Stochastic Proximal Gradient Algorithm

Gersende FORT

LTCI, CNRS & Telecom ParisTech
Paris, France

Journée "Sélection dans les modèles mixtes"
Grenoble, Novembre 2014

Travail en collaboration avec Yves Atchadé (Univ. Michigan, USA) et Eric Moulines (Telecom ParisTech).

Plan

Régression logistique à effets aléatoires

- Observations : N observations binaires $\mathbf{Y} \in \{0,1\}^N$
- Modèle à données cachées : Conditionnellement à $\mathbf{U}$, pour tout $i = 1, \cdots, N$,

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

où

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

- Modèle paramétrique :

  $\mathbf{X} \in \mathbb{R}^{N\times p}$, $\mathbf{Z} \in \mathbb{R}^{N\times q}$, connues déterministes.
  $\mathbf{U} \in \mathbb{R}^q$ vecteur aléatoire.
  $\beta_{\text{true}} \in \mathbb{R}^p$,$\sigma_{\text{true}} > 0$ paramètres inconnus.

Maximum de vraisemblance pénalisée

- Log-vraisemblance : Sous l'hypothèse $\mathbf{U} \sim \mathcal{N}_q(0,I)$, en posant

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

la log-vraisemblance des observations $\mathbf{Y}$ en θ est donnée par

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

- Avec contraintes : Dans un contexte de grande dimension $N << p$, introduction de pénalités

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

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

Minimisation sous contraintes (1/2)

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

Quelle régularité sur la log-vraisemblance $\theta \mapsto \ell(\theta)$?

- Fonction :

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

avec

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

- Gradient :

$$\nabla \ell(\theta) = \int \left\{ \sum_{i=1}^{N} \{ Y_i - F(\mathbf{X}_{i\cdot}\beta + \sigma(\mathbf{Zu})_i) \} \begin{bmatrix} \mathbf{X}_{i\cdot} \\ (\mathbf{Zu})_i \end{bmatrix} \right\} \pi_\theta(\mathbf{u}) \mathbf{du}$$

avec

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

Minimisation sous contraintes (1/2)

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

Quelle régularité sur la log-vraisemblance $\theta \mapsto \ell(\theta)$?

- Fonction :

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

avec

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

- Hessien : pas de "signe" particulier mais

$$\exists L, \forall \theta, \theta' \qquad \|\nabla \ell(\theta) - \nabla \ell(\theta')\| \leq L \|\theta - \theta'\|$$

Minimisation sous contraintes (1/2)

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

Quelle régularité sur la log-vraisemblance $\theta \mapsto \ell(\theta)$?

- Fonction :

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

avec

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

$\hookrightarrow$ Fonction non explicite, définie comme une espérance
Fonction pas nécessairement convexe
Gradient non explicite, défini comme une espérance
Gradient lipschitzien

Minimisation sous contraintes (2/2)

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

Quelle régularité sur la fonction $\theta \mapsto g(\theta)$?

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

$$\tilde{g}_C(\theta) = \begin{cases} 0 & \text{si } \theta \in C \\ +\infty & \text{sinon} \end{cases} \qquad C \text{ convexe fermé}$$

$\hookrightarrow$ Fonction convexe,
Peut valoir $+\infty$ (en certains points)
Semi-continue inférieurement, pas nécessairement dérivable.

Conclusion

- Chercher le maximum de vraisemblance pénalisé dans ce modèle de regression logistique à effet mixte nécessite la résolution d'un problème d'optimisation sous contraintes.
- Plus précisément, résoudre

$$\text{argmin}_{\theta \in \Theta} \left(-\ell(\theta) + g(\theta) \right)$$

dans un contexte

- $-\ell$ pas nécessairement convexe, de gradient Lipschitz

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

- $-\ell, \nabla \ell$ ne sont pas explicites mais sont définis par des intégrales.
- g est convexe, pouvant valoir $+\infty$, semi-continue inférieurement

Plan

Problème considéré

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

sous les hypothèses

(A0) Θ espace euclidien de dimension finie, avec norme $\|\cdot\|$ et produit scalaire $\langle \cdot, \cdot \rangle$.

