Stochastic Proximal Gradient Algorithm

Eric Moulines

Institut Mines-Télécom / Telecom ParisTech / Laboratoire Traitement et Communication de l'Information

Joint work with: Y. Atchadé, G. Fort

High-dimensional logistic regression with random effects

- Observations : N observations $\mathbf{Y} \in \{0,1\}^N$
- Random effect : Conditionally to $\mathbf{U}$, for all $i = 1, \cdots, N$,

$$Y_i \overset{\text{ind.}}{\sim} \mathcal{B}\left(\frac{\exp(\eta_i)}{1+\exp(\eta_i)}\right)$$

where

$$\begin{bmatrix} \eta_1 \\ \cdots \\ \eta_N \end{bmatrix} = \mathbf{X}\beta_{\mathsf{true}} + \sigma_{\mathsf{true}}\mathbf{Z}\mathbf{U}$$

- The regressors $\mathbf{X} \in \mathbb{R}^{N \times p}$ and the factor loadings $\mathbf{Z} \in \mathbb{R}^{N \times q}$, known.
- Objective: estimate $\beta_{\mathsf{true}} \in \mathbb{R}^p, \sigma_{\mathsf{true}} > 0$.

Penalized likelihood

- log-likelihood : Taking $\mathbf{U} \sim \mathcal{N}_q(0, I)$, setting

$$\theta = (\beta, \sigma) \qquad\qquad F(\eta) = \frac{\exp(\eta)}{1 + \exp(\eta)}$$

the log-likelihood of the observations $\mathbf{Y}$ (with respect to θ) is

$$\ell(\theta) = \log \int \prod_{i=1}^{N} \left\{F\left(\mathbf{X}_{i\cdot}\beta + \sigma(\mathbf{Z}\mathbf{U})_i\right)\right\}^{Y_i} \left\{1 - F\left(\mathbf{X}_{i\cdot}\beta + \sigma(\mathbf{Z}\mathbf{U})_i\right)\right\}^{1-Y_i} \phi(\mathbf{u})\mathrm{d}\mathbf{u}$$

- Elastic net penalty

$$g_{\lambda,\theta}(\theta) = \lambda \left(\frac{1-\alpha}{2}\|\beta\|_2^2 + \alpha\|\beta\|_1\right)$$

$$\tilde{g}_{\mathcal{C}}(\theta) = \begin{cases} 0 & \text{si } \theta \in \mathcal{C} \\ +\infty & \text{otherwise} \end{cases}$$

Penalized likelihood

$$\min_{\theta \in \Theta} \left(f(\theta) + g(\theta) \right) \ , \quad f(\theta) = -\ell(\theta) \ ,$$

with

$$\ell(\theta) = \log \int \exp\left(\ell_c(\theta|\mathbf{u})\right) \ \phi(\mathbf{u}) \mathrm{d}\mathbf{u}$$

$$\ell_c(\theta|\mathbf{u}) = \sum_{i=1}^{N} \left\{ Y_i \left(\mathbf{X}_{i\cdot}\beta + \sigma(\mathbf{ZU})_i \right) - \ln\left(1 + \exp\left(\mathbf{X}_{i\cdot}\beta + \sigma(\mathbf{ZU})_i\right)\right) \right\}$$

Gradient :

$$\nabla\ell(\theta) = \int \nabla\ell_c(\theta|\mathbf{u}) \pi_\theta(\mathbf{u}) \mathrm{d}\mathbf{u}$$

where $\pi_\theta(\mathbf{u})$ is the posterior distribution of the random effect given the observations

$$\pi_\theta(\mathbf{u}) = \exp\left(\ell_c(\theta|\mathbf{u}) - \ell(\theta)\right) \ \phi(\mathbf{u})$$

Penalized likelihood

$$\min_{\theta \in \Theta} \left(f(\theta) + g(\theta) \right) , \quad f(\theta) = -\ell(\theta)$$

where

$$g_{\lambda,\theta}(\theta) = \lambda \left(\frac{1-\alpha}{2} \|\beta\|_2^2 + \alpha \|\beta\|_1 \right) + \mathbb{I}_{\mathcal{C}}(\theta)$$

$$\mathbb{I}_{\mathcal{C}}(\theta) = \begin{cases} 0 & \text{if } \theta \in \mathcal{C} \\ +\infty & \text{otherwise} \end{cases} \qquad \mathcal{C} \text{ compact convex set}$$

$\hookrightarrow$ proper convex,
lower-semi continuous, not differentiable.

Wrap-up

Solve

$$\operatorname{argmin}_{\theta\in\Theta}\left(f(\theta)+g(\theta)\right)$$

where

- $f(\theta)=-\ell(\theta)$ not necessarily convex, gradient Lipschitz

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

- $f,\nabla f$ are intractable (but ∇f is the conditional expectation of the complete data likelihood).
- g is closed convex but non-smooth.

Definition

- Definition: Proximal mapping associated with closed convex function g and stepsize γ

$$\mathrm{prox}_\gamma(\theta) = \mathrm{Argmin}_{\vartheta \in \Theta} \left(g(\vartheta) + (2\gamma)^{-1} \|\vartheta - \theta\|_2^2 \right)$$

- The uniqueness of the minimizer stems from the strong convexity of the function $\vartheta \mapsto g(\vartheta) + 1/(2\gamma)\|\vartheta - \theta\|_2^2$
- If $g = \mathbb{I}_{\mathcal{K}}$, where $\mathcal{K}$ is a closed convex set, then prox_γ is the Euclidean projection on $\mathcal{K}$

$$\mathrm{prox}_\gamma(\theta) = \mathrm{Argmin}_{\vartheta \in \mathcal{K}} \|\vartheta - \theta\|_2^2 = P_{\mathcal{K}}(\theta)$$

- The proximal operator may be seen as a generalisation of the projection on closed convex sets.

