepsize sequence $\{\gamma_n, n \geq 0\}$ *and a batch size sequence* $\{m_n, n \geq 0\}$.

Given the current value θ_n,

1. *Sample a Markov chain* $\{X_{j,n}, j \geq 0\}$ *from a MCMC sampler with kernel* $P_{\theta_n}(x, \mathrm{d}x')$, *and unique invariant distribution* $\mathrm{d}\pi_{\theta_n}$.
2. *Set*

$$H_{n+1} = \frac{1}{m_{n+1}} \sum_{j=1}^{m_{n+1}} H_{\theta_n}(X_{j,n}).$$

3. *Update the value of the parameter*

$$\theta_{n+1} = \mathrm{Prox}_{\gamma_{n+1},g}\left(\theta_n - \gamma_{n+1} H_{n+1}\right)$$

MCPG or SAPG

If in addition,

$$H_\theta(x) = \Phi(\theta) + \langle \Psi(\theta), S(x) \rangle$$

which implies

$$\nabla f(\theta) = \Phi(\theta) + \left\langle \Psi(\theta), \int S(x)\, \pi_\theta(x) \mu(\mathrm{d}x) \right\rangle,$$

Then, for H_{n+1}, two strategies analogy with MCEM / SAEM

1. Monte Carlo - Proximal Gradient Algorithms

$$\nabla f(\theta_n) \approx H_{n+1} \stackrel{\text{def}}{=} \Phi(\theta_n) + \left\langle \Psi(\theta_n), \frac{1}{m_{n+1}} \sum_{j=1}^{m_{n+1}} S(X_{j,n}) \right\rangle$$

2. Stochastic Approximation - Proximal Gradient Algorithms

$$\nabla f(\theta_n) \approx H_{n+1} \stackrel{\text{def}}{=} \Phi(\theta_n) + \langle \Psi(\theta_n), S_{n+1} \rangle$$

where

$$S_{n+1} = (1 - \delta_{n+1})\, S_n + \delta_{n+1}\, \frac{1}{m_{n+1}} \sum_{j=1}^{m_{n+1}} S(X_{j,n}).$$

(*) Penalized Expectation-Maximization (EM) vs Proximal-Gradient

- EM Dempster et al. (1977) is a Majorize-Minimize algorithm for the computation of the ML estimate in latent variable models.
- (Stochastic) EM algorithms

$$\begin{aligned}\tau_{n+1} &= \text{argmax}_\theta \int \log p_\theta(x)\ \pi_{\theta_n}(\mathrm{d}x) \\ &= \text{argmax}_\theta \left\{\Phi(\theta) + \langle \Psi(\theta), S_{n+1}\rangle\right\}\end{aligned}$$

with

$$S_{n+1} = \int S(x)\ \pi_{\tau_n}(\mathrm{d}x) \qquad \text{EM}$$

$$S_{n+1} = \frac{1}{m_{n+1}} \sum_{j=1}^{m_{n+1}} S(X_{j,n}) \qquad \text{Monte Carlo EM} \quad \text{Wei and Tanner (1990)}$$

$$S_{n+1} = (1-\delta_{n+1})S_n + \frac{\delta_{n+1}}{m_{n+1}} \sum_{j=1}^{m_{n+1}} S(X_{j,n}) \qquad \text{Stoch. Approx. EM} \quad \text{Delyon et al. (1999)}$$

(*) Penalized Expectation-Maximization (EM) vs Proximal-Gradient

- EM Dempster et al. (1977) is a Majorize-Minimize algorithm for the computation of the ML estimate in latent variable models.
- Generalized (Stochastic) EM algorithms

$$\begin{aligned}\tau_{n+1} &= \text{argmax}_\theta \int \log p_\theta(x)\ \pi_{\theta_n}(\mathrm{d}x) \\ &= \text{argmax}_\theta \left\{\Phi(\theta) + \langle \Psi(\theta), S_{n+1}\rangle\right\}\end{aligned}$$

$$\tau_{n+1} \text{ s.t. } \Phi(\tau_{n+1}) + \langle \Psi(\tau_{n+1}), S_{n+1}\rangle \geq \Phi(\tau_n) + \langle \Psi(\tau_n), S_{n+1}\rangle$$

with

$$S_{n+1} = \int S(x)\ \pi_{\tau_n}(\mathrm{d}x) \qquad \text{EM}$$

$$S_{n+1} = \frac{1}{m_{n+1}} \sum_{j=1}^{m_{n+1}} S(X_{j,n}) \qquad \text{Monte Carlo EM} \quad \text{Wei and Tanner (1990)}$$

$$S_{n+1} = (1-\delta_{n+1}) S_n + \frac{\delta_{n+1}}{m_{n+1}} \sum_{j=1}^{m_{n+1}} S(X_{j,n}) \qquad \text{Stoch. Approx. EM} \quad \text{Delyon et al. (1999)}$$

(*) Penalized Expectation-Maximization (EM) vs Proximal-Gradient

- EM Dempster et al. (1977) is a Majorize-Minimize algorithm for the computation of the ML estimate in latent variable models.
- Generalized Penalized (Stochastic) EM algorithms

$$\begin{aligned}\tau_{n+1} &= \text{argmax}_\theta \int \log p_\theta(x)\ \pi_{\theta_n}(\mathrm{d}x) - g(\theta) \\ &= \text{argmax}_\theta \left\{\Phi(\theta) + \langle \Psi(\theta), S_{n+1}\rangle\right\} - g(\theta)\end{aligned}$$

$$\tau_{n+1} \text{ s.t. } \Phi(\tau_{n+1}) + \langle \Psi(\tau_{n+1}), S_{n+1}\rangle - g(\tau_{n+1}) \geq \Phi(\tau_n) + \langle \Psi(\tau_n), S_{n+1}\rangle - g(\tau_n)$$

with

$$S_{n+1} = \int S(x)\ \pi_{\tau_n}(\mathrm{d}x) \qquad \text{EM}$$

$$S_{n+1} = \frac{1}{m_{n+1}} \sum_{j=1}^{m_{n+1}} S(X_{j,n}) \qquad \text{Monte Carlo EM} \quad \text{Wei and Tanner (1990)}$$

$$S_{n+1} = (1-\delta_{n+1}) S_n + \frac{\delta_{n+1}}{m_{n+1}} \sum_{j=1}^{m_{n+1}} S(X_{j,n}) \qquad \text{Stoch. Approx. EM} \quad \text{Delyon et al. (1999)}$$

