Inférence pénalisée dans les modèles à vraisemblance non explicite par des algorithmes gradient-proximaux perturbés

Gersende Fort

Institut de Mathématiques de Toulouse,
CNRS and Univ. Paul Sabatier
Toulouse, France

Based on joint works with

- Yves Atchadé (Univ. Michigan, USA)
- Eric Moulines (Ecole Polytechnique, France)

↪ `On Perturbed Proximal-Gradient algorithms (JMLR, 2016)`

- Edouard Ollier (ENS Lyon, France)
- Adeline Samson (Univ. Grenoble Alpes, France).

↪ `Penalized inference in Mixed Models by Proximal Gradients methods (work in progress)`

- Jean-François Aujol (IMB, Bordeaux, France)
- Charles Dossal (IMB, Bordeaux, France).

↪ `Acceleration for perturbed Proximal Gradient algorithms (work in progress)`

Motivation : Pharmacokinetic (1/2)

- N patients.
- For patient i, observations $\{Y_{ij}, 1 \leq j \leq J\}$: evolution of the concentration at times $t_{ij}, 1 \leq j \leq J$.
- Initial dose D.

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

Motivation : Pharmacokinetic (1/2)

- N patients.
- For patient i, observations $\{Y_{ij}, 1 \leq j \leq J\}$: evolution of the concentration at times $t_{ij}, 1 \leq j \leq J$.
- Initial dose D.

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

Statistical analysis:

- estimation of $(\beta, \sigma^2, \Omega)$, 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:

- The distribution of $\{Y_{ij}, X_i; 1 \leq i \leq N, 1 \leq j \leq J\}$ has an explicit expression.
- The distribution of $\{Y_{ij}; 1 \leq i \leq N, 1 \leq j \leq J\}$ does not have an explicit expression; at least, the marginal distribution of the previous one.

Outline

Penalized Maximum Likelihood inference with untractable Likelihood

- N observations : $\mathsf{Y} = (Y_1, \cdots, Y_N)$
- A parametric statistical model $\theta \in \Theta \subseteq \mathbb{R}^d$

$$\theta \mapsto L(\theta) \qquad \text{likelihood of the observations}$$

- A penalty constraint on the parameter θ: $\theta \mapsto g(\theta)$ for sparsity constraints on θ. Usually, g non-smooth and convex.

Goal: Computation of

$$\theta \mapsto \operatorname{argmin}_{\theta \in \Theta} \left(-\frac{1}{N} \log L(\theta) + g(\theta) \right)$$

when the likelihood L has no closed form expression, and can not be evaluated.

Example 1: Latent variable model

- The log-likelihood of the observations Y is of the form

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

  where μ is a positive σ-finite measure on a set X.
- x are the missing/latent data; (x, Y) are the complete data.

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 ?

Gradient of the likelihood in a latent variable model

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

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

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

Gradient of the likelihood in a latent variable model

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

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

$$\nabla_\theta \left\{ -\frac{1}{N} \log L(\theta) \right\} = \int H_\theta(x)\ \pi_\theta(\mathrm{d}x)$$

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

Approximation of the gradient

$$\nabla_\theta \left\{ -\frac{1}{N} \log L(\theta) \right\} = \int_{\mathsf{X}} H_\theta(x)\ \pi_\theta(\mathrm{d}x)$$

1. Quadrature techniques: poor behavior w.r.t. the dimension of X
2. use i.i.d. samples from π_θ to define a Monte Carlo approximation: not possible, 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

$$\nabla_\theta \left\{ -\frac{1}{N} \log L(\theta) \right\} = \int_{\mathsf{X}} H_\theta(x)\ \pi_\theta(\mathrm{d}x)$$

1. Quadrature techniques: poor behavior w.r.t. the dimension of X
2. use i.i.d. samples from π_θ to define a Monte Carlo approximation: not possible, 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

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

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

Example 2: Discrete graphical model (Markov random field)

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 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.

Likelihood and its gradient in Markov random field

- 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.

Likelihood and its gradient in Markov random field

- 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.

- 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.

Approximation of the gradient

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

The Gibbs measure

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

is known up to the constant Z_θ.

Exact sampling from π_θ can be approximated by MCMC samplers (Gibbs-type samplers such as Swendsen-Wang, ...)

A biased approximation of the gradient is available.

To summarize,

Problem:

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

when

- $\theta \in \Theta \subseteq \mathbb{R}^d$
- g convex non-smooth function (explicit).
- f is C^1, 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.
∇f is Lipschitz

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

f is not necessarily convex.