Proximal operator

Lemma

If $\theta = (\theta_1, \theta_2, \dots, \theta_p)$ *and* $g(\theta) = \sum_{i=1}^{p} g_i(\theta_i)$, *then*

$$prox_{\gamma g}(\theta) = (\mathit{prox}_{\gamma g_1}(\theta_1), \mathit{prox}_{\gamma g_2}(\theta_2), \dots, \mathit{prox}_{\gamma g_p}(\theta_p))$$

Proximal operator

Lemma

If $\theta = (\theta_1, \theta_2, \ldots, \theta_p)$ *and* $g(\theta) = \sum_{i=1}^p g_i(\theta_i)$, *then*

$$prox_{\gamma g}(\theta) = (\mathrm{prox}_{\gamma g_1}(\theta_1), \mathrm{prox}_{\gamma g_2}(\theta_2), \ldots, \mathrm{prox}_{\gamma g_p}(\theta_p))$$

$$\begin{aligned}\mathrm{Argmin}_{(\vartheta_1,\ldots,\theta_p)} \sum_{i=1}^p g_i(\vartheta_i) + 2\gamma^{-1} \sum_{i=1}^p \|\vartheta_i - \theta_i\|^2 \\ = \sum_{i=1}^p \mathrm{Argmin}_{\vartheta_i}\, g_i(\vartheta_i) + (2\gamma)^{-1} \|\vartheta_i - \theta_i\|^2\end{aligned}$$

A characterization of the proximal operator

Theorem

Let g be a convex function on Θ, $(\theta, p) \in \Theta^2$,

$$p = \mathrm{prox}_{\gamma g}(\theta) \iff \text{for all } \vartheta \in \Theta, \quad g(p) + \gamma^{-1} \langle \vartheta - p, \theta - p \rangle \leq g(\vartheta)$$

i.e. p is the unique element of Θ satisfying $\gamma^{-1}(\theta - p) \in \partial g(p)$.

A characterization of the proximal operator

Theorem

Let g be a convex function on Θ, $(\theta, p) \in \Theta^2$,

$$p = \text{prox}_{\gamma g}(\theta) \iff \text{for all } \vartheta \in \Theta, \quad g(p) + \gamma^{-1} \langle \vartheta - p, \theta - p \rangle \leq g(\vartheta)$$

i.e. p is the unique element of Θ satisfying $\gamma^{-1}(\theta - p) \in \partial g(p)$.

For all $\alpha \in [0, 1)$

$$g(p) \leq \alpha g(\vartheta) + (1 - \alpha) g(p) \\ + (2\gamma)^{-1} \|\alpha\vartheta + (1 - \alpha)p - \theta\|^2 - (2\gamma)^{-1} \|p - \theta\|^2$$

Conclude by letting $\alpha \to 0$. Follows also from the characterization of the subdifferential, $0 \in \partial g(p) + \gamma^{-1}(p - \theta)$.

Examples

- If $g(\theta) = (1/2)\theta' A\theta + b'\theta + c$ then

$$\mathrm{prox}_{\gamma g}(\theta) = (A + (2\gamma)^{-1} I)^{-1}(\theta - (2\gamma)^{-1} b)$$

- If $g(\theta) = \|\theta\|$, then

$$\mathrm{prox}_{\gamma g}(\theta) = \begin{cases} (1 - (2\gamma)^{-1}/\|\theta\|)\theta & \|\theta\| \geq 2\gamma \\ 0 & \text{otherwise} \end{cases}$$

Proximal operator: LASSO and Elastic net

- If $g(\theta) = \sum_{i=1}^{p} \lambda_i |\theta_i|$ then prox_g is shrinkage (soft threshold) operation

$$[S_{\lambda,\gamma}(\theta)]_i = \begin{cases} \theta_i - \gamma\lambda_i & \theta_i \geq \gamma\lambda_i \\ 0 & |\theta_i| \leq \gamma\lambda_i \\ \theta_i + \gamma\lambda_i & \theta_i \leq -\gamma\lambda_i \end{cases}$$

- If $g(\theta) = \lambda\left((1-\alpha)/2\|\theta\|_2^2 + \alpha\|\theta\|_1\right)$

$$(\mathrm{Prox}_\gamma(\tau))_i = \frac{1}{1+\gamma\lambda(1-\alpha)} \begin{cases} \tau_i - \gamma\lambda\alpha & \text{if } \tau_i \geq \gamma\lambda\alpha \\ \tau_i + \gamma\lambda\alpha & \text{if } \tau_i \leq -\gamma\lambda\alpha \\ 0 & \text{otherwise} \end{cases}$$

Fixed points of the proximal operator

Theorem

Let g be a proper convex function on Θ. The fixed points $\{\theta \in \Theta, \mathrm{Prox}_{\gamma g}(\theta) = \theta\}$ coincide with the minimum of g

Fixed points of the proximal operator

Theorem

Let g be a proper convex function on Θ. The fixed points $\{\theta \in \Theta, \mathrm{Prox}_{\gamma g}(\theta) = \theta\}$ coincide with the minimum of g

If $p = \mathrm{prox}_{\gamma g}(\theta)$, then $\gamma^{-1}(\theta - p) \in \partial g(p)$ which implies that $g(p) + \gamma^{-1} \langle \theta - p, \vartheta - p \rangle \leq g(\vartheta)$. Then,

$$
\begin{aligned}
& p = \mathrm{prox}_{\gamma g}(p) \\
& \Longleftrightarrow \text{for all } \vartheta \in \Theta, \gamma^{-1} \langle \vartheta - p, p - p \rangle + g(p) \leq g(\vartheta) \\
& \Longleftrightarrow \text{for all } \vartheta \in \Theta, g(p) \leq g(\vartheta) \ .
\end{aligned}
$$

Firm non-expansiveness

Theorem

If g is a proper convex function, then $\mathrm{prox}_{\gamma g}$ and $(I - \mathrm{prox}_{\gamma g})$ are firmly non-expansive (or co-coercive with constant 1), i.e. for all $\theta, \vartheta \in \Theta$,

$$\|p - q\|^2 + \|(\theta - p) - (\vartheta - q)\|^2 \leq \|\theta - \vartheta\|^2 ,$$
$$\iff \quad \langle p - q, \theta - \vartheta \rangle \geq \|p - q\|^2 .$$

where $p = \mathrm{prox}_{\gamma g}(\theta)$ and $q = \mathrm{prox}_{\gamma g}(\vartheta)$.