(A1) $g : \Theta \to (-\infty, +\infty]$ convexe, non identiquement égale à $+\infty$, et semi-continue inférieurement.
La fonction $\ell : \Theta \to \mathbb{R}$ est C^1 et il existe L tq pour tout θ,$\theta' \in \Theta$,

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

où $\nabla \ell$ désigne le gradient de ℓ.

Algorithme Gradient Proximal (1/3)

- Algorithme de type gradient pour des cas où la fonction objectif n'est pas différentiable.
- Peut se comprendre comme une approche "majoration-minimisation".

De la relation conséquence du gradient lipschitzien

$$-\ell(\vartheta) \leq -\ell(\theta) - \langle \nabla \ell(\theta), \vartheta - \theta \rangle + \frac{L}{2} \|\theta - \vartheta\|^2$$

on déduit pour tout $\gamma \in (0, 1/L]$

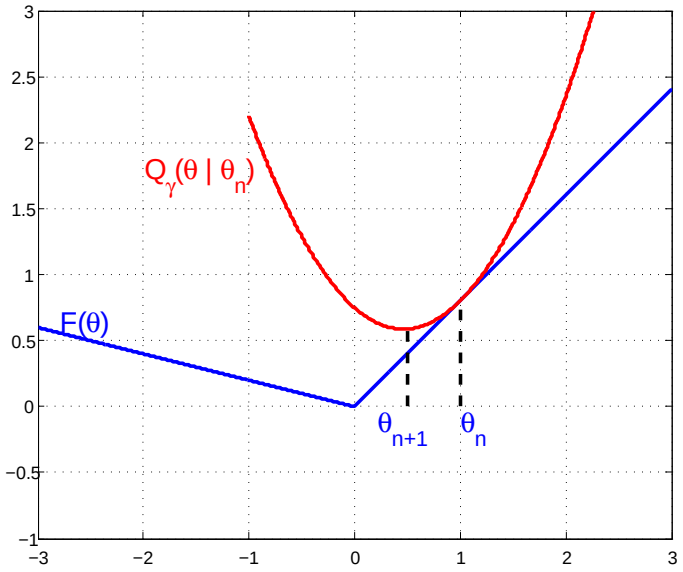
$$F(\vartheta) \leq -\ell(\theta) - \langle \nabla \ell(\theta), \vartheta - \theta \rangle + \frac{1}{2\gamma} \|\theta - \vartheta\|^2 + g(\vartheta)$$

Posons

$$\mathcal{Q}_\gamma(\vartheta|\theta) = -\ell(\theta) - \langle \nabla \ell(\theta), \vartheta - \theta \rangle + \frac{1}{2\gamma} \|\theta - \vartheta\|^2 + g(\vartheta)$$

Alors pour tout $\theta, \vartheta \in \Theta$ et $\gamma \in (0, 1/L]$

$$F(\vartheta) \leq \mathcal{Q}_\gamma(\vartheta|\theta) \qquad F(\theta) = \mathcal{Q}_\gamma(\theta|\theta)$$



$$F(\vartheta) \leq \mathcal{Q}_\gamma(\vartheta|\theta_n) \qquad F(\theta_n) = \mathcal{Q}_\gamma(\theta_n|\theta_n)$$

Algorithme Gradient Proximal (2/3)

D'où l'algorithme

$$\begin{aligned}\theta_{n+1} &= \min_{\vartheta\in\Theta} Q_\gamma\left(\vartheta|\theta_n\right)\\ &= \min_{\vartheta\in\Theta}\left\{g(\vartheta)+\frac{1}{2\gamma}\|\vartheta-(\theta_n+\gamma\nabla\ell(\theta_n))\|^2\right\}\end{aligned}$$

où $\gamma\in(0,1/L]$.

