

# Convergence d'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)

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

↔ Acceleration for perturbed Proximal Gradient algorithms (work in progress)

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

↔ Penalized inference in Mixed Models by Proximal Gradient methods (work in progress)

## Motivation : Pharmacokinetic (1/2)

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

### Model:

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

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

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

## Motivation : Pharmacokinetic (1/2)

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

### Model:

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

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

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

### Example of model $\mathcal{F}$ : monocompartmental, oral administration

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

For each patient  $i$ ,

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

## Motivation : Pharmacokinetic (1/2)

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

### Model:

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

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

$Z_i$  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  $\beta$
- selection of the covariates based on  $\hat{\beta}$ .

↪ Penalized Maximum Likelihood

## Motivation : Pharmacokinetic (2/2)

### Model:

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

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

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

### Likelihoods:

- Likelihood: not explicit.
- Complete likelihood: the distribution of  $\{Y_{ij}, X_i; 1 \leq i \leq N, 1 \leq j \leq J\}$  has an explicit expression.
- ML: here, the likelihood is not concave.

## Outline

Penalized Maximum Likelihood inference in models with untractable likelihood

Example 1: Latent variable models

Example 2: Discrete graphical model (Markov random field)

Numerical methods for Penalized ML in such models: Perturbed Proximal Gradient algorithms

Convergence analysis

Conclusion

## Penalized Maximum Likelihood inference with untractable Likelihood

- $N$  observations :  $Y = (Y_1, \dots, Y_N)$
- A parametric statistical model  $\theta \in \Theta \subseteq \mathbb{R}^d$  the dependance upon  $Y$  is omitted

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

- A penalty term on the parameter  $\theta$ :  $\theta \mapsto g(\theta) \geq 0$  for sparsity constraints on  $\theta$ . Usually,  $g$  non-smooth and convex.

Goal: Computation of

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

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

$$\theta \mapsto \log L(\theta) \quad L(\theta) = \int_{\mathbf{X}} p_{\theta}(x) \mu(d\mathbf{x}),$$

where  $\mu$  is a positive  $\sigma$ -finite measure on a set  $\mathbf{X}$ .

- $x$  collect the missing/latent 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).

↔ What about the gradient of the (log)-likelihood ?

## Gradient of the likelihood in a latent variable model

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

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

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

## Gradient of the likelihood in a latent variable model

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

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

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

### The gradient of the log-likelihood

$$\nabla_{\theta} \{\log L(\theta)\} = \int \partial_{\theta} \log p_{\theta}(x) \pi_{\theta}(\mathbf{d}x)$$

is an *untractable expectation* w.r.t. the conditional distribution of the latent variable given the observations  $Y$ .

For all  $(x, \theta)$ ,  $H_{\theta}(x)$  can be evaluated.

## Approximation of the gradient

$$\nabla_{\theta} \{\log L(\theta)\} = \int_{\mathbf{X}} \partial_{\theta} \log p_{\theta}(x) \pi_{\theta}(\mathbf{d}x)$$

- 1 Quadrature techniques: poor behavior w.r.t. the dimension of  $\mathbf{X}$
- 2 use i.i.d. samples from  $\pi_{\theta}$  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  $\pi_{\theta}$ , and define a Monte Carlo approximation. MCMC samplers provide such a chain.

## Approximation of the gradient

$$\nabla_{\theta} \{\log L(\theta)\} = \int_{\mathcal{X}} \partial_{\theta} \log p_{\theta}(x) \pi_{\theta}(\mathrm{d}x)$$

- 1 Quadrature techniques: poor behavior w.r.t. the dimension of  $\mathcal{X}$
- 2 use i.i.d. samples from  $\pi_{\theta}$  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  $\pi_{\theta}$ , and define a Monte Carlo approximation. MCMC samplers provide such a chain.

### Stochastic approximation of the gradient

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

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

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

## 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

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

where  $B$  is a symmetric function.

- $\theta$  is a symmetric  $p \times p$  matrix.
- the normalizing constant (partition function)  $Z_\theta$  can not be computed - sum over  $|X|^p$  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

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

with

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

The gradient of the (log)-likelihood is untractable.



## Approximation of the gradient

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

The Gibbs measure

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

is known up to the constant  $Z_{\theta}$ .

Exact **sampling from  $\pi_{\theta}$  can be approximated** by MCMC samplers (Gibbs-type samplers such as Swendsen-Wang, ...)