Firm non-expansiveness

Theorem

If g is a proper convex function, then $\mathrm{prox}_{\gamma g}$ and $(I - \mathrm{prox}_{\gamma g})$ are firmly non-expansive (or co-coercive with constant 1), i.e. for all $\theta, \vartheta \in \Theta$,

$$\|p-q\|^2 + \|(\theta - p) - (\vartheta - q)\|^2 \leq \|\theta - \vartheta\|^2 \, ,$$
$$\iff \quad \langle p - q, \theta - \vartheta \rangle \geq \|p-q\|^2 \, .$$

where $p = \mathrm{prox}_{\gamma g}(\theta)$ and $q = \mathrm{prox}_{\gamma g}(\vartheta)$.

$$\gamma^{-1} \langle q - p, \theta - p \rangle + g(p) \leq g(q) \quad \gamma^{-1} \langle p - q, \vartheta - q \rangle + g(q) \leq g(p)$$

Adding these two equations yield

$$\langle p - q, (\theta - p) - (\vartheta - q) \rangle \geq 0 \, .$$

Conclude by writing $\|\theta - \vartheta\|^2 = \|p - q + (\theta - p) - (\vartheta - q)\|^2$.

Proximal gradient algorithm

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

where

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

Majorization-Minimization interpretation

- Since f is gradient Lipshitz, for all $\gamma \in (0, 1/L]$

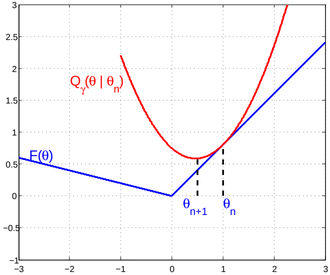
$$F(\vartheta) = f(\vartheta) + g(\vartheta) \leq f(\theta) + \langle \nabla f(\theta), \vartheta - \theta \rangle + \frac{1}{2\gamma} \|\theta - \vartheta\|^2 + g(\vartheta)$$

- Consider the following surrogate function

$$Q_\gamma(\vartheta|\theta) = f(\theta) + \langle \nabla f(\theta), \vartheta - \theta \rangle + \frac{1}{2\gamma} \|\theta - \vartheta\|^2 + g(\vartheta)$$

- For all $\theta \in \Theta$, $\vartheta \mapsto Q_\gamma(\vartheta|\theta)$ is strongly convex and has a unique minimum and

$$F(\vartheta) \leq Q_\gamma(\vartheta|\theta) \qquad F(\theta) = Q_\gamma(\theta|\theta)$$



$$F(\vartheta) \leq Q_\gamma(\vartheta|\theta_n) \qquad F(\theta_n) = Q_\gamma(\theta_n|\theta_n)$$

Majorization-Minimization interpretation

The proximal gradient algorithm is a special instance of the Majorization-Minimization framework !

$$\begin{aligned} Q_\gamma(\vartheta|\theta) &\stackrel{\text{def}}{=} f(\theta) + \langle \nabla f(\theta), \vartheta - \theta \rangle + \frac{1}{2\gamma} \|\vartheta - \theta\|^2 + g(\vartheta) \\ &= f(\theta) + \frac{1}{2\gamma} \|\vartheta - (\theta - \gamma \nabla f(\theta))\|^2 - \frac{\gamma}{2} \|\nabla f(\theta)\|^2 + g(\vartheta) \ , \end{aligned}$$

The iterates of the proximal gradient algorithms may be rewritten as $\theta_{n+1} = T_{\gamma_{n+1}}(\theta_n)$ with the point-to-point map T_γ defined by

$$\begin{aligned} T_\gamma(\theta) &\stackrel{\text{def}}{=} \mathrm{Prox}_\gamma \left(\theta - \gamma \nabla f(\theta)\right) \\ &= \mathrm{argmin}_{\vartheta \in \mathrm{Dom}(g)} Q_\gamma(\vartheta|\theta) \ . \end{aligned}$$

Proximal gradient

- If $g(\theta) = 0$, $\hookrightarrow$ gradient proximal = classical stochastic gradient

$$\theta_{n+1} = \theta_n - \gamma_{n+1} \nabla f(\theta_n)$$

Proximal gradient

- If $g(\theta) = 0$, $\hookrightarrow$ gradient proximal = classical stochastic gradient

$$\theta_{n+1} = \theta_n - \gamma_{n+1} \nabla f(\theta_n)$$

- If $g(\theta) = 0$ if $\theta \in \mathcal{C}$ and $g(\theta) = +\infty$ otherwise where $\mathcal{C}$ is a closed convex set,

$$\mathrm{Prox}_\gamma(\tau) = \min_{\theta \in \mathcal{C}} \|\tau - \theta\|^2$$

$\hookrightarrow$ gradient proximal = projected gradient

$$\theta_{n+1} = \Pi_{\mathcal{C}} \left(\theta_n - \gamma_{n+1} \nabla f(\theta_n) \right)$$

Proximal gradient for the elastic net penalty

If $g(\theta) = \lambda \left(\frac{1-\alpha}{2} \|\theta\|_2^2 + \alpha \|\theta\|_1 \right)$

$$(\mathrm{Prox}_\gamma(\tau))_i = \frac{1}{1+\gamma\lambda(1-\alpha)} \begin{cases} \tau_i - \gamma\lambda\alpha & \text{if } \tau_i \geq \gamma\lambda\alpha \\ \tau_i + \gamma\lambda\alpha & \text{if } \tau_i \leq -\gamma\lambda\alpha \\ 0 & \text{otherwise} \end{cases}$$

$\hookrightarrow$ Proximal gradient= soft-thresholded gradient

$$\theta_{n+1} = \mathcal{S}_{\alpha,\lambda,\gamma_{n+1}} \left(\theta_n - \gamma_{n+1} \nabla f(\theta_n) \right)$$

Assumptions

$$\textbf{(P)} \quad \min_{\theta \in \Theta} F(\theta) \qquad F(\theta) = f(\theta) + g(\theta),$$

Assumptions

(A0) Θ finite dimensional euclidean space

(A1) $g: \ \Theta \to (-\infty, +\infty]$ closed convex
$f : \Theta \to \mathbb{R}$ is continuously differentiable and ∇f is gradient Lipshitz: for all $\theta, \theta' \in \Theta$,

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

Stationary points of the proximal gradient

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

where T_γ is the proximal map,

$$T_\gamma(\theta) \stackrel{\text{def}}{=} \operatorname{Prox}_\gamma\left(\theta - \gamma \nabla f(\theta)\right) = \operatorname{argmin}_{\vartheta \in \operatorname{Dom}(g)} Q_\gamma(\vartheta | \theta) .$$

Theorem

Under A0 and A1:

$$\mathcal{L} = \{\theta : \theta = \operatorname{Prox}_{\gamma g}(\theta - \gamma \nabla f(\theta))\} = \{\theta \in \operatorname{Dom}(g) : 0 \in \nabla f(\theta) + \partial g(\theta)\}.$$

If in addition f is convex then $\mathcal{L}$ is the set of the global minimizers of F.