(*) Penalized Expectation-Maximization (EM) vs Proximal-Gradient

For the computation of

$$\text{argmax}_\theta \left(\ell(\theta) - g(\theta)\right), \qquad \ell(\theta) \stackrel{\text{def}}{=} \int \exp\left(\Phi(\theta) + \langle \Psi(\theta), S(x)\rangle\right) \mu(\mathrm{d}x)$$

- Generalized Penalized Stochastic EM define a sequence $(\tau_n)_n$ s.t.

  $$\Phi(\tau_{n+1}) + \langle \Psi(\tau_{n+1}), S_{n+1}\rangle - g(\tau_{n+1}) \geq \Phi(\tau_n) + \langle \Psi(\tau_n), S_{n+1}\rangle - g(\tau_n)$$

  for different definitions of S_{n+1}
- Monte Carlo - Proximal Gradient and Stochastic Approximation -Proximal Gradient define a sequence $(\theta_n)_n$ s.t.

  $$\Phi(\theta_{n+1}) + \langle \Psi(\theta_{n+1}), S_{n+1}\rangle - g(\theta_{n+1}) \geq \Phi(\theta_n) + \langle \Psi(\theta_n), S_{n+1}\rangle - g(\theta_n)$$

  for different definitions S_{n+1}.

In all cases, S_{n+1} is a Monte Carlo-based approximation of

$$\int S(x) \, \frac{\exp(\Phi(\theta_n) + \langle S(x), \Psi(\theta_n)\rangle)}{Z_{\theta_n}} \mu(\mathrm{d}x)$$

Outline

The assumptions

$$\text{argmin}_{\theta\in\Theta} F(\theta) \qquad \text{with } F(\theta) = f(\theta) + g(\theta)$$

where

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

- the function f: $\mathbb{R}^d \to \mathbb{R}$ is a smooth **convex** function

  i.e. f is continuously differentiable and there exists $L > 0$ such that

$$\|\nabla f(\theta) - \nabla f(\theta')\| \leq L \, \|\theta - \theta'\| \qquad \forall \theta, \theta' \in \mathbb{R}^d$$

- $\Theta \subseteq \mathbb{R}^d$ is the domain of g: $\Theta = \{\theta \in \mathbb{R}^d : g(\theta) < \infty\}$.
- The set $\text{argmin}_\Theta F$ is a non-empty subset of Θ.

Convergence: Existing results in the literature

There exist results under (some of) the assumptions

i.i.d. Monte Carlo approx, $\inf_n \gamma_n > 0$, $\sum_n \|H_{n+1} - \nabla f(\theta_n)\| < \infty$,

i.e. results for

- unbiased sampling. Almost no conditions for the biased sampling, such as the MCMC one.
- non vanishing stepsize sequence $\{\gamma_n, n \geq 0\}$.
- increasing batch size: when H_{n+1} is a Monte Carlo sum i.e.

$$H_{n+1} = \frac{1}{m_{n+1}} \sum_{j=1}^{m_{n+1}} H_{\theta_n}(X_{j,n}),$$

the assumptions imply that $\lim_n m_n = +\infty$ at some rate.

Combettes (2001) Elsevier Science.
Combettes-Wajs (2005) Multiscale Modeling and Simulation.
Combettes-Pesquet (2015, 2016) SIAM J. Optim, arXiv
Lin-Rosasco-Villa-Zhou (2015) arXiv
Rosasco-Villa-Vu (2014,2015) arXiv
Schmidt-Leroux-Bach (2011) NIPS

Convergence of the perturbed proximal gradient algorithm (1/3)

$$\theta_{n+1} = \mathrm{Prox}_{\gamma_{n+1},g}\left(\theta_n - \gamma_{n+1}\ H_{n+1}\right) \qquad \text{with } H_{n+1} \approx \nabla f(\theta_n)$$

Set: $\mathcal{L} = \mathrm{argmin}_\Theta(f+g)$ $\qquad$ $\eta_{n+1} = H_{n+1} - \nabla f(\theta_n)$

Theorem (Atchadé, F., Moulines (2015))

Assume

- *g convex, lower semi-continuous; f convex, C^1 and its gradient is Lipschitz with constant L; $\mathcal{L}$ is non empty.*
- *$\sum_n \gamma_n = +\infty$ and $\gamma_n \in (0, 1/L]$.*
- *Convergence of the series*

$$\sum_n \gamma_{n+1}^2 \|\eta_{n+1}\|^2, \qquad \sum_n \gamma_{n+1}\eta_{n+1}, \qquad \sum_n \gamma_{n+1}\left\langle \mathsf{T}_n, \eta_{n+1}\right\rangle$$

where $\mathsf{T}_n = \mathrm{Prox}_{\gamma_{n+1},g}(\theta_n - \gamma_{n+1}\nabla f(\theta_n))$.

Then there exists $\theta_\star \in \mathcal{L}$ such that $\lim_n \theta_n = \theta_\star$.

Convergence of the perturbed proximal gradient algorithm (2/3)

This convergence result

- for the convex case: f and g are convex.

Convergence of the perturbed proximal gradient algorithm (2/3)

This convergence result

- for the convex case: f and g are convex.
- is a deterministic result.
  Covered: deterministic and random approximations H_{n+1} of $\nabla f(\theta_n)$.