Algorithme Gradient Proximal:

$$\theta_{n+1}=\mathrm{Prox}_\gamma\left(\theta_n+\gamma\nabla\ell(\theta_n)\right)$$

avec

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

Algorithme Gradient Proximal (2/3)

D'où l'algorithme

$$\begin{aligned}\theta_{n+1} &= \min_{\vartheta\in\Theta} Q_\gamma\left(\vartheta|\theta_n\right)\\ &= \min_{\vartheta\in\Theta}\left\{g(\vartheta)+\frac{1}{2\gamma}\|\vartheta-(\theta_n+\gamma\nabla\ell(\theta_n))\|^2\right\}\end{aligned}$$

où $\gamma\in(0,1/L]$.

Algorithme Gradient Proximal: à pas décroissants

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

avec

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

Algorithme Gradient Proximal (3/3)

Forme explicite du Prox dans certains cas. Par exemple,

- Lorsque $g(\theta) = 0$, $\hookrightarrow$ gradient proximal $=$ algo. de gradient

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

Algorithme Gradient Proximal (3/3)

Forme explicite du Prox dans certains cas. Par exemple,

- Lorsque $g(\theta) = 0$, $\hookrightarrow$ gradient proximal $=$ algo. de gradient

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

- Lorsque $g(\theta) = \begin{cases} 0 & \text{si } \theta \in \mathsf{C} \\ +\infty & \text{sinon} \end{cases}$ pour un ensemble C fermé convexe,

$$\mathrm{Prox}_\gamma(\tau) = \min_{\theta \in \mathsf{C}} \|\tau - \theta\|^2 \qquad \text{i.e. projection orthogonale sur C}$$

$\hookrightarrow$ gradient proximal $=$ algo. de gradient projeté

$$\theta_{n+1} = \Pi_{\mathsf{C}} \left(\theta_n + \gamma_{n+1} \nabla \ell(\theta_n) \right)$$

Algorithme Gradient Proximal (3/3)

Forme explicite du Prox dans certains cas. Par exemple,

- Lorsque $g(\theta) = 0$, $\hookrightarrow$ gradient proximal $=$ algo. de gradient

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

- Lorsque $g(\theta) = \begin{cases} 0 & \text{si } \theta \in \mathsf{C} \\ +\infty & \text{sinon} \end{cases}$ pour un ensemble C fermé convexe,

$$\mathrm{Prox}_\gamma(\tau) = \min_{\theta \in \mathsf{C}} \|\tau - \theta\|^2 \qquad \text{i.e. projection orthogonale sur C}$$

$\hookrightarrow$ gradient proximal $=$ algo. de gradient projeté

$$\theta_{n+1} = \Pi_{\mathsf{C}} \left(\theta_n + \gamma_{n+1} \nabla \ell(\theta_n) \right)$$

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

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

$\hookrightarrow$ gradient proximal $=$ algo. de gradient seuillé

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

Convergence du Gradient-Proximal (1/4)

Quel ensemble limite $\mathcal{L}$?

$$\theta_{n+1} = \mathrm{Prox}_{\gamma}\left(\theta_n + \gamma \nabla \ell(\theta_n)\right)$$

Lemma

Sous A0 et A1:

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

Si de plus $-\ell$ est convexe

$\mathcal{L}$ *est l'ensemble des minimiseurs de* F

Convergence du Gradient-Proximal (2/4)

Fonction de Lyapunov associée à un mapping continue.

- Par définition de la fonction $\mathcal{Q}_\gamma(\theta|\vartheta)$, on a :

$$F(\theta_{n+1}) \leq \mathcal{Q}_\gamma(\theta_{n+1}|\theta_n) \leq \mathcal{Q}_\gamma(\theta_n|\theta_n) = F(\theta_n)$$

- Régularité de $\theta \mapsto \mathrm{Prox}_\gamma(\theta + \gamma\nabla\ell(\theta))$: il exists C tq

$$\|\mathrm{Prox}_\gamma(\theta + \gamma\nabla\ell(\theta)) - \mathrm{Prox}_\gamma(\theta' + \gamma\nabla\ell(\theta'))\| \leq C\|\theta - \theta'\|$$