Fixed points of the proximal map

Denote $F(\theta) = f(\theta) + g(\theta)$. Then

$$\begin{aligned} 0 \in \partial F(\theta) &\iff \quad 0 \in \partial \gamma F(\theta) \\ &\iff \quad 0 \in \gamma \nabla f(\theta) + \partial \gamma g(\theta) \\ &\iff \quad \theta - \gamma \nabla f(\theta) \in (\theta + \gamma \partial g(\theta)) \end{aligned}$$

Fixed points of the proximal map

Denote $F(\theta) = f(\theta) + g(\theta)$. Then

$$\begin{aligned} 0 \in \partial F(\theta) \iff & \quad 0 \in \partial \gamma F(\theta) \\ \iff & \quad 0 \in \gamma \nabla f(\theta) + \partial \gamma g(\theta) \\ \iff & \quad \theta - \gamma \nabla f(\theta) \in (\theta + \gamma \partial g(\theta)) \end{aligned}$$

Recall that, for any ϑ

$$p = \mathrm{prox}_{\gamma g}(\vartheta) \iff (\vartheta - p) \in \gamma \partial g(p) \iff \vartheta \in p + \gamma \partial g(p).$$

Fixed points of the proximal map

Denote $F(\theta) = f(\theta) + g(\theta)$. Then

$$\begin{aligned} 0 \in \partial F(\theta) \iff & \quad 0 \in \partial \gamma F(\theta) \\ \iff & \quad 0 \in \gamma \nabla f(\theta) + \partial \gamma g(\theta) \\ \iff & \quad \theta - \gamma \nabla f(\theta) \in (\theta + \gamma \partial g(\theta)) \end{aligned}$$

Recall that, for any ϑ

$$p = \mathrm{prox}_{\gamma g}(\vartheta) \iff (\vartheta - p) \in \gamma \partial g(p) \iff \vartheta \in p + \gamma \partial g(p).$$

Hence, taking $p \leftarrow \theta$ and $\vartheta \leftarrow \theta - \gamma \nabla f(\theta)$

$$0 \in \partial F(\theta) \iff \theta = T_\gamma(\theta)$$

Lyapunov function

$$Q_\gamma(\vartheta|\theta) = f(\theta) + \langle \nabla f(\theta), \vartheta - \theta \rangle + \frac{1}{2\gamma}\|\theta - \vartheta\|^2 + g(\vartheta)$$

- For all $\theta \in \Theta$, $F \circ T_\gamma(\theta) \leq F(\theta)$:

$$F \circ T_\gamma(\theta) \leq Q_\gamma(T_\gamma(\theta)|\theta) \leq Q_\gamma(\theta|\theta) = F(\theta)$$

- The function $\vartheta \mapsto Q_\gamma(\vartheta|\theta)$ is strictly convex (even if f is not convex !). If $F \circ T_\gamma(\theta) = F(\theta) = Q_\gamma(\theta|\theta)$, the unique minimum is $\theta = T_\gamma(\theta)$.
- Since the proximal operator is non-expansive and ∇f is Lipshitz, there exists $C < \infty$ such that

$$\|T_\gamma(\theta) - T_\gamma(\theta')\| \leq C\|\theta - \theta'\|$$

- F is a Lyapunov function for the proximal mapping.

Lyapunov function (convex case)

Global convergence result

Theorem ($\cdots$; AFM,14)

Assume A0-A1, and set $\gamma \in (0, 1/L]$. *If* $\{\theta_n,\ n \in \mathbb{N}\} \subset \mathcal{K}$ *where* $\mathcal{K}$ *is compact then:*

(i) $\mathcal{L} \neq \emptyset$, *the limiting points of* $\{\theta_n,\ n \in \mathbb{N}\}$ *belong to* $\mathcal{L} \cap \mathcal{K}$.

(ii) *there exists* $\theta_\star \in \mathcal{L} \cap \mathcal{K}$ *such that* $\lim_n F(\theta_n) = F(\theta_\star)$.

(iii) $\|\theta_{n+1} - \theta_n\| \to 0$.

- Can always enforce that $\{\theta_n,\ n \in \mathbb{N}\} \subset \mathcal{K}$ (project)
- If the level set $\mathcal{K} = \{F \leq \theta_0\}$ is compact, then $\{\theta_n,\ n \in \mathbb{N}\}_i n \mathcal{K}$.
- Using (iii): either $\{\theta_n,\ n \in \mathbb{N}\}$ converges or $\mathcal{L}$ is a continuum.
- Using (ii,iii): $\{\theta_n,\ n \in \mathbb{N}\}$ converges as soon as $\{\theta \in \mathcal{L} : F(\theta) = f_\star\}$is finite.

The convex case

Theorem ($\cdots$; AFM,14)

Assume A0-A1 and f convex; set $\gamma \in (0, 1/L]$. Assume that

(i) $\{\theta_n,\ n \in \mathbb{N}\} \subset \mathcal{K}$

(ii) *$\mathcal{L}$ is non-empty*

Then, there exists $\theta_\star \in \mathcal{L}$ such that $\lim_n \theta_n = \theta_\star$. In addition, $F(\theta_k) - F(\theta_\star)$ decreases to zero as $1/k$.

Wrap up

$$\textbf{(P)} \quad (\arg)\min_{\theta\in\Theta} \left\{ f(\theta) + g(\theta) \right\},$$

- the objective function always converge $\{F(\theta_n), n \geq 0\}$
- f is convex: then $\{\theta_n,\ n \in \mathbb{N}\}$ converges to $\theta_\star$, where $\theta_\star$ is a minimizer of F.
- $F(\theta_n) - F(\theta_\star) = O(1/n)$.