Convergence of the perturbed proximal gradient algorithm (2/3)

This convergence result

- for the convex case: f and g are convex.
- is a deterministic result.
  Covered: deterministic and random approximations H_{n+1} of $\nabla f(\theta_n)$.
  Among random approximations:

1. Applications in Computational Statistics

$$H_{n+1} = \Xi\left(X_{1,n}, \cdots, X_{m_{n+1},n}; \theta_n\right)$$

Convergence of the perturbed proximal gradient algorithm (2/3)

This convergence result

- for the convex case: f and g are convex.
- is a deterministic result.
  Covered: deterministic and random approximations H_{n+1} of $\nabla f(\theta_n)$.
  Among random approximations:

1. Applications in Computational Statistics
2. Applications in learning - "finite sum context" :

(objective) $$\operatorname{argmin}_\theta \left(\frac{1}{N} \sum_{i=1}^{N} f_i(\theta) + g(\theta) \right)$$

(Approx. Gdt) $$H_{n+1} = \frac{1}{|I_{n+1}|} \sum_{i \in I_{n+1}} \nabla f_i(\theta_n)$$

(Stochastic) the set I_{n+1}

Proof / Convergence of the perturbed proximal gradient algorithm (3/3)

Its proof relies on

1. a deterministic Lyapunov inequality

$$\|\theta_{n+1}-\theta_\star\|^2 \leq \|\theta_n-\theta_\star\|^2 - \underbrace{2\gamma_{n+1}\left(F(\theta_{n+1})-\min F\right)}_{\text{non-negative}} \underbrace{-2\gamma_{n+1}\left\langle \mathsf{T}_n-\theta_\star, \eta_{n+1}\right\rangle + 2\gamma_{n+1}^2\|\eta_{n+1}\|^2}_{\text{signed noise}}$$

2. (an extension of) the Robbins-Siegmund lemma

Let $\{v_n, n \geq 0\}$ and $\{\chi_n, n \geq 0\}$ be non-negative sequences and $\{\xi_n, n \geq 0\}$ be such that $\sum_n \xi_n$ exists. If for any $n \geq 0$,

$$v_{n+1} \leq v_n - \chi_{n+1} + \xi_{n+1}$$

then $\sum_n \chi_n < \infty$ and $\lim_n v_n$ exists.

Proof / Convergence of the perturbed proximal gradient algorithm (3/3)

Its proof relies on

1. a deterministic Lyapunov inequality

$$\|\theta_{n+1}-\theta_\star\|^2 \leq \|\theta_n-\theta_\star\|^2 - \underbrace{2\gamma_{n+1}\left(F(\theta_{n+1})-\min F\right)}_{\text{non-negative}} \underbrace{-2\gamma_{n+1}\left\langle \mathsf{T}_n-\theta_\star, \eta_{n+1}\right\rangle + 2\gamma_{n+1}^2\|\eta_{n+1}\|^2}_{\text{signed noise}}$$

2. (an extension of) the Robbins-Siegmund lemma

   Let $\{v_n, n \geq 0\}$ and $\{\chi_n, n \geq 0\}$ be non-negative sequences and $\{\xi_n, n \geq 0\}$ be such that $\sum_n \xi_n$ exists. If for any $n \geq 0$,

   $$v_{n+1} \leq v_n - \chi_{n+1} + \xi_{n+1}$$

   then $\sum_n \chi_n < \infty$ and $\lim_n v_n$ exists.

Note: deterministic lemma, signed noise.

Convergence: when H_{n+1} is a Monte-Carlo approximation (1/3)

In the case

$$\nabla f(\theta_n) \approx H_{n+1} = \frac{1}{m_{n+1}} \sum_{j=1}^{m_{n+1}} H_{\theta_n}(X_{j,n}),$$

$$X_{j+1,n} | \text{past} \sim P_{\theta_n}(X_{j,n}, \cdot) \qquad \pi_\theta P_\theta = \pi_\theta;$$

Convergence: when H_{n+1} is a Monte-Carlo approximation (1/3)

In the case

$$\nabla f(\theta_n) \approx H_{n+1} = \frac{1}{m_{n+1}} \sum_{j=1}^{m_{n+1}} H_{\theta_n}(X_{j,n}),$$

$$X_{j+1,n} | \text{past} \sim P_{\theta_n}(X_{j,n}, \cdot) \qquad \pi_\theta P_\theta = \pi_\theta;$$

let us check the condition "$\sum_n \gamma_n \eta_n < \infty$ w.p.1":

$$\begin{aligned}\sum_n \gamma_{n+1}\eta_{n+1} &= \sum_n \gamma_{n+1} \left(H_{n+1} - \nabla f(\theta_n)\right) \\ &= \sum_n \gamma_{n+1} \left\{H_{n+1} - \mathbb{E}\left[H_{n+1}|\mathcal{F}_n\right]\right\} + \sum_n \gamma_{n+1} \underbrace{\left\{\mathbb{E}\left[H_{n+1}|\mathcal{F}_n\right] - \nabla f(\theta_n)\right\}}_{\substack{\text{unbiased MC: null} \\ \text{biased MC: } O_{Lp}(1/m_n)}}\end{aligned}$$

Convergence: when H_{n+1} is a Monte-Carlo approximation (1/3)

In the case

$$\nabla f(\theta_n) \approx H_{n+1} = \frac{1}{m_{n+1}} \sum_{j=1}^{m_{n+1}} H_{\theta_n}(X_{j,n}),$$

$$X_{j+1,n} | \text{past} \sim P_{\theta_n}(X_{j,n}, \cdot) \qquad \pi_\theta P_\theta = \pi_\theta;$$

let us check the condition "$\sum_n \gamma_n \eta_n < \infty$ w.p.1":

$$\sum_n \gamma_{n+1}\eta_{n+1} = \sum_n \gamma_{n+1} \left(H_{n+1} - \nabla f(\theta_n)\right)$$

$$= \sum_n \gamma_{n+1} \left\{H_{n+1} - \mathbb{E}\left[H_{n+1}|\mathcal{F}_n\right]\right\} + \sum_n \gamma_{n+1} \underbrace{\left\{\mathbb{E}\left[H_{n+1}|\mathcal{F}_n\right] - \nabla f(\theta_n)\right\}}_{\substack{\text{unbiased MC: null} \\ \text{biased MC: } O_{Lp}(1/m_n)}}$$

The most technical case: the biased case with constant batch size $m_n = m$

Solution $\widehat{H}_\theta$ to the Poisson equation: $H_\theta - \pi_\theta H_\theta = \widehat{H}_\theta - P_\theta \widehat{H}_\theta$

$H_{n+1} - \nabla f(\theta_n) =$ martingale increment + remainder

Regularity in θ of $t \mapsto \widehat{H}_t$.