**A biased approximation of the gradient is available.**

To summarize,

**Problem:**

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

when

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

To summarize,

**Problem:**

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

when

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

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

- 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

Penalized Maximum Likelihood inference in models with untractable likelihood

**Numerical methods for Penalized ML in such models: Perturbed Proximal Gradient algorithms**

**Algorithms**

**Numerical illustration**

Convergence analysis

Conclusion

## The Proximal-Gradient algorithm (1/2)

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

### The Proximal Gradient algorithm

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

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

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)

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

### The Proximal Gradient algorithm

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

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

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

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

## The proximal-gradient algorithm (2/2)

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

### The Proximal Gradient algorithm

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

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

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

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

## The perturbed proximal-gradient algorithm

### The Perturbed Proximal Gradient algorithm

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

$$\theta_{n+1} = \text{Prox}_{\gamma_{n+1}, g}(\theta_n - \gamma_{n+1} \mathbf{H}_{n+1})$$

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



## Monte Carlo-Proximal Gradient algorithm

In the case:

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

### 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  $\theta_n$ ,

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

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 value of the parameter

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

## Stochastic Approximation-Proximal Gradient algorithm

If in addition,

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

which implies

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

### The SA-Proximal Gradient algorithm

Choose two stepsize sequences  $\{\gamma_n, \delta_n, n \geq 0\}$  and a batch size sequence  $\{m_n, n \geq 0\}$

Given the current value  $\theta_n$ ,

- 1 Sample a Markov chain  $\{X_{j,n}, j \geq 0\}$  from a MCMC sampler with kernel  $P_{\theta_n}(x, dx')$ , and unique invariant distribution  $\pi_{\theta_n} d\mu$ .
- 2 Set  $H_{n+1} = \Phi(\theta_n) + \Psi(\theta_n)S_{n+1}$  with

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

- 3 Update the value of the parameter

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

## (\*) 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

$$\tau_{n+1} = \operatorname{argmax}_{\theta} \int \log p_{\theta}(x) \pi_{\theta}(x) \mathrm{d}\mu(x) = \operatorname{argmax}_{\theta} \{A(\theta) + \langle B(\theta), S_{n+1} \rangle\}$$

with

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

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

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

## (\*) 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**

$$\tau_{n+1} = \operatorname{argmax}_{\theta} \int \log p_{\theta}(x) \pi_{\theta}(x) \mathrm{d}\mu(x) = \operatorname{argmax}_{\theta} \{A(\theta) + \langle B(\theta), S_{n+1} \rangle\}$$

$$A(\tau_{n+1}) + \langle B(\tau_{n+1}), S_{n+1} \rangle \geq A(\tau_n) + \langle B(\tau_n), S_{n+1} \rangle$$

with

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

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

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

## (\*) 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**

$$\tau_{n+1} = \operatorname{argmax}_{\theta} \int \log p_{\theta}(x) \pi_{\theta}(x) \mathrm{d}\mu(x) = \operatorname{argmax}_{\theta} \{A(\theta) + \langle B(\theta), S_{n+1} \rangle\}$$

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

with

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

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

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

## (\* 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.
- (Stochastic) EM algorithms

$$\tau_{n+1} = \operatorname{argmax}_{\theta} \int \log p_{\theta}(x) \pi_{\theta}(x) \mathrm{d}\mu(x) = \operatorname{argmax}_{\theta} \{A(\theta) + \langle B(\theta), S_{n+1} \rangle\}$$

with

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

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

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

- MC-Prox Gdt and SA-Prox GDT are Generalized Penalized EM algorithms (in the convex case).

## Numerical illustration (1/3): pharmacokinetic

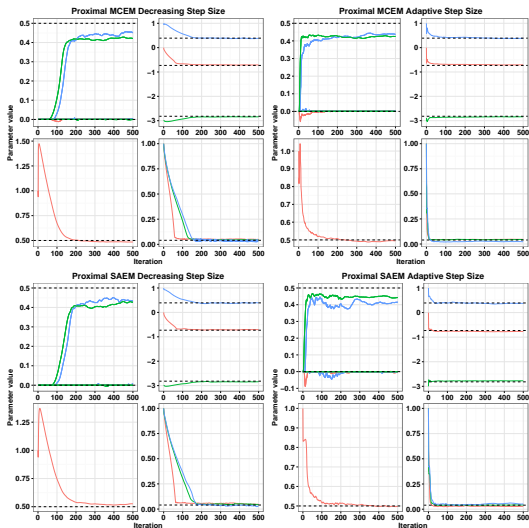
For the implementation of the algorithm

- **Penalty term:**  $g(\theta) = \lambda \|\beta\|_1$ . How to choose  $\lambda$  ?

$$\hookrightarrow \lambda = \operatorname{argmin}_{\lambda_1, \dots, \lambda_L} \text{E-BIC}(\hat{\beta}_\lambda)$$

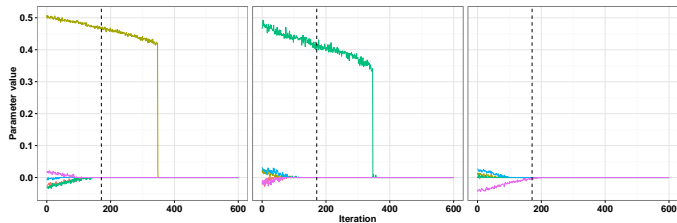
- **Stepsize sequences:** constant or vanishing stepsize sequence  $\{\gamma_n, n \geq 0\}$  ?  
(and  $\delta_n$  for the SA-Prox Gdt algorithm)
- **Monte Carlo approximation:** fixed or increasing batch size ?