Outline

The Proximal-Gradient algorithm (1/2)

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

The Proximal Gradient 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)

The Proximal-Gradient algorithm (1/2)

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

The Proximal Gradient 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)

- A generalization of the gradient algorithm to a composite objective function.
- An iterative algorithm (from the Majorize-Minimize optim method) which produces a sequence $\{\theta_n, n \geq 0\}$ such that

$$F(\theta_{n+1}) \leq F(\theta_n).$$

The proximal-gradient algorithm (2/2)

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

The Proximal Gradient algorithm

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

where

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

About the Prox-step:

- when $g = 0$: $\operatorname{Prox}(\tau) = \tau$
- when g is the projection on a compact set: the algorithm is the projected gradient.
- in some cases, Prox is explicit (e.g. elastic net penalty).
- Otherwise, numerical approximation:

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

The perturbed proximal-gradient algorithm

The Perturbed Proximal Gradient algorithm

Given a deterministic sequence $\{\gamma_n, n \geq 0\}$,

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

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

Algorithm Monte Carlo-Proximal Gradient for Penalized ML

When the gradient of the log-likelihood is of the form

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

The MC-Proximal Gradient algorithm

Given the current value θ_n,

1. *Sample a Markov chain $\{X_{j,n}, j \geq 0\}$ from a MCMC sampler with target distribution $\pi_{\theta_n}\, \mathrm{d}\mu$*
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 current value of the parameter*
$$\theta_{n+1} = \mathrm{Prox}_{\gamma_{n+1}, g}\left(\theta_n - \gamma_{n+1} H_{n+1}\right)$$

Algorithm Stochastic Approximation-Proximal gradient for Penalized ML

If in addition,

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

which implies

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

The SA-Proximal Gradient algorithm

Given the current value θ_n,

1. *Sample a Markov chain $\{X_{j,n}, j \geq 0\}$ from a MCMC sampler with target distribution $\pi_{\theta_n}\, \mathrm{d}\mu$*
2. *Set $H_{n+1} = \Phi(\theta_n) + \Psi(\theta_n) S_{n+1}$ with*

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

3. *Update the current value of the parameter*

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

(*) Penalized Expectation-Maximization vs Proximal-Gradient

- EM Dempster et al. (1977) is a Majorize-Minimize algorithm for the computation of the ML estimate in a latent variable model.
- The Proximal Gradient algorithm is a Penalized Generalized-EM algorithm:

$$g(\theta_{n+1}) + Q(\theta_{n+1}, \theta_n) \leq g(\theta_n) + Q(\theta_n, \theta_n)$$

where

$$Q(\theta, \theta') \stackrel{\text{def}}{=} \int \log p_\theta(x)\, \pi_{\theta'}(x) \mathsf{d}\mu(x), \qquad \pi_\theta(x) \stackrel{\text{def}}{=} \frac{p_\theta(x)}{\int p_\theta(z) \mathsf{d}\mu(z)}$$

(*) Penalized Expectation-Maximization vs Proximal-Gradient

- EM Dempster et al. (1977) is a Majorize-Minimize algorithm for the computation of the ML estimate in a latent variable model.
- The Proximal Gradient algorithm is a Penalized Generalized-EM algorithm:

$$g(\theta_{n+1}) + Q(\theta_{n+1}, \theta_n) \leq g(\theta_n) + Q(\theta_n, \theta_n)$$

where

$$Q(\theta, \theta') \overset{\text{def}}{=} \int \log p_\theta(x)\, \pi_{\theta'}(x) \mathrm{d}\mu(x), \qquad \pi_\theta(x) \overset{\text{def}}{=} \frac{p_\theta(x)}{\int p_\theta(z) \mathrm{d}\mu(z)}$$