Perturbed proximal gradient

- Exact algorithm :

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

- Pertubed algorithm :

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

  where H_{n+1} is a proxy for $\nabla f(\theta_n)$.

- Problem find sufficient conditions on the perturbation $H_{n+1} - \nabla f(\theta_n)$ to preserve convergence.

Convergence (1/2)

- The Lyapunov condition is no longer satisfied:

$$\begin{aligned} &F(\theta_{n+1}) - F(\theta_n) \\ &= F(\mathrm{Prox}_{\gamma_{n+1}}(\theta_n + \gamma_{n+1} H_{n+1})) - F(\mathrm{Prox}_{\gamma_{n+1}}(\theta_n + \gamma_{n+1} \nabla f(\theta_n))) \\ &\qquad + F(\mathrm{Prox}_{\gamma_{n+1}}(\theta_n - \gamma_{n+1} \nabla f(\theta_n))) - F(\theta_n) \end{aligned}$$

- Under A0-A1, for any compact set $\mathcal{K}$

$$\lim_{n\to\infty} \left| F(\theta_{n+1}) - F(\mathrm{Prox}_{\gamma_{n+1}}(\theta_n - \gamma \nabla f(\theta_n))) \right| \mathbb{1}_{\theta_n \in \mathcal{K}} = 0$$

as soon as $\lim_n \{H_{n+1} - \nabla f(\theta_n)\} \mathbb{1}_{\theta_n \in \mathcal{K}} = 0$.

Convergence (2/2)

Résultat de convergence, cas général

Theorem (AFM,14)

Assume A0-A1, and set $\gamma \in (0, 1/L]$.
If $\mathcal{L} \neq \emptyset$, $\limsup_n \|\theta_n\| < \infty$ *et* $\lim_n \{H_{n+1} - \nabla f(\theta_n)\} = 0$, *then* $\{F(\theta_n), n \geq 0\}$ *converge to a connected component of* $F(\mathcal{L})$.
If the interior of $F(\mathcal{L}) = \emptyset$, *then there exists* $\theta_\star \in \mathcal{L}$ *such that*

(a) $\lim_n F(\theta_n) = F(\theta_\star)$,

(b) *the sequence* $\{\theta_n, n \geq 0\}$ *converges to* $\mathcal{L} \cap \{\theta : F(\theta) = F(\theta_\star)\}$.

A basic inequality

Lemma

Assume A0 and A1 and let $\gamma \in (0, 1/L]$. *Then for all* $\theta, \vartheta \in \Theta$,

$$\begin{aligned} -2\gamma\left(F(\mathrm{Prox}_\gamma(\theta)) - F(\vartheta)\right) \geq \left\|\mathrm{Prox}_\gamma(\theta) - \vartheta\right\|^2 \\ + 2\left\langle \mathrm{Prox}_\gamma(\theta) - \vartheta, \vartheta - \gamma\nabla f(\vartheta) - \theta\right\rangle \ . \end{aligned}$$

A basic inequality

Lemma

Assume A0 and A1 and let $\gamma \in (0, 1/L]$. Then for all $\theta, \vartheta \in \Theta$,

$$- 2\gamma \left(F(\mathrm{Prox}_\gamma(\theta)) - F(\vartheta)\right) \geq \left\|\mathrm{Prox}_\gamma(\theta) - \vartheta\right\|^2 + 2 \left\langle \mathrm{Prox}_\gamma(\theta) - \vartheta, \vartheta - \gamma \nabla f(\vartheta) - \theta \right\rangle \ .$$

Set $p = \mathrm{Prox}_{\gamma g}(\theta)$ and $\vartheta \in \Theta$.

$$f(p) - f(\vartheta) - \langle \nabla f(\vartheta), p - \vartheta \rangle \leq (2\gamma)^{-1} \|p - \vartheta\|^2 \ .$$

On the other hand, $\gamma^{-1}(\theta - p) \in \partial g(\theta)$. Therefore

$$g(p) + \gamma^{-1} \langle \theta - p, \vartheta - p \rangle \leq g(\vartheta) \ .$$

Combine.

The 3 points inequalities

Lemma

Assume A0 and A1, f closed convex and $\gamma \in (0, 1/L]$. Then for all $\theta, \vartheta \in \Theta$,

$$-2\gamma \left(F(\mathrm{Prox}_\gamma(\theta)) - F(\vartheta)\right) \geq \left\|\mathrm{Prox}_\gamma(\theta) - \vartheta\right\|^2 + 2\left\langle \mathrm{Prox}_\gamma(\theta) - \vartheta, \xi - \gamma \nabla f(\xi) - \theta \right\rangle - \left\|\vartheta - \xi\right\|^2 .$$

The 3 points inequalities

Lemma

Assume A0 and A1, f closed convex and $\gamma \in (0, 1/L]$. Then for all $\theta, \vartheta \in \Theta$,

$$- 2\gamma \left(F(\mathrm{Prox}_\gamma(\theta)) - F(\vartheta)\right) \geq \left\|\mathrm{Prox}_\gamma(\theta) - \vartheta\right\|^2 + 2 \left\langle \mathrm{Prox}_\gamma(\theta) - \vartheta, \xi - \gamma \nabla f(\xi) - \theta \right\rangle - \left\|\vartheta - \xi\right\|^2 .$$

$$\begin{aligned} f(p) - f(\vartheta) &\leq f(\xi) + \langle \nabla f(\xi), p - \xi \rangle + (2\gamma)^{-1} \|p - \xi\|^2 - f(\vartheta) \\ &\leq \{f(\xi) + \langle \nabla f(\xi), \vartheta - \xi \rangle - f(\vartheta)\} + \langle \nabla f(\xi), p - \vartheta \rangle + (2\gamma)^{-1} \|p - \xi\|^2 \end{aligned}$$

Combine with $g(p) + \gamma^{-1} \langle \theta - p, \vartheta - p \rangle \leq g(\vartheta)$ and conclude as before.