Convergence: when H_{n+1} is a Monte-Carlo approximation (2/3)

Increasing batch size: $\lim_n m_n = +\infty$

Conditions on the step sizes and batch sizes

$$\sum_n \gamma_n = +\infty, \qquad \sum_n \frac{\gamma_n^2}{m_n} < \infty; \qquad \sum_n \frac{\gamma_n}{m_n} < \infty \text{ (biased case)}$$

Conditions on the Markov kernels: *There exist* $\lambda \in (0,1)$, $b < \infty$, $p \geq 2$ *and a measurable function* $W : \mathsf{X} \to [1, +\infty)$ *such that*

$$\sup_{\theta \in \Theta} |H_\theta|_W < \infty, \qquad \sup_{\theta \in \Theta} P_\theta W^p \leq \lambda W^p + b.$$

In addition, for any $\ell \in (0, p]$, *there exist* $C < \infty$ *and* $\rho \in (0,1)$ *such that for any* $x \in \mathsf{X}$,

$$\sup_{\theta \in \Theta} \|P_\theta^n(x, \cdot) - \pi_\theta\|_{W^\ell} \leq C \rho^n W^\ell(x). \tag{1}$$

Condition on Θ: Θ *is bounded.*

Convergence: when H_{n+1} is a Monte-Carlo approximation (3/3)

Fixed batch size: $m_n = m$

Condition on the step size:

$$\sum_n \gamma_n = +\infty \qquad \sum_n \gamma_n^2 < \infty \qquad \sum_n |\gamma_{n+1} - \gamma_n| < \infty$$

Condition on the Markov chain: *same as in the case "increasing batch size" and there exists a constant C such that for any $\theta, \theta' \in \Theta$*

$$|H_\theta - H_{\theta'}|_W + \sup_x \frac{\|P_\theta(x,\cdot) - P_{\theta'}(x,\cdot)\|_W}{W(x)} \leq C \|\theta - \theta'\|.$$

Condition on the Prox:

$$\sup_{\gamma \in (0,1/L]} \sup_{\theta \in \Theta} \gamma^{-1} \|\mathrm{Prox}_{\gamma,g}(\theta) - \theta\| < \infty.$$

Condition on Θ: Θ is bounded.

Rates of convergence (1/3) : the problem

For non negative weights a_k, find an upper bound of

$$\sum_{k=1}^{n} \frac{a_k}{\sum_{\ell=1}^{n} a_\ell} F(\theta_k) - \min F$$

It provides

- an upper bound for the cumulative regret ($a_k = 1$)
- an upper bound for an averaging strategy when F is convex since

$$F\left(\sum_{k=1}^{n} \frac{a_k}{\sum_{\ell=1}^{n} a_\ell} \theta_k\right) - \min F \leq \sum_{k=1}^{n} \frac{a_k}{\sum_{\ell=1}^{n} a_\ell} F(\theta_k) - \min F.$$

Rates of convergence (2/3): a deterministic control

Theorem (Atchadé, F., Moulines (2016))

For any $\theta_\star \in \operatorname{argmin}_\Theta F$,

$$\sum_{k=1}^{n} \frac{a_k}{A_n} F(\theta_k) - \min F \leq \frac{a_0}{2\gamma_0 A_n} \|\theta_0 - \theta_\star\|^2 + \frac{1}{2A_n} \sum_{k=1}^{n} \left(\frac{a_k}{\gamma_k} - \frac{a_{k-1}}{\gamma_{k-1}} \right) \|\theta_{k-1} - \theta_\star\|^2 + \frac{1}{A_n} \sum_{k=1}^{n} a_k \gamma_k \|\eta_k\|^2 - \frac{1}{A_n} \sum_{k=1}^{n} a_k \langle \mathsf{T}_{k-1} - \theta_\star, \eta_k \rangle$$

where

$$A_n = \sum_{\ell=1}^{n} a_\ell, \qquad \eta_k = H_k - \nabla f(\theta_{k-1}), \qquad \mathsf{T}_k = \operatorname{Prox}_{\gamma_k, g}(\theta_{k-1} - \gamma_k \nabla f(\theta_{k-1})).$$

Rates (3/3): when H_{n+1} is a Monte Carlo approximation, bound in L^q

$$\left\| F\left(\frac{1}{n}\sum_{k=1}^{n}\theta_k\right) - \min F \right\|_{L^q} \leq \left\| \frac{1}{n}\sum_{k=1}^{n} F(\theta_k) - \min F \right\|_{L^q} \leq u_n$$

$u_n = O(1/\sqrt{n})$

with fixed size of the batch and (slowly) decaying stepsize

$$\gamma_n = \frac{\gamma_\star}{n^a}, a \in [1/2, 1] \qquad\qquad m_n = m_\star.$$

With averaging: optimal rate, even with slowly decaying stepsize $\gamma_n \sim 1/\sqrt{n}$.