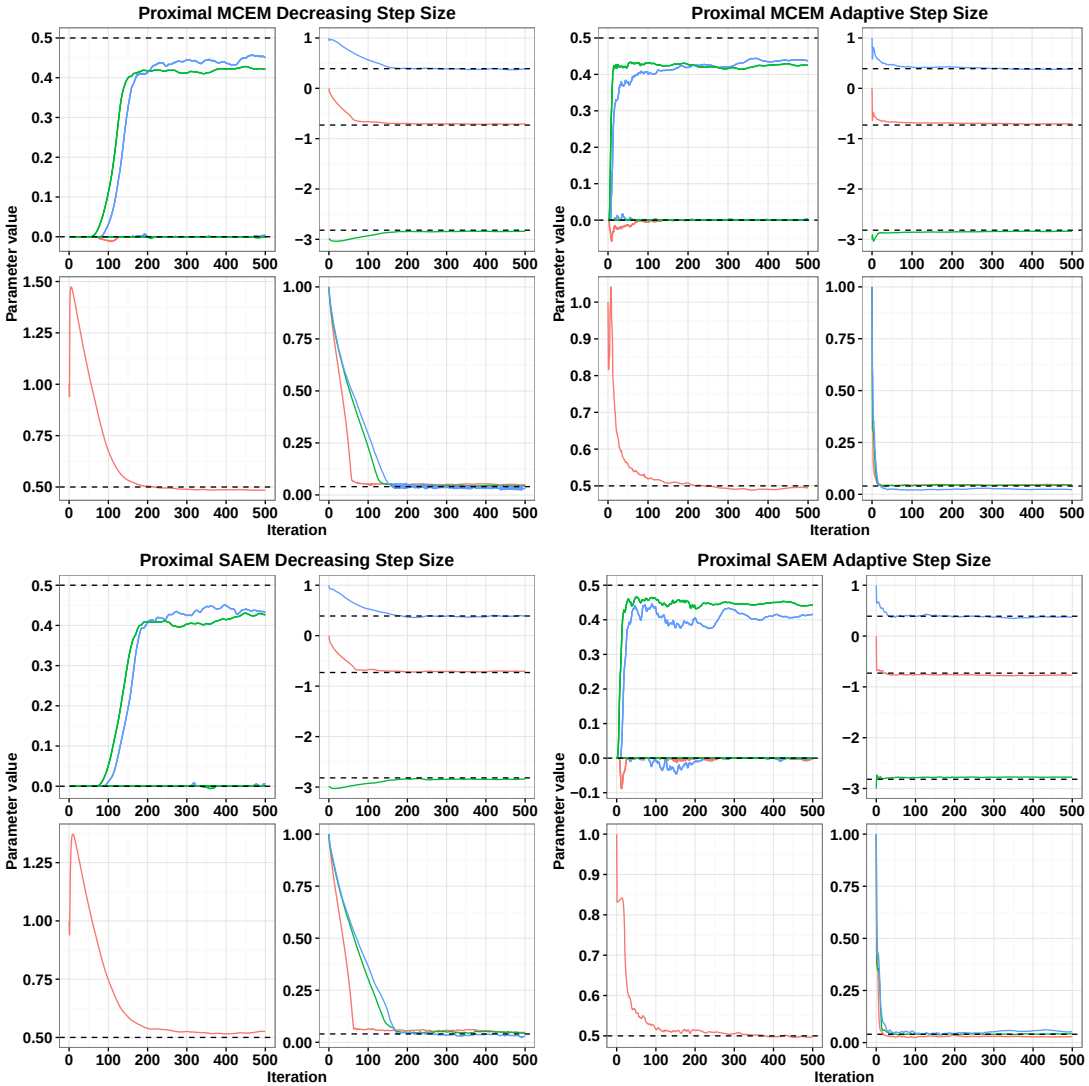
- MC-Proximal Gradient corresponds to the Penalized Generalized-MCEM algorithm Wei and Tanner (1990)
- SA-Proximal Gradient corresponds to the Penalized Generalized-SAEM algorithm Delyon et al. (1999)

Numerical illustration (1/3): pharmacokinetic

For the implementation of the algorithm

- Penalty term: $g(\theta) = \lambda\|\beta\|_1$. How to choose λ ? Weighted norm ?
- Step size sequences: constant or vanishing stepsize sequence $\{\gamma_n, n \geq 0\}$? (and δ_n for the SA-Prox Gdt algorithm)
- Monte Carlo approximation: fixed or increasing batch size ?

