# Stochastic Perturbations of Proximal-Gradient methods for nonsmooth convex optimization: the price of Markovian perturbations

Gersende Fort

LTCI, CNRS and Telecom ParisTech
Paris, France

Based on joint works with

- Eric Moulines (Ecole Polytechnique, France)
- Yves Atchadé (Univ. Michigan, USA)
- Jean-François Aujol (Univ. Bordeaux, France) and Charles Dossal (Univ. Bordeaux, France)

$\hookrightarrow$ `On Perturbed Proximal-Gradient algorithms (2016-v3, arXiv)`

# Outline

Application: Penalized Maximum Likelihood inference in latent variable models

Stochastic Gradient methods (case $g = 0$)

Stochastic Proximal Gradient methods

Rates of convergence

High-dimensional logistic regression with random effects

# Penalized Maximum Likelihood inference, latent variable model

- $N$ observations : $\mathsf{Y} = (Y_1, \cdots, Y_N)$
- A negative normalized log-likelihood of the observations Y, in a latent variable model

$$\theta \mapsto -\frac{1}{N} \log L(\mathsf{Y}, \theta) \qquad L(\mathsf{Y}, \theta) = \int p_\theta(x, \mathsf{Y}) \, \mu(\mathsf{d}x)$$

  where $\theta \in \Theta \subset \mathbb{R}^d$.
- A penalty term on the parameter $\theta$: $\theta \mapsto g(\theta)$ for sparsity constraints on $\theta$; usually non-smooth and convex.

**Goal: Computation of**

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

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

# Latent variable model: example (Generalized Linear Mixed Models)

## GLMM

- $Y_1, \cdots, Y_N$: indep. observations from a Generalized Linear Model.
- Linear predictor

$$\eta_i = \underbrace{\sum_{k=1}^{p} X_{i,k}\beta_k}_{\text{fixed effect}} + \underbrace{\sum_{\ell=1}^{q} Z_{i,\ell}\mathsf{U}_\ell}_{\text{random effect}}$$

where

$X, Z$: covariate matrices
$\beta \in \mathbb{R}^p$: fixed effect parameter
$\mathsf{U} \in \mathbb{R}^q$: **random** effect parameter

# Latent variable model: example (Generalized Linear Mixed Models)

## GLMM

- $Y_1, \cdots, Y_N$: indep. observations from a Generalized Linear Model.
- Linear predictor

$$\eta_i = \underbrace{\sum_{k=1}^{p} X_{i,k}\beta_k}_{\text{fixed effect}} + \underbrace{\sum_{\ell=1}^{q} Z_{i,\ell}\mathsf{U}_\ell}_{\text{random effect}}$$

where

$X, Z$: covariate matrices
$\beta \in \mathbb{R}^p$: fixed effect parameter
$\mathsf{U} \in \mathbb{R}^q$: **random** effect parameter

## Example: logistic regression

- $Y_1, \cdots, Y_N$ binary independent observations: Bernoulli r.v. with mean $p_i = \exp(\eta_i)/(1 + \exp(\eta_i))$

$$(Y_1, \cdots, Y_N)|\mathsf{U} \equiv \prod_{i=1}^{N} \frac{\exp(Y_i \eta_i)}{1 + \exp(\eta_i)}$$

- Gaussian random effect: $\mathsf{U} \sim \mathcal{N}_q$.

## Gradient of the log-likelihood

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

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

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

## Gradient of the log-likelihood

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

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

$$\nabla_\theta \log L(\mathsf{Y}, \theta) = \frac{\int \partial_\theta p_\theta(x, \mathsf{Y})\, \mu(\mathrm{d}x)}{\int p_\theta(z, \mathsf{Y})\, \mu(\mathrm{d}z)} = \int \partial_\theta \log p_\theta(x, \mathsf{Y}) \quad \underbrace{\frac{p_\theta(x, \mathsf{Y})\, \mu(\mathrm{d}x)}{\int p_\theta(z, \mathsf{Y})\, \mu(\mathrm{d}z)}}_{\text{the a posteriori distribution}}$$