$u_n = O(\ln n/n)$

with increasing batch size and constant stepsize

$$\gamma_n = \gamma_\star \qquad\qquad m_n \propto n.$$

Rate after $O(n^2)$ *Monte Carlo samples !*

Nesterov Acceleration (1/3) - solving complex programming problem with convergence rate $O(1/n^2)$ (1983)

- First order

$$\dot{X}_t + \nabla\phi(X_t) = 0$$

Time-discretization:

$$\frac{x_{n+1} - x_n}{\gamma_{n+1}} + \nabla\phi(x_n) = 0.$$

- Second order Inertial Gradient-Like systems

$$\ddot{X}_t + \frac{\alpha}{t}\dot{X}_t + \nabla\phi(X_t) = 0$$

$$\phi(X_t) - \min\phi = O(1/t^2) \qquad \text{Attouch et al. (2015)}$$

seen as the continuous time version of FISTA, satisfying

$$\phi(x_n) - \min\phi = O(1/n^2) \qquad \text{Beck and Teboulle (2009)}$$

Acceleration (2/3)

Let $\{t_n, n \geq 0\}$ be a positive sequence s.t.

$$\gamma_{n+1} t_n (t_n - 1) \leq \gamma_n t_{n-1}^2$$

Nesterov acceleration of the Proximal Gradient algorithm

$$\theta_{n+1} = \mathrm{Prox}_{\gamma_{n+1}, g} \left(\tau_n - \gamma_{n+1} \nabla f(\tau_n) \right)$$
$$\tau_{n+1} = \theta_{n+1} + \frac{t_n - 1}{t_{n+1}} \left(\theta_{n+1} - \theta_n \right)$$

Nesterov(2004), Tseng(2008), Beck-Teboulle(2009)

Zhu-Orecchia (2015); Attouch-Peypouquet(2015); Bubeck-Lee-Singh(2015); Su-Boyd-Candes(2015)

(deterministic) Proximal-gradient $\quad F(\theta_n) - \min F = O\left(\frac{1}{n}\right)$

(deterministic) Accelerated Proximal-gradient $\quad F(\theta_n) - \min F = O\left(\frac{1}{n^2}\right)$

Acceleration (3/3) Aujol-Dossal-F.-Moulines, work in progress

Perturbed Nesterov acceleration: some convergence results

Choose γ_n, m_n, t_n *s.t.*

$$\gamma_n \in (0, 1/L]\,, \qquad \lim_n \gamma_n t_n^2 = +\infty, \qquad \sum_n \gamma_n t_n (1 + \gamma_n t_n) \frac{1}{m_n} < \infty$$

Then there exists $\theta_\star \in \operatorname{argmin}_\Theta F$ *s.t* $\lim_n \theta_n = \theta_\star$.
In addition

$$F(\theta_{n+1}) - \min F = O\left(\frac{1}{\gamma_{n+1} t_n^2}\right)$$

Schmidt-Le Roux-Bach (2011); Dossal-Chambolle(2014); Aujol-Dossal(2015)

γ_n	m_n	t_n	rate	NbrMC
γ	n^3	n	n^{-2}	n^4
$\gamma/\sqrt{n}$	n^2	n	$n^{-3/2}$	n^3

Table: Control of $F(\theta_n) - \min F$

Outline

Inference in graph

Hereafter, we compare MCPG and SAPG and more precisely

- for MCPG, the role of

$$\gamma_n \sim \frac{\gamma_\star}{n^a} \qquad m_n,$$

  and for SAPG, the role of

$$\gamma_n \sim \frac{\gamma_\star}{n^a} \qquad \delta_n \sim \frac{\delta_\star}{n^b} \qquad m_n.$$

- Boxplots are obtained from 50 independent runs.
- Each run of each algorithm produces a sequence $\{\theta_n, n \geq 0\}$. We analyze
  - the convergence of the L_1-norm, to illustrate the convergence of the sequence
  - the convergence of the size of the support: $\#(k : |\theta_{n,k}| > 0)$
  - for each component of the vector θ_n, the frequency of being in the support (frequency over the 50 runs)

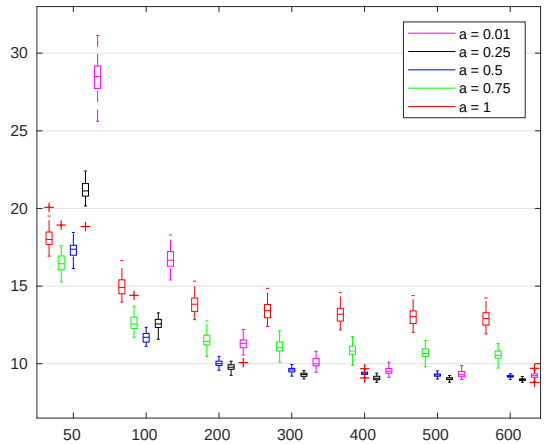
SAPG Here $m_n = 500$.

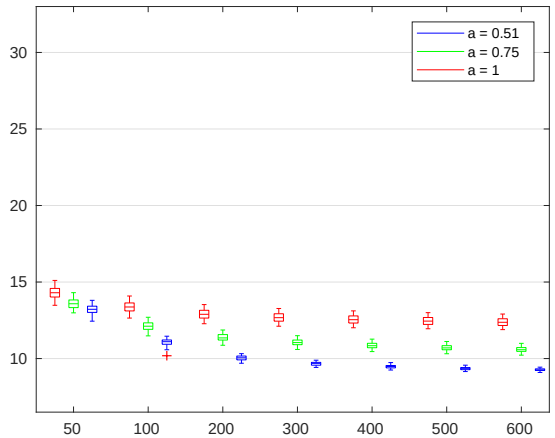
MCPG Different cases for the batch size

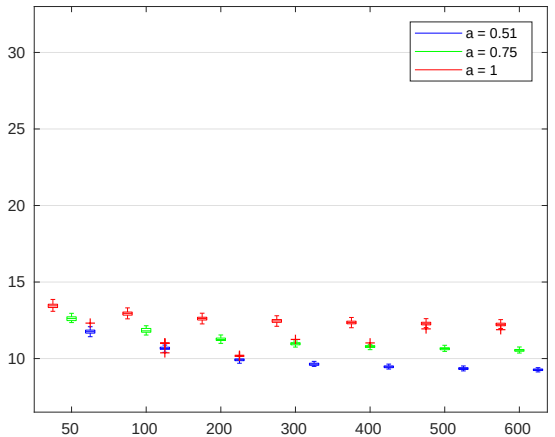
- constant : $m_n = 3\ 000$
- square-root growth
$$m_n = 150\sqrt{n}$$
- linear growth
$$m_n = \max(200, 10n).$$

Here the constants are chosen so that after 600 iterations, all the algorithms used the same total number of Monte Carlo draws (here $1.8\,e6$).