Numerical illustration (2/3): pharmacokinetic



Numerical illustration (3/3): pharmacokinetic

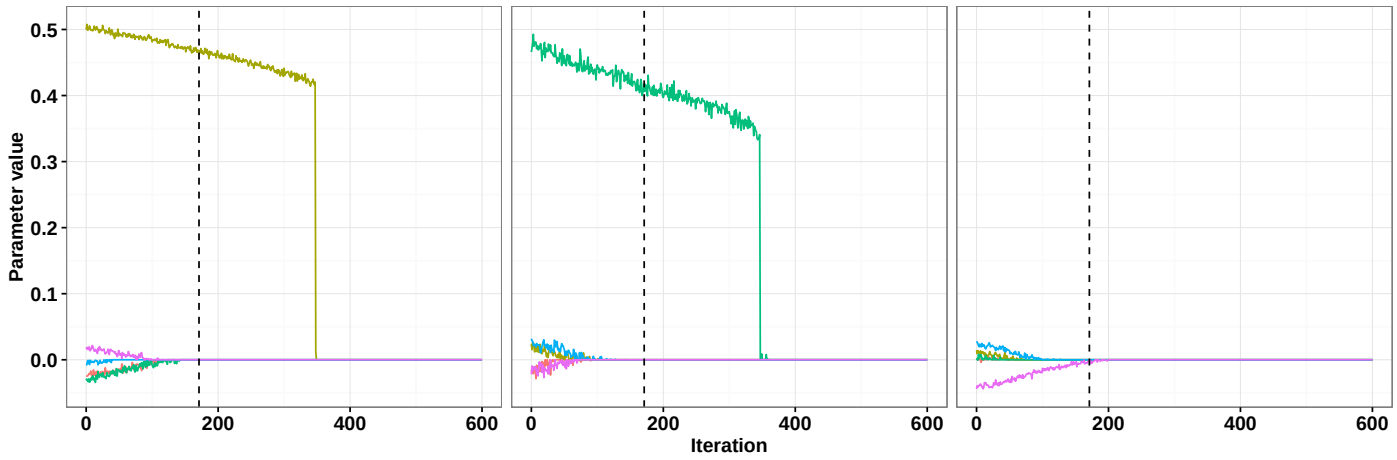


Figure: Low dimension simulation setting. Regularization path of the covariate parameters for the clearance (left), absorption constant (middle) and volume of distribution (right) parameters. Black dashed line corresponds to the λ value selected by $EBIC$. Each colored curve corresponds to a covariate.

Outline

Problem:

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

The Perturbed Proximal Gradient algorithm

Given a $(0, 1/L]$*-valued deterministic sequence* $\{\gamma_n, n \geq 0\}$*, set for* $n \geq 0$*,*

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

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

Which conditions on the sequence $\{\gamma_n, n \geq 0\}$ and on the perturbations $H_{n+1} - \nabla f(\theta_n)$, for this algorithm to converge to the minima of F ?

The assumptions

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

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 $\operatorname{argmin}_\Theta F$ is a non-empty subset of Θ.

Existing results in the literature

There exist results under (some of) the assumptions

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

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.

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

then $\lim_n m_n = +\infty$ 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

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

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

In the case

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

and $\{X_{j,n}, j \geq 0\}$ is a non-stationary Markov chain with unique stationary distribution π_{θ_n}: i.e. for any $n \geq 0$,

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

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

In the case

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

and $\{X_{j,n}, j \geq 0\}$ is a non-stationary Markov chain with unique stationary distribution π_{θ_n}: i.e. for any $n \geq 0$,

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

let us check a condition:

$$\begin{aligned}
\sum_n \gamma_{n+1}\eta_{n+1} &= \sum_n \gamma_{n+1}\left(H_{n+1} - \nabla f(\theta_n)\right) \\
&= \sum_n \gamma_{n+1}\left\{H_{n+1} - \mathbb{E}\left[H_{n+1}|\mathcal{F}_n\right]\right\} + \sum_n \gamma_{n+1} \underbrace{\mathbb{E}\left[H_{n+1}|\mathcal{F}_n\right] - \nabla f(\theta_n)}_{\text{bias, non vanishing when } m_n = m} \\
&= \sum_n \gamma_{n+1}\text{Martingale increment} + \sum_n \gamma_{n+1}\text{Remainder}
\end{aligned}$$

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

In the case

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

and $\{X_{j,n}, j \geq 0\}$ is a non-stationary Markov chain with unique stationary distribution π_{θ_n}: i.e. for any $n \geq 0$,

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

let us check a condition:

$$\begin{aligned}
\sum_n \gamma_{n+1}\eta_{n+1} &= \sum_n \gamma_{n+1}\left(H_{n+1} - \nabla f(\theta_n)\right) \\
&= \sum_n \gamma_{n+1}\left\{H_{n+1} - \mathbb{E}\left[H_{n+1}|\mathcal{F}_n\right]\right\} + \sum_n \gamma_{n+1} \underbrace{\mathbb{E}\left[H_{n+1}|\mathcal{F}_n\right] - \nabla f(\theta_n)}_{\text{bias, non vanishing when } m_n = m} \\
&= \sum_n \gamma_{n+1}\text{Martingale increment} + \sum_n \gamma_{n+1}\text{Remainder}
\end{aligned}$$

$\hookrightarrow$ the most technical case is "biased approximation" with "fixed batch size"

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 $\theta, \theta' \in \Theta$*

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

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

Acceleration (1)

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 (2) 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 $F(\theta_n) - \min F$