The basic inequality

Lemma

Assume A0-A1, f convex; Let $\theta_\star$ be a minimizer of F and set $F_\star \stackrel{\mathrm{def}}{=} F(\theta_\star)$. Then

$$\begin{aligned}\left\|\theta_{n+1}-\theta_\star\right\|^2 \leq \left\|\theta_n-\theta_\star\right\|^2 &- 2\gamma_{n+1}\left(F(\theta_n)-F_\star\right) \\ &+ 2\gamma_{n+1}^2\left\|\eta_{n+1}\right\|^2 - 2\gamma_{n+1}\left\langle T_{\gamma_{n+1}}(\theta_n)-\theta_\star, \eta_{n+1}\right\rangle \ .\end{aligned}$$

The basic inequality

Lemma

Assume A0-A1, f convex; Let $\theta_\star$ be a minimizer of F and set $F_\star \stackrel{\text{def}}{=} F(\theta_\star)$. Then

$$\|\theta_{n+1} - \theta_\star\|^2 \leq \|\theta_n - \theta_\star\|^2 - 2\gamma_{n+1}\left(F(\theta_n) - F_\star\right) + 2\gamma_{n+1}^2 \|\eta_{n+1}\|^2 - 2\gamma_{n+1}\left\langle T_{\gamma_{n+1}}(\theta_n) - \theta_\star, \eta_{n+1}\right\rangle \; .$$

The "3 points lemma" applied with $\theta \leftarrow \theta_n - \gamma_{n+1} H_{n+1}$, $\xi \leftarrow \theta_n$, $\vartheta \leftarrow \theta_\star$, $\gamma \leftarrow \gamma_{n+1}$

$$\|\theta_{n+1} - \theta_\star\|^2 \leq \|\theta_n - \theta_\star\|^2 - 2\gamma_{n+1}\left(F(\theta_n) - F_\star\right) - 2\gamma_{n+1}\left\langle \theta_{n+1} - \theta_\star, \eta_{n+1}\right\rangle \; .$$

Write $\theta_{n+1} - \theta_\star = \theta_{n+1} - T_{\gamma_{n+1}}(\theta_n) + T_{\gamma_{n+1}}(\theta_n) - \theta_\star$ and use that $\mathrm{Prox}_{\gamma g}$ is nonexpansive.

Convergence of the parameter

Theorem

Assume A0 and A1, f convex and $\gamma_n \in (0, 1/L]$ for any $n \geq 1$.

(i) *For any $\theta_\star$ in $\mathcal{L}$ and for any $n \geq m \geq 0$,*

$$\|\theta_{n+1} - \theta_\star\|^2 \leq \|\theta_m - \theta_\star\|^2 - 2 \sum_{k=m}^{n} \gamma_{k+1} \left\langle T_{\gamma_{k+1}}(\theta_k) - \theta_\star, \eta_{k+1} \right\rangle + 2 \sum_{k=m}^{n} \gamma_{k+1}^2 \|\eta_{k+1}\|^2 .$$

(ii) *Assume that for any $\theta_\star \in \mathcal{L}$, the two series in the RHS of the previous equation converge. Then, for any $\theta_\star$ in $\mathcal{L}$, $\lim_n \|\theta_n - \theta_\star\|$ exists. If in addition $\sum_n \gamma_n = +\infty$, then there exists $\theta_\infty \in \mathcal{L}$ such that $\lim_n \theta_n = \theta_\infty$.*

Convergence of the criterion

Theorem

Assume A0 and A1, f convex and $\gamma_n \in (0, 1/L]$ for any $n \geq 1$. For any non-negative sequence $\{a_n,\ n \in \mathbb{N}\}$, any minimizer $\theta_\star$ of F and any $n \geq 1$,

$$A_n^{-1} \sum_{k=1}^{n} a_k F(\theta_k) - F(\theta_\star) \leq B_n$$

where

$$B_n \stackrel{\text{def}}{=} \frac{1}{2A_n} \sum_{k=2}^{n} \left(\frac{a_k}{\gamma_k} - \frac{a_{k-1}}{\gamma_{k-1}} \right) \|\theta_{k-1} - \theta_\star\|^2 + \frac{a_1}{2\gamma_1 A_n} \|\theta_1 - \theta_\star\|^2 - \frac{1}{A_n} \sum_{k=1}^{n} a_k \left\{ \langle T_{\gamma_k}(\theta_{k-1}) - \theta_\star, \eta_k \rangle + \gamma_k \|\eta_k\|^2 \right\} .$$

Monte Carlo Approximation(1/2)

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

- Numerical integration - OK when the dimension is small (cubature or QMC)
- Importance Sampling

$$\int H_\theta(x)\, \pi_\theta(x)\mathrm{d}x = \int H_\theta(x)\, \frac{\pi_\theta(x)}{\pi_\star(x)}\, \pi_\star(x)\mathrm{d}x \approx \frac{1}{m_{n+1}} \sum_{k=1}^{m_{n+1}} H_{\theta_n}(X_k) \frac{\pi_{\theta_n}(X_k)}{\pi_\star(X_k)}$$

- MCMC

$$\int H_\theta(x)\, \pi_\theta(x)\mathrm{d}x \approx \frac{1}{m_{n+1}} \sum_{k=1}^{m_{n+1}} H_{\theta_n}(X_{n,k})$$

where $\{X_{n,k}, k \geq 0\}$ is a Markov chain with stationary distribution $\pi_{\theta_n}(\mathrm{d}x)$.

Monte Carlo Approximation (2/2)

$$\nabla f(\theta) = \int H_\theta(x)\, \pi_\theta(x) \mathrm{d}x \approx \frac{1}{m_{n+1}} \sum_{k=1}^{m_{n+1}} H_{\theta_n}(X_{n,k})$$

- Typically the approximation is biased $\mathbb{E}\left[H_{n+1} \,|\, \mathcal{F}_n \right] \neq \nabla f(\theta_n)$
- Nevertheless, in most cases

$$\left| \mathbb{E}\left[H_{n+1} - \nabla f(\theta_n) \,|\, \mathcal{F}_n \right] \right| \leq \frac{C_{\theta_n}}{m_{n+1}} \quad \mathbb{E}\left[\left\| H_{n+1} - \nabla f(\theta_n) \right\|^2 \middle| \mathcal{F}_n \right] \leq \frac{\tilde{C}_{\theta_n}}{m_{n+1}}$$

$\hookrightarrow$ How to choose the step sizes γ_n, and the size of the batches m_n ?