### The gradient of the log-likelihood

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

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

# Approximation of the gradient

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

1. Quadrature techniques: poor behavior w.r.t. the dimension of $\mathcal{X}$
2. Monte Carlo approximation with i.i.d. samples: not possible, in general.
3. Markov chain Monte Carlo approximations: sample a Markov chain $\{X_{m,\theta}, m \geq 0\}$ with stationary distribution $\pi_\theta(\mathsf{d}x)$ and set

$$\int_{\mathcal{X}} H_\theta(x) \; \pi_\theta(\mathsf{d}x) \approx \frac{1}{M} \sum_{m=1}^{M} H_\theta(X_{m,\theta})$$

# Approximation of the gradient

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

1. Quadrature techniques: poor behavior w.r.t. the dimension of $\mathcal{X}$
2. Monte Carlo approximation with i.i.d. samples: not possible, in general.
3. Markov chain Monte Carlo approximations: sample a Markov chain $\{X_{m,\theta}, m \geq 0\}$ with stationary distribution $\pi_\theta(\mathrm{d}x)$ and set

$$\int_{\mathcal{X}} H_\theta(x) \; \pi_\theta(\mathrm{d}x) \approx \frac{1}{M} \sum_{m=1}^{M} H_\theta(X_{m,\theta})$$

## Stochastic approximation of the gradient

- *a biased approximation*

$$\mathbb{E}\left[ \frac{1}{M} \sum_{m=1}^{M} H_\theta(X_{m,\theta}) \right] \neq \int H_\theta(x) \; \pi_\theta(\mathrm{d}x).$$

- *if the chain is ergodic "enough", the bias vanishes when $M \to \infty$.*

# To summarize,

## Problem:

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

when

- $\theta \in \Theta \subseteq \mathbb{R}^d$
- $g$ convex non-smooth function (explicit).
- $f$ is $C^1$ and its gradient is of the form

$$\nabla f(\theta) = \int H_\theta(x)\, \pi_\theta(\mathrm{d}x) \approx \frac{1}{M} \sum_{m=1}^{M} H_\theta(X_{m,\theta})$$

where $\{X_{m,\theta}, m \geq 0\}$ is the output of a MCMC sampler with target $\pi_\theta$.

# To summarize,

## Problem:

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

when

- $\theta \in \Theta \subseteq \mathbb{R}^d$
- $g$ convex non-smooth function (explicit).
- $f$ is $C^1$ and its gradient is of the form

$$\nabla f(\theta) = \int H_\theta(x)\, \pi_\theta(\mathrm{d}x) \approx \frac{1}{M} \sum_{m=1}^{M} H_\theta(X_{m,\theta})$$

where $\{X_{m,\theta}, m \geq 0\}$ is the output of a MCMC sampler with target $\pi_\theta$.

## Difficulties:

- **biased** stochastic perturbation of the gradient
- gradient-based methods in the Stochastic Approximation framework (a **fixed** number of Monte Carlo samples)
- weaker conditions on the stochastic perturbation.

# Outline

Application: Penalized Maximum Likelihood inference in latent variable models

Stochastic Gradient methods (case $g = 0$)

Stochastic Proximal Gradient methods

Rates of convergence

High-dimensional logistic regression with random effects

# Perturbed gradient algorithm

**Algorithm:**

*Given a stepsize/learning rate sequence* $\{\gamma_n, n \geq 0\}$*:*

*Initialisation:* $\theta_0 \in \Theta$

*Repeat:*

- *compute* $H_{n+1}$*, an approximation of* $\nabla f(\theta_n)$
- *set* $\theta_{n+1} = \theta_n - \gamma_{n+1} H_{n+1}.$

M. Benaïm. Dynamics of stochastic approximation algorithms. Séminaire de Probabilités de Strasbourg (1999)

A. Benveniste, M. Métivier and P. Priouret, Adaptive Algorithms and Stochastic Approximations, Springer-Verlag, New York, 1990.