MCPG: boxplot of $n \mapsto \|\theta_n\|_1$ along 600 itérations





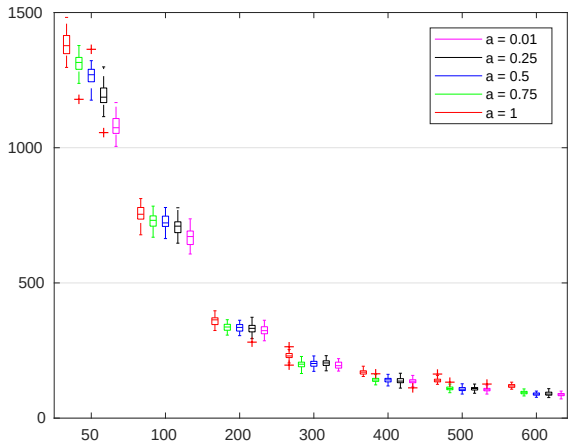


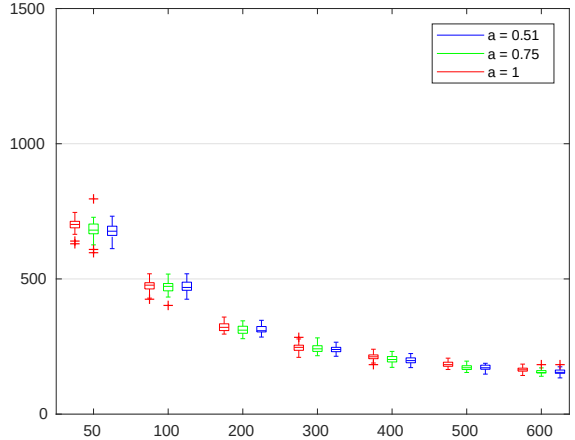
Different batch size :
$m_n = O(n)$ (left), $m_n = O(\sqrt{n})$ (center), $m_n = m$ (right)

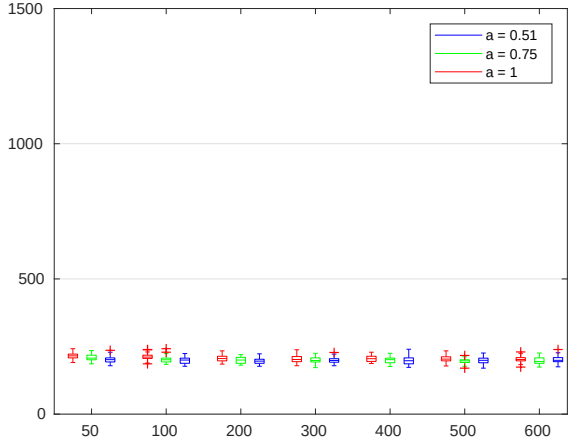
Different rates for the stepsize : $\gamma_n = O(n^{-a})$

$\hookrightarrow$ it is better to have a slow decaying rate of the stepsize $1/\sqrt{n}$ to be efficient both in the transient phase and on the convergence phase.