## Numerical illustration (2/3): pharmacokinetic





## Numerical illustration (3/3): pharmacokinetic



**Figure:** 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  $\lambda$  value selected by *EBIC*. Each colored curve corresponds to a covariate.

## Outline

Penalized Maximum Likelihood inference in models with untractable likelihood

Numerical methods for Penalized ML in such models: Perturbed Proximal Gradient algorithms

Convergence analysis

Conclusion

## The assumptions

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

where

- the function  $g: \mathbb{R}^d \rightarrow [0, \infty]$  is **convex, non smooth**, not identically equal to  $+\infty$ , and lower semi-continuous
- the function  $f: \mathbb{R}^d \rightarrow \mathbb{R}$  is a **smooth convex function**  
i.e.  $f$  is continuously differentiable and there exists  $L > 0$  such that

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

- $\Theta \subseteq \mathbb{R}^d$  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  $\Theta$ .

## Existing results in the literature

There exist results under (some of) the assumptions

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

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)

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

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

## 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

$$\sum_n \gamma_{n+1}^2 \|\eta_{n+1}\|^2, \quad \sum_n \gamma_{n+1} \eta_{n+1}, \quad \sum_n \gamma_{n+1} \langle \mathbf{T}_n, \eta_{n+1} \rangle$$

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

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:

- Applications in Computational Statistics

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



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

( $X_i$ 's)      the indices  $i \in I_{n+1}$

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

Its proof relies on

- ① a deterministic Lyapunov inequality

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

- 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

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

- 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

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

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

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

In the case

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

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

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

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

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

In the case

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

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

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

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

The most technical case: the biased case with constant batch size  $m_n = m$ Solution  $\hat{H}_{\theta}$  to the Poisson equation:  $H_{\theta} - \pi_{\theta} H_{\theta} = \hat{H}_{\theta} - P_{\theta} \hat{H}_{\theta}$  $H_{n+1} - \nabla f(\theta_n) =$  martingale increment + remainderRegularity in  $\theta$  of  $t \mapsto \hat{H}_t$ .

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*

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

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

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

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

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

*Condition on  $\Theta$ :*  $\Theta$  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:*

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

*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:*

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

*Condition on  $\Theta$ :*  $\Theta$  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  $\theta_\star \in \operatorname{argmin}_\Theta F$ ,

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

where

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

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

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

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

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

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

$$\begin{aligned}\theta_{n+1} &= \text{Prox}_{\gamma_{n+1}, g}(\tau_n - \gamma_{n+1} \nabla f(\tau_n)) \\ \tau_{n+1} &= \theta_{n+1} + \frac{t_n - 1}{t_{n+1}} (\theta_{n+1} - \theta_n)\end{aligned}$$

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

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

(deterministic) Proximal-gradient

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

(deterministic) Accelerated Proximal-gradient

$$F(\theta_n) - \min F = O\left(\frac{1}{n^2}\right)$$

## Acceleration (2) Aujol-Dossal-F.-Moulines, work in progress

### Perturbed Nesterov acceleration: some convergence results

Choose  $\gamma_n, m_n, t_n$  s.t.

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

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)

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

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

## Outline

Penalized Maximum Likelihood inference in models with untractable likelihood

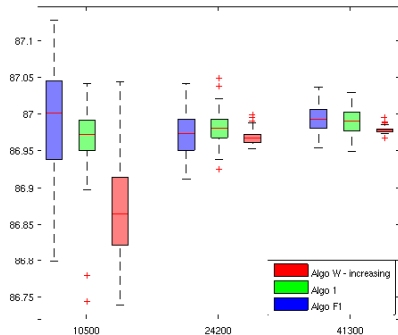
Numerical methods for Penalized ML in such models: Perturbed Proximal Gradient algorithms

Convergence analysis

Conclusion

## Conclusion (1/2): acceleration ?

- with or without the acceleration: complexity  $O(1/\sqrt{n})$ .
- acceleration: longer Markov chains, few iterations.



## Conclusion (2/2): weaken the assumptions

- $\theta \in \mathbb{R}^d \rightarrow \theta$  in a Hilbert space
- $\Theta$  bounded  $\rightarrow$  no boundedness condition on  $\Theta$
- $f$  convex  $\rightarrow f$  non convex