V. Borkar. Stochastic Approximation: a dynamical systems viewpoint. Cambridge Univ. Press (2008).

M. Duflo, Random Iterative Systems, Appl. Math. 34, Springer-Verlag, Berlin, 1997.

H. Kushner, G. Yin. Stochastic Approximation and Recursive Algorithms and Applications. Springer Book (2003).

## Sufficient conditions for the convergence

Set $\mathcal{L} = \{\theta \in \Theta : \nabla f(\theta) = 0\}, \qquad \eta_{n+1} = H_{n+1} - \nabla f(\theta_n).$

**Theorem** (Andrieu-Moulines-Priouret(2005); F.-Moulines-Schreck-Vihola(2016))

*Assume*

- *the level sets of $f$ are compact subsets of $\Theta$ and $\mathcal{L}$ is in a level set of $f$.*
- $\sum_n \gamma_n = +\infty$ *and* $\sum_n \gamma_n^2 < \infty$.
- $\sum_n \gamma_n \eta_{n+1} \mathbb{1}_{\theta_n \in \mathcal{K}} < \infty$ *for any compact subset $\mathcal{K}$ of $\Theta$.*

*Then*

(i) *there exists a compact subset $\mathcal{K}_\star$ of $\Theta$ s.t. $\theta_n \in \mathcal{K}_\star$ for all $n$.*

(ii) $\{f(\theta_n), n \geq 0\}$ *converges to a connected component of $f(\mathcal{L})$.*

*If in addition $\nabla f$ is locally lipschitz and $\sum_n \gamma_n^2 \|\eta_n\|^2 \mathbb{1}_{\theta_n \in \mathcal{K}} < \infty$, then $\{\theta_n, n \geq 0\}$ converges to a connected component of $\{\theta : \nabla f(\theta) = 0\}$.*

# When $H_{n+1}$ is a Monte Carlo approximation (1)

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

Two strategies:

(1) Stochastic Approximation (fixed batch size)

$$H_{n+1} = H_{\theta_n}(X_{1,n}),$$

(2) Monte Carlo assisted optimization (increasing batch size)

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

where $\{X_{m,n}\}_m$ "approximate" the target $\pi_{\theta_n}(\mathrm{d}x)$.

# When $H_{n+1}$ is a Monte Carlo approximation (2)

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

- With i.i.d. Monte Carlo:

$$\mathbb{E}\left[H_{n+1} | \mathcal{F}_n\right] = \nabla f(\theta_n) \qquad \text{unbiased approximation}$$

- With Markov chain Monte Carlo approximation

$$\mathbb{E}\left[H_{n+1} | \mathcal{F}_n\right] \neq \nabla f(\theta_n) \qquad \text{Biased approximation !}$$

## When $H_{n+1}$ is a Monte Carlo approximation (2)

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

- With i.i.d. Monte Carlo:

$$\mathbb{E}\left[H_{n+1} | \mathcal{F}_n\right] = \nabla f(\theta_n) \qquad \text{unbiased approximation}$$

- With Markov chain Monte Carlo approximation

$$\mathbb{E}\left[H_{n+1} | \mathcal{F}_n\right] \neq \nabla f(\theta_n) \qquad \text{Biased approximation !}$$

and the bias:

$$\left|\mathbb{E}\left[H_{n+1} | \mathcal{F}_n\right] - \nabla f(\theta_n)\right| = O_{L^p}\left(\frac{1}{M_{n+1}}\right)$$

does not vanish when the size of the batch is fixed.

## When $H_{n+1}$ is a Monte Carlo approximation (3)

$$\theta_{n+1} = \theta_n - \gamma_{n+1} H_{n+1}$$

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

### MCMC approx. and fixed batch size

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

### i.i.d. MC approx. / MCMC approx with increasing batch size

$$\sum_n \gamma_n = +\infty \qquad \sum_n \frac{\gamma_n^2}{M_n} < \infty \qquad \sum_n \frac{\gamma_n}{M_n} < \infty \text{ (case MCMC)}$$

# A remark on the proof

$$\sum_{n=1}^{N} \gamma_{n+1} \left( H_{n+1} - \nabla f(\theta_n) \right) = \sum_{n=1}^{N} \gamma_{n+1} \left( \underbrace{\Delta_{n+1}}_{\text{martingale increment}} + \underbrace{R_{n+1}}_{\text{remainder term}} \right)$$

$$= \text{Martingale} + \text{Remainder}$$

## How to define $\Delta_{n+1}$ ?

**unbiased** MC approx $\Delta_{n+1} = H_{n+1} - \nabla f(\theta_n)$

**biased** MC approx with **increasing batch size** $\Delta_{n+1} = H_{n+1} - \mathbb{E}\left[H_{n+1} | \mathcal{F}_n\right]$

**biased** MC approx with **fixed batch size** technical !

Stochastic Approximation with MCMC inputs: see e.g.