	$\gamma_n \equiv n^{-\mathsf{c}}$	$a_n \equiv n^{\mathsf{a}}$	$m_n \equiv n^{\mathsf{b}}$	Rate	MC
no bias	0	$(0, \infty)$	1	$1/n$	$1/\delta^2$
$(C_1 = 0)$	$[0, 1/2]$	$(-\mathsf{c}, +\infty)$	$1 - 2\mathsf{c}$	$1/n^{1-\mathsf{c}}$	$1/\delta^2$
	$[0, 1)$	$-\mathsf{c}$	$((1 - 2\mathsf{c})_+, \infty)$	$1/n^{1-\mathsf{c}}$	$1/\delta^{(1+\mathsf{b})/(1-\mathsf{c})}$
with bias	0	$(0, \infty)$	1	$1/n$	$1/\delta^2$
$(C_1 > 0)$	$[0, 1)$	$(0, \infty)$	$1 - \mathsf{c}$	$1/n^{1-\mathsf{c}}$	$1/\delta^{(2-\mathsf{c})/(1-\mathsf{c})}$
	$[0, 1)$	$-\mathsf{c}$	$(1 - \mathsf{c}, \infty)$	$1/n^{1-\mathsf{c}}$	$1/\delta^{(1+\mathsf{b})/(1-\mathsf{c})}$
det.	$[0, 1)$	$(-1, \infty)$	-	$1/n^{1-\mathsf{c}}$	-

Table: [Averaged Perturbed Proximal Gradient] Values of $(\mathsf{a}, \mathsf{b}, \mathsf{c})$ in order to reach the rate of convergence Rate. The column MC reports the number of Monte Carlo samples in this strategy to reach a precision $\mathcal{E}_n = O(\delta)$. As a reference, the last row reports the rate when $\eta_n = 0$.

Perturbed FISTA

Let $\theta_0 \in \mathrm{Dom}(g)$, $\{t_n,\ n \in \mathbb{N}\}$ and $\{\gamma_n,\ n \in \mathbb{N}\}$ be positive sequences. For $n \geq 1$, given $(\theta_0, \ldots, \theta_n)$:

1. Compute
$$\vartheta_n = \theta_n + t_n^{-1}(t_{n-1} - 1)(\theta_n - \theta_{n-1}) \ .$$
where
$$t_{n+1} = \frac{1 + \sqrt{1 + 4t_n^2}}{2} .$$
2. Obtain H_{n+1} an approximation of $\nabla f(\vartheta_n)$, and set
$$\theta_{n+1} = \mathrm{Prox}_{\gamma_{n+1}} \left(\vartheta_n - \gamma_{n+1} H_{n+1} \right) \ .$$

Wrap-up

- Penalized log-likelihood

$$\min_{(\beta,\sigma)\in\mathbb{R}^p\times\mathbb{R}^+} f(\theta)+g(\theta)\;, \quad f(\theta)=-\ell(\theta)\;,$$

- g closed convex; $-\ell$ is gradient Lipshitz but non convex
- $\nabla f(\theta)=\int H_\theta(\mathbf{u})\pi_\theta(\mathbf{u})\mathrm{d}\mathbf{u}$ with

$$H_\theta(\mathbf{u}) = -\sum_{i=1}^{N}(Y_i - F(x_i'\beta+\sigma z_i'\mathbf{u}))\begin{bmatrix} x_i \\ z_i'\mathbf{u}\end{bmatrix}$$

MCMC approximation of the gradient

Data augmentation approach

$$\nabla \ell(\theta) = \int_{\mathbb{R}^q \times \mathbb{R}^N} H_\theta(\mathbf{u}) \tilde{\pi}_\theta(\mathbf{u}, \mathbf{w}) \; \mathrm{d}\mathbf{u}\mathrm{d}\mathbf{w},$$

$$\tilde{\pi}_\theta(\mathbf{u}, \mathbf{w}) = \left(\prod_{i=1}^N \bar{\pi}_{\mathrm{PG}} \left(w_i; x_i' \beta + \sigma z_i' \mathbf{u} \right) \right) \pi_\theta(\mathbf{u}) \mathrm{d}\mathbf{u}$$

where $\bar{\pi}_{\mathrm{PG}}$ is the Polya-Gamma density.

- Polson algorithm (2012) Given $(\mathbf{u}^t, \mathbf{w}^t)$,
  - (i) sample $\mathbf{u}^{t+1} \sim \mathcal{N}_q \left(\mu_\theta(\mathbf{w}^t); \Gamma_\theta(\mathbf{w}^t) \right)$
  - (ii) sample $\mathbf{w}^{t+1} \sim \prod_{i=1}^N \bar{\pi}_{\mathrm{PG}}(w_i; |\mathbf{X}_{i\cdot} \beta + \sigma (\mathbf{Z}\mathbf{u}^{t+1})_i|)$

$$\Gamma_\theta(\mathbf{w}) = \left(I + \sigma^2 \sum_{i=1}^N w_i \mathbf{Z}_{i\cdot}' \mathbf{Z}_{i\cdot} \right)^{-1}, \qquad \mu_\theta(\mathbf{w}) = \sigma \Gamma_\theta(\mathbf{w}) \sum_{i=1}^N \left((Y_i - 1/2) - w_i \mathbf{X}_{i\cdot} \beta \right) \mathbf{Z}_{\cdot i}.$$

Toy example (1/2)

- $N = 500$ observations; $p = 1000$ regressors.
- Correlated design $\mathbf{X}$: $\mathbf{X}_{\cdot,n+1} = 0.8\mathbf{X}_{\cdot,n} + \sqrt{1-0.8^2}\mathcal{N}_q(0,I)$.
- Moderate sparsity: approximately 20 non zeros coefficients β_{true} : .