MCPG :boxplot of the length of the support θ_n, along 600 itérations





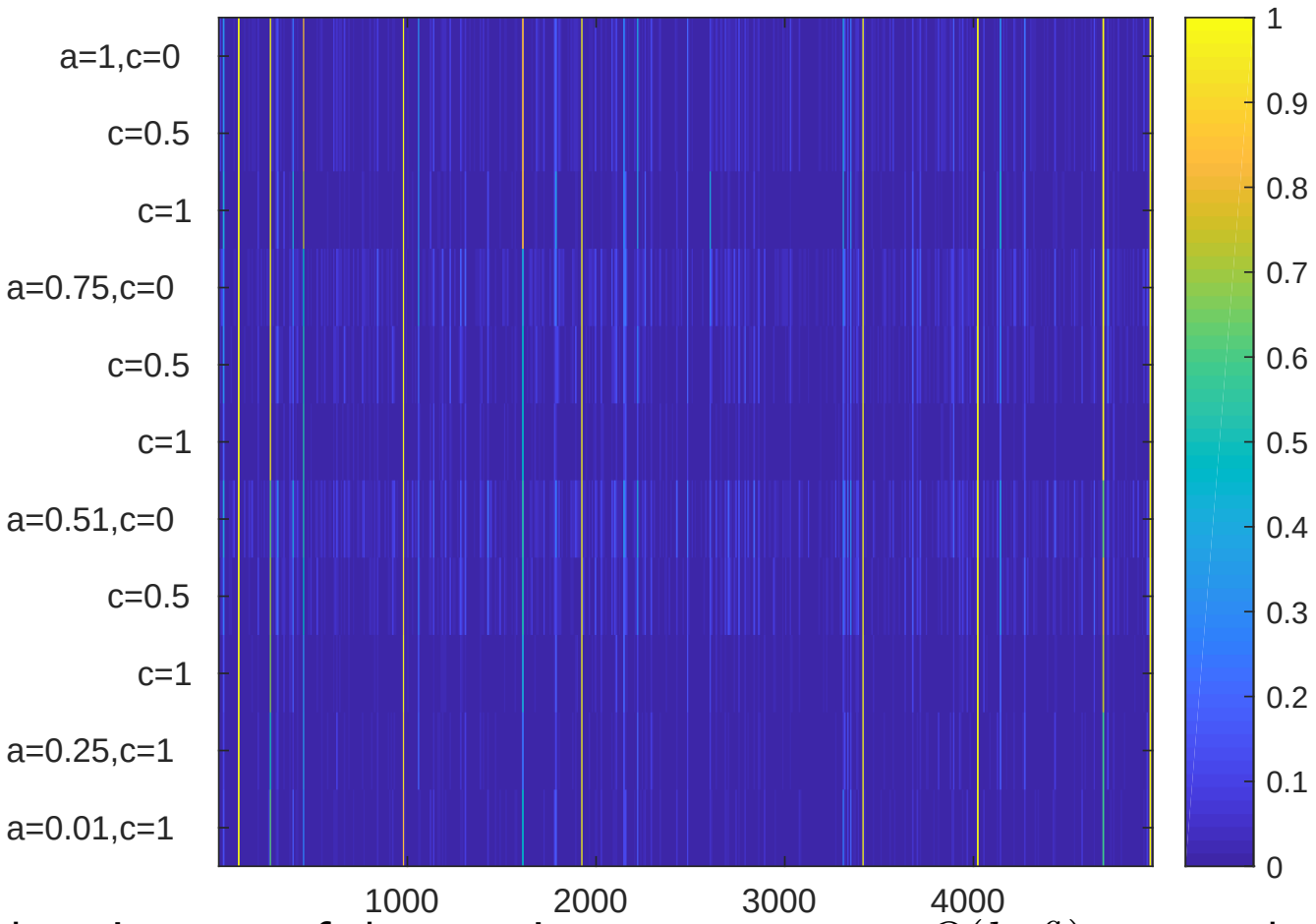


Different batch sizes :
$m_n = O(n)$ (left), $m_n = O(\sqrt{n})$ (center), $m_n = m$ (right)

Different stepsizes : $\gamma_n = O(n^{-a})$

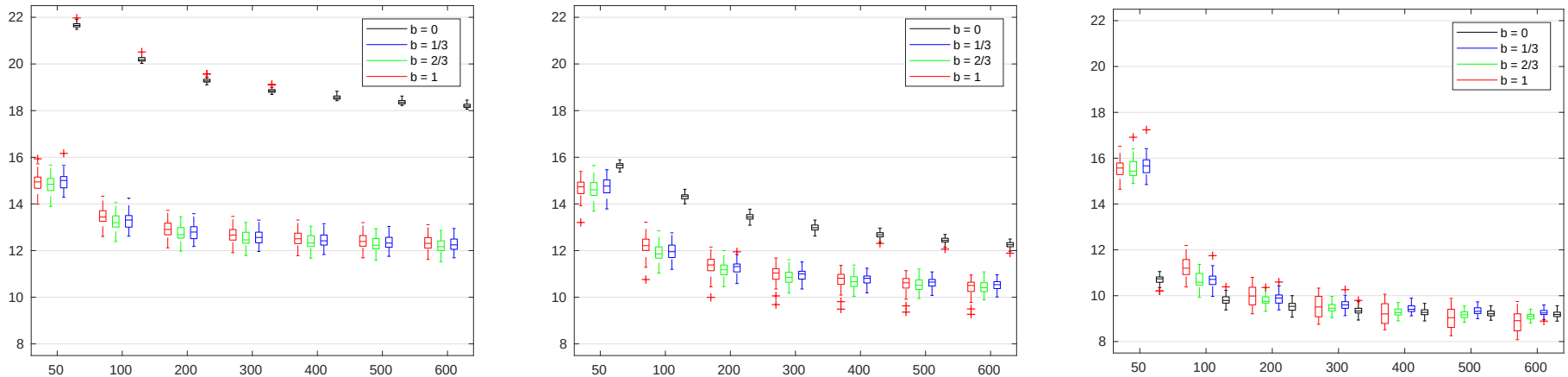
$\hookrightarrow$ it is better to choose a slow decaying rate

At convergence, support / MCPG. θ_n contains about. 4800 components



For different decaying rate of the stepsize sequence $\gamma_k = O(k^{-a})$ and increasing rate of the batch size $m_k = O(k^c)$

SAPG: boxplot of $n \mapsto \|\theta_n\|_1$ along 600 iterations

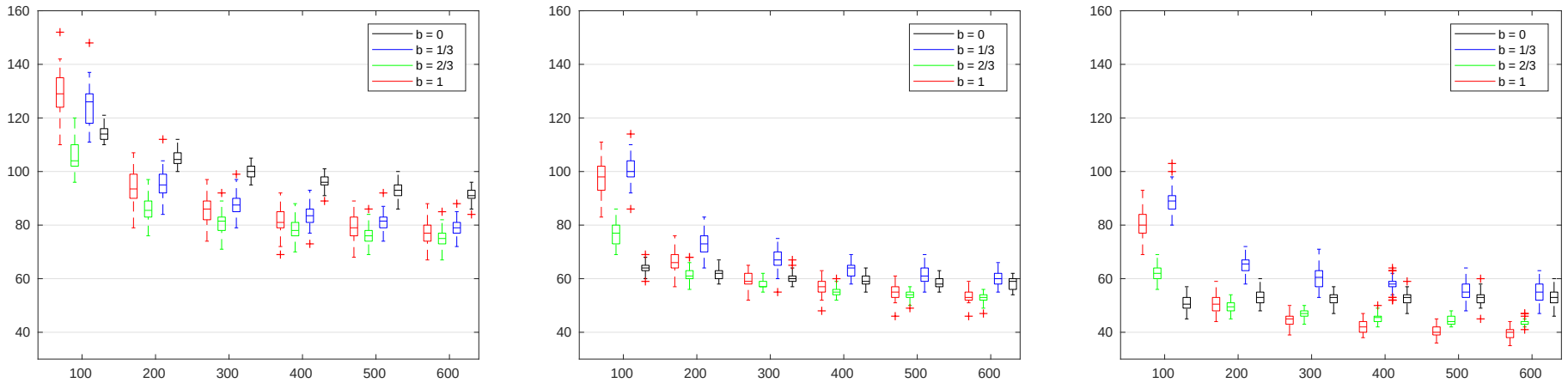


Different decaying rates for γ_n :
$O(1/n)$ (left) $O(n^{-3/4})$ (center) $O(n^{-0.5})$ (right)

Different decaying rate for : $\delta_n = O(n^{-b})$

$\hookrightarrow$ it is better to have a slow decaying rate of γ_n; and in that case, a constant δ_n.