Benveniste-Metivier-Priouret (1990) Springer-Verlag.

Duflo (1997) Springer-Verlag.

Andrieu-Moulines-Priouret (2005) SIAM Journal on Control and Optimization.

F.-Moulines-Priouret (2012) Annals of Statistics.

F.-Jourdain-Lelièvre-Stoltz (2015,2016) Mathematics of Computation, Statistics and Computing.

F.-Moulines-Schreck-Vihola (2016) SIAM Journal on Control and Optimization.

# Outline

Application: Penalized Maximum Likelihood inference in latent variable models

Stochastic Gradient methods (case $g = 0$)

Stochastic Proximal Gradient methods

Rates of convergence

High-dimensional logistic regression with random effects

# Problem:

A gradient-based method for solving

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

when

- $g$ is non-smooth and convex
- $f$ is $C^1$ and
$$\nabla f(\theta) = \int_{\mathsf{X}} H_\theta(x)\, \pi_\theta(\mathrm{d}x).$$
- Available: Monte Carlo approximation of $\nabla f(\theta)$ through Markov chain samples.

# The setting, hereafter

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

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 $g$: $\Theta = \{\theta : g(\theta) < \infty\}$.

# The proximal-gradient algorithm

## The Proximal Gradient algorithm

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

*where*

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

Proximal map: Moreau(1962); Parikh-Boyd(2013);

Proximal Gradient algorithm: Nesterov(2004); Beck-Teboulle(2009)

### About the Prox-step:

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

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

# The perturbed proximal-gradient algorithm

## The Perturbed Proximal Gradient algorithm

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

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

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. fixed stepsize, increasing batch size and unverifiable conditions for MCMC sampling

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*

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

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

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

# When $H_{n+1}$ is a Monte Carlo approximation

$$\theta_{n+1} = \mathrm{Prox}_{\gamma_{n+1},g}\left(\theta_n - \gamma_{n+1} H_{n+1}\right)$$
$$H_{n+1} = \frac{1}{M_{n+1}} \sum_{j=1}^{M_{n+1}} H_{\theta_n}(X_{j,n}) \approx \nabla f(\theta_n)$$

**MCMC approx. and fixed batch size**

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

**i.i.d. MC approx. / MCMC approx with increasing batch size**

$$\sum_n \gamma_n = +\infty \qquad \sum_n \frac{\gamma_n^2}{M_n} < \infty \qquad \sum_n \frac{\gamma_n}{M_n} < \infty \text{ (case MCMC)}$$

↪ Same conditions as in the Stochastic Gradient algorithm

# Outline

Application: Penalized Maximum Likelihood inference in latent variable models

Stochastic Gradient methods (case $g = 0$)

Stochastic Proximal Gradient methods

Rates of convergence

High-dimensional logistic regression with random effects

# 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

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

# A deterministic control

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

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

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

*where*

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

## 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*

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

*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*

$$\gamma_n = \gamma_\star \qquad M_n = m_\star n.$$

*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 (1983); Beck-Teboulle (2009)

AllenZhu-Orecchia (2015); Attouch-Peypouquet(2015); Bubeck-TatLee-Singh(2015); Su-Boyd-Candes(2015)

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

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

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

## Perturbed Nesterov acceleration: some convergence results

*Choose* $\gamma_n, M_n, t_n$ *s.t.*

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

*Then there exists* $\theta_\star \in \mathrm{argmin}_\Theta F$ *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

Application: Penalized Maximum Likelihood inference in latent variable models

Stochastic Gradient methods (case $g = 0$)