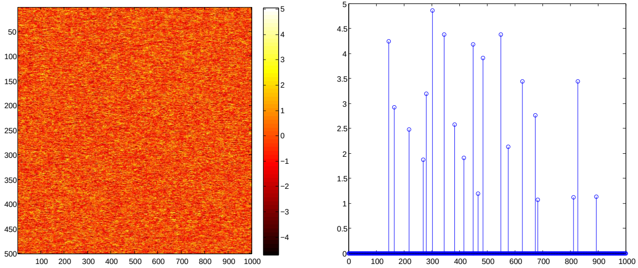


Figure: (left) Design matrix $\mathbf{X}$. (right) regression coefficient

Toy example (2/2)

- Random effect dimension: $q = 5$.
- $\mathbf{U} \sim \mathcal{N}_q(0, I)$.
- Binary factor $\mathbf{Z}$.

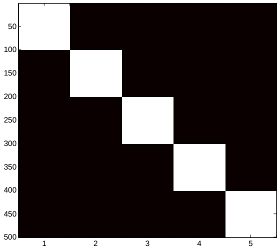
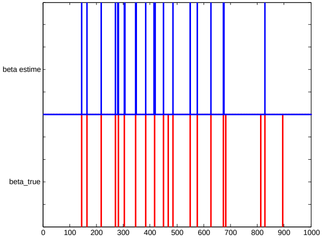


Figure: Factor loading $\mathbf{Z}$: 1 (white) et 0 (black)

Parameter convergence

- 150 iterations of the algorithms are performed with $\gamma_n = 5.10^{-2}$ and $m_n = 200 + n$.
- The support of β_∞ and β_{true} are displayed.



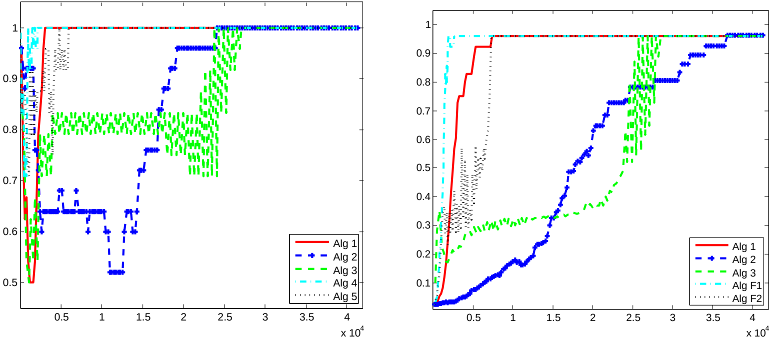
Parameter convergence

- We investigate the sensitivity and the precision defined as

$$\mathrm{SEN}_n = \frac{\sum_i \mathbb{1}_{|\beta_{n,i}|>0} 1_{|\beta_{\infty,i}|>0}}{\sum_i \mathbb{1}_{|\beta_{\infty,i}|>0}}$$

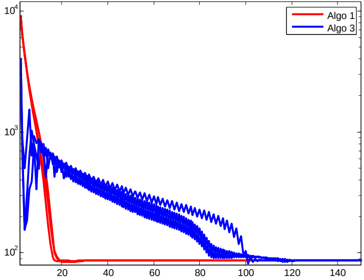
$$\mathrm{PRE}_n = \frac{\sum_i \mathbb{1}_{|\beta_{n,i}|>0} 1_{|\beta_{\infty,i}|>0}}{\sum_i \mathbb{1}_{|\beta_{n,i}|>0}}$$

- We consider 3 possible MC proximal gradient with $\gamma_n = \gamma = 0.005$, $m_n = 200 + n$ (Algo 1), $\gamma_n = \gamma = 0.001$, $m_n = 200 + n$ (Algo 2) and $\gamma_n = 0.05/\sqrt{n}$ and $m_n = 270 + \lceil \sqrt{n} \rceil$ (Algo 3).
- 2 MC FISTA with $\gamma_n = \gamma = 0.001$, $m_n = 45 + \lceil n^{3.1}/6000 \rceil$ (Algo F1); $\gamma_n = 0.005 \wedge (0.1/n)$, $m_n = 155 + \lceil n^{2.1}/100 \rceil$ (Algo F2).



The sensitivity SEN_n [left] and the precision PRE_n [right] along a path, versus the total number of Monte Carlo samples up to time n

Objective



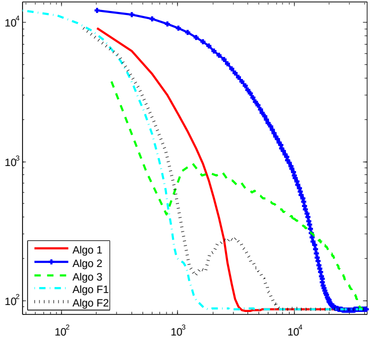


Figure: [left] $n \mapsto F(\theta_n)$ for several independent runs. [right] $\mathbb{E}\left[F(\theta_n)\right]$ versus the total number of Monte Carlo samples up to iteration n

Objective with averaging

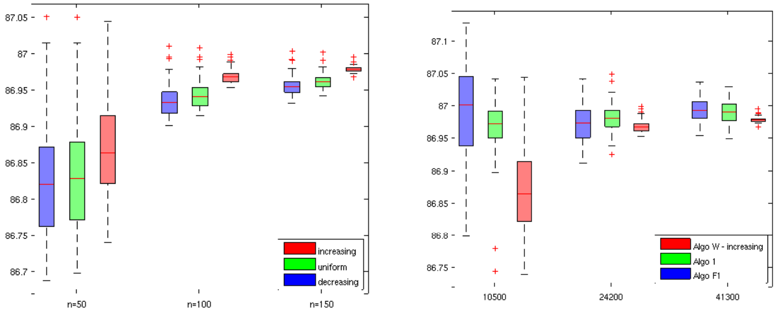


Figure: [left] Algo W: boxplot of $F(\bar{\theta}_n)$ for $n = 50, 100, 150$ with - from left to right - the decreasing, the uniform and the increasing weight sequence. [right] Boxplot of $F(\theta_n)$ and $F(\bar{\theta}_n)$ with n chosen such that the total number of Monte Carlo samples up to time n is about $10\,500, 24\,200, 41\,300$.

Take-home message

- Efficient and globally converging procedure for penalized likelihood inference in incomplete data models are available with convex sparsity-inducing penalty (provided that computing the proximal operator is easy)
- Minibatch algorithms combined to an averaging procedure allow to obtain numerically efficient algorithms.
- Thanks for your attention... and patience !