SAPG : boxplot of the size of the support θ_n, along 600 iterations

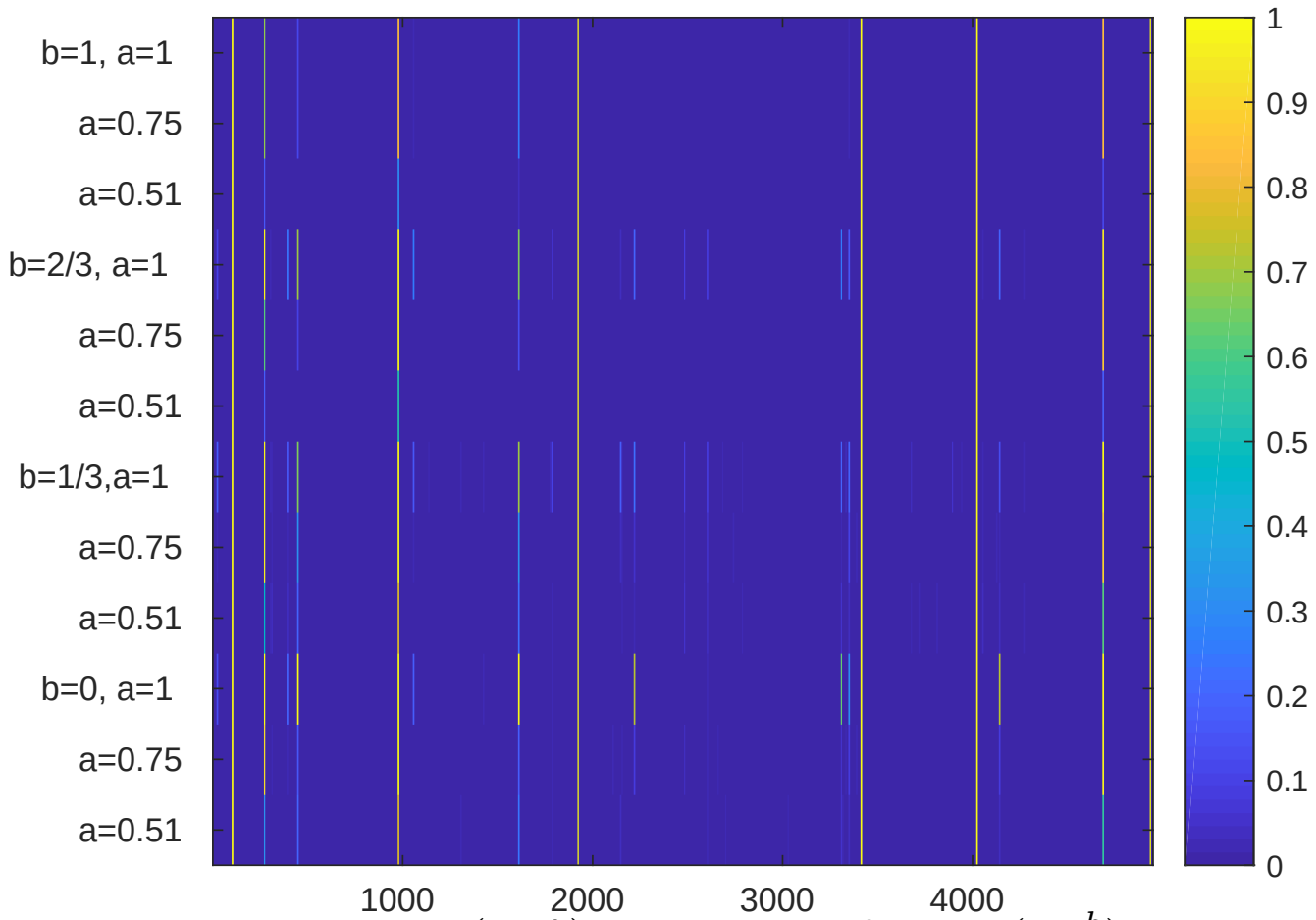


Different decaying rates of γ_n
$O(1/n)$ (left) $O(n^{-3/4})$ (center) $O(n^{-0.5})$ (right)

Different decaying rates of : $\delta_n = O(n^{-b})$

$\hookrightarrow$ a slow decaying rate of γ_n is better; and in that case, a constant δ_n for the transient phase and a rapid decay for the convergent phase.

Support at convergence, SAPG



For different decaying rates $\gamma_k = O(k^{-a})$ and of $\delta_k = O(k^{-b})$
probability to be in the support at iteration $n = 600$, computed over 50 indep

MCPG ou SAPG

- on ne voit pas grande différence dans l'évolution de $\|\theta_n\|$.
- En revanche, SAPG donne des vecteurs beaucoup plus creux : voir la taille du support à convergence, et la proba des composantes actives à l'itération finale. Justement parce qu'il permet d'avoir une meilleure appoximation du gradient en utilisant tous les tirages produite depuis le début. On voit bien que MCPG souffre tant que le nombre de tirages MC est petit (voir la taille du support, dans le cas $m_n = O(n)$).