Stochastic Proximal Gradient methods

Rates of convergence

High-dimensional logistic regression with random effects

# Logistic regression with random effects

## The model

- Given $\mathsf{U} \in \mathbb{R}^q$,

$$Y_i \sim \mathcal{B}\left(\frac{\exp(x_i'\beta + \sigma z_i'\mathsf{U})}{1 + \exp(x_i'\beta + \sigma z_i'\mathsf{U})}\right), \qquad i = 1, \cdots, N.$$

- $\mathsf{U} \sim \mathcal{N}_q(0, I)$
- Unknown parameters: $\beta \in \mathbb{R}^p$ and $\sigma^2 > 0$.

## Stochastic approximation of the gradient of $f$

$$\nabla f(\theta) = \int H_\theta(u) \pi_\theta(\mathsf{d}u)$$

with

$$\pi_\theta(u) \propto \mathcal{N}(0, I)[u] \prod_{i=1}^{N} \frac{\exp\left(Y_i(x_i'\beta + \sigma z_i'u)\right)}{1 + \exp(x_i'\beta + \sigma z_i'u)}$$

$\hookrightarrow$ sampled by MCMC Polson-Scott-Windle (2013)

## Numerical illustration

- The Data set simulated: $N = 500$ observations, a sparse covariate vector $\beta_{\mathrm{true}} \in \mathbb{R}^{1000}$, $q = 5$ random effects.
- Penalty term elastic net on $\beta$, and $\sigma > 0$.
- Comparison of 5 algorithms

  Algo1 fixed batch size: $\gamma_n = 0.01/\sqrt{n}$ $M_n = 275$
  Algo2 fixed batch size: $\gamma_n = 0.5/n$ $M_n = 275$

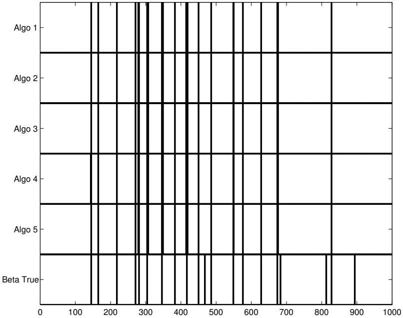
  Algo3 increasing batch size: $\gamma_n = 0.005$ $M_n = 200 + n$
  Algo4 increasing batch size: $\gamma_n = 0.001$ $M_n = 200 + n$

  Algo5 increasing batch size: $\gamma_n = 0.05/\sqrt{n}$ $M_n = 270 + \sqrt{n}$

  After 150 iterations, the algorithms use the same number of MC draws.

# A sparse limiting value

Displayed: for each algorithm, the non-zero entries of the limiting value $\beta_\infty \in \mathbb{R}^{1000}$ of a path $(\beta_n)_n$




Algo1 $\gamma_n = 0.01/\sqrt{n}$ $M_n = 275$
Algo2 $\gamma_n = 0.5/n$ $M_n = 275$

Algo3 $\gamma_n = 0.005$ $M_n = 200 + n$
Algo4 $\gamma_n = 0.001$ $M_n = 200 + n$

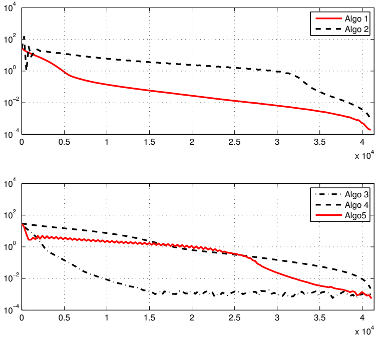
Algo5 $\gamma_n = 0.05/\sqrt{n}$ $M_n = 270 + \sqrt{n}$

# Relative error

Displayed: For each algorithm, relative error

$$\frac{\|\beta_n - \beta_{150}\|}{\|\beta_{150}\|}$$

as a function of the total number of MC draws up to time $n$.




(⋆) Algo1 $\gamma_n = 0.01/\sqrt{n}$ $M_n = 275$

Algo2 $\gamma_n = 0.5/n$ $M_n = 275$

(⋆) Algo3 $\gamma_n = 0.005$ $M_n = 200 + n$

Algo4 $\gamma_n = 0.001$ $M_n = 200 + n$

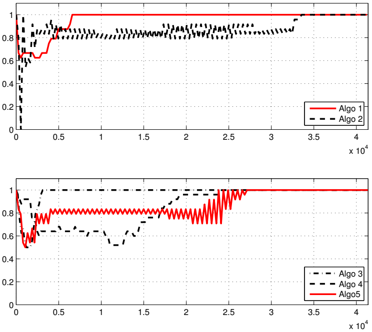
Algo5 $\gamma_n = 0.05/\sqrt{n}$ $M_n = 270 + \sqrt{n}$

# Recovery of the sparsity structure of $\beta_\infty (= \beta_{150})$ (1)

Displayed: For each algorithm, the sensitivity

$$\frac{\sum_{i=1}^{1000} \mathbb{1}_{|\beta_{n,i}|>0} \mathbb{1}_{|\beta_{\infty,i}|>0}}{\sum_{i=1}^{1000} \mathbb{1}_{|\beta_{\infty,i}|>0}}$$

as a function of the total number of MC draws up to time $n$.




($\star$) Algo1 $\gamma_n = 0.01/\sqrt{n}$ $M_n = 275$

Algo2 $\gamma_n = 0.5/n$ $M_n = 275$

($\star$) Algo3 $\gamma_n = 0.005$ $M_n = 200 + n$

Algo4 $\gamma_n = 0.001$ $M_n = 200 + n$

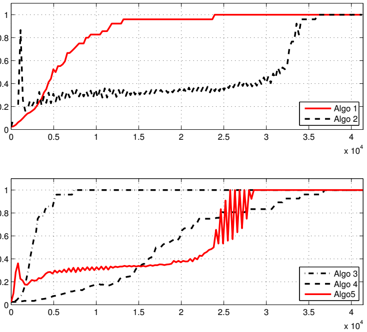
Algo5 $\gamma_n = 0.05/\sqrt{n}$ $M_n = 270 + \sqrt{n}$

# Recovery of the sparsity structure of $\beta_\infty (= \beta_{150})$ (2)

Displayed: For each algorithm, the precision

$$\frac{\sum_{i=1}^{1000} \mathbb{1}_{|\beta_{n,i}|>0} \mathbb{1}_{|\beta_{\infty,i}|>0}}{\sum_{i=1}^{1000} \mathbb{1}_{|\beta_{n,i}|>0}}$$

as a function of the total number of MC draws up to time $n$.




($\star$) Algo1 $\gamma_n = 0.01/\sqrt{n}$ $M_n = 275$

Algo2 $\gamma_n = 0.5/n$ $M_n = 275$

($\star$) Algo3 $\gamma_n = 0.005$ $M_n = 200 + n$

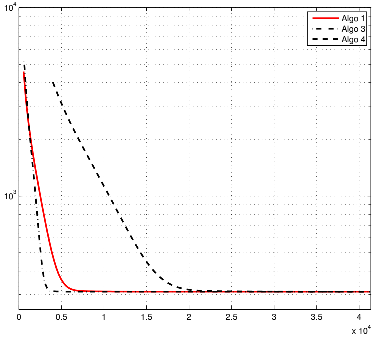
Algo4 $\gamma_n = 0.001$ $M_n = 200 + n$

Algo5 $\gamma_n = 0.05/\sqrt{n}$ $M_n = 270 + \sqrt{n}$

## Convergence of $\mathbb{E}\left[F(\theta_n)\right]$

In this example, the mixed effects are chosen so that $F(\theta)$ can be approximated.

Displayed: For some algorithm, a Monte Carlo approximation of $\mathbb{E}\left[F(\theta_n)\right]$ over 50 indep. runs as a function of the total number of MC draws up to time $n$.




($\star$) Algo1 $\gamma_n = 0.01/\sqrt{n}$ $M_n = 275$

($\star$) Algo3 $\gamma_n = 0.005$ $M_n = 200 + n$

Algo4 $\gamma_n = 0.001$ $M_n = 200 + n$