Convergence du Gradient-Proximal (3/4)

Résultat général de convergence

Proposition (· · · ; AFM,14)

Sous A0-A1, pour $\gamma \in (0,1/L]$. *Si il existe* $\mathcal{K}$ *compact tel que* $\theta_n \in \mathcal{K}$ *pour tout* n, *alors:*

(i) $\mathcal{L}$ *est non vide, et les valeurs d'adhérence de* $\{\theta_n, n \geq 0\}$ *sont dans* $\mathcal{L} \cap \mathcal{K}$.

(ii) *Il existe* $\theta_\star \in \mathcal{L} \cap \mathcal{K}$ *tel que* $\lim_n F(\theta_n) = F(\theta_\star)$.

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

- La suite peut rester dans un compact par construction (par ex. si g est la projection sur $\mathcal{K}$)
- La suite reste dans un compact si $\{F \leq \theta_0\}$ est compact.

- Par (iii): soit $\{\theta_n, n \geq 0\}$ converge soit $\mathcal{L}$ est un continuum.
- Par (ii,iii): $\{\theta_n\}$ converge dès que $\{\theta \in \mathcal{L} : F(\theta) = f_\star\}$ est fini.

Convergence du Gradient-Proximal (4/4)

Résultat de convergence, dans le cas convexe

Proposition (· · · ; AFM,14)

Sous A0-A1 et $-\ell$ convexe; pour $\gamma \in (0,1/L]$. Supposons une des conditions équivalentes

(i) *il existe un compact $\mathcal{K}$ tel que $\theta_n \in \mathcal{K}$ pour tout n.*

(ii) *$\mathcal{L}$ est non vide.*

Alors il existe $\theta_\star \in \mathcal{L}$ tel que $\lim_n \theta_n = \theta_\star$.

Conclusion

$$\textbf{(P)} \quad (\arg)\min_{\theta\in\Theta} \{-\ell(\theta) + g(\theta)\},$$

- L'algorithme Gradient-Proximal construit $\{\theta_n, n \geq 0\}$ tq sous certaines hypothèses

  - (cas convexe et non convexe) convergence de la fonction objectif $\{F(\theta_n), n \geq 0\}$
  - (cas convexe) converge vers $\theta_\star$, $\theta_\star$ minimiseur de F.
  - (cas convexe) vitesses de convergence de $\{F(\theta_n), n \geq 0\}$ vers $F(\theta_\star)$ en $1/n$.

- Sa mise en oeuvre nécessite de

  - bien choisir les pas γ_n
  - savoir calculer Prox_γ
  - calculer $\nabla\ell(\theta_n)$

$\hookrightarrow$ dans l'application statistique considérée au début, $\nabla\ell(\theta_n)$ n'est pas calculable. Quelle alternative ?

Plan

Algorithme Gradient-Proximal perturbé

- Algorithme exact :

$$\theta_{n+1} = \text{Prox}_{\gamma_{n+1}} \left(\theta_n + \gamma_{n+1} \nabla \ell(\theta_n)\right)$$

- Algorithme perturbé :

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

où H_{n+1} est une approximation de $\nabla \ell(\theta_n)$.

Dans la suite,

- on étudie la suite $\{\theta_n\}$ produite par l'algorithme perturbé.

- On cherche des conditions suffisantes sur la perturbation $H_{n+1} - \nabla \ell(\theta_n)$ pour que l'algorithme perturbé converge.

Convergence de l'algorithme (1/2)

- La perturbation fait perdre la propriété de Lyapunov:

$$
\begin{aligned}
&F(\theta_{n+1}) - F(\theta_n) \\
&= F(\theta_{n+1}) - F(\mathrm{Prox}_{\gamma_{n+1}}(\theta_n + \gamma_{n+1}\nabla\ell(\theta_n))) \\
&\qquad + F(\mathrm{Prox}_{\gamma_{n+1}}(\theta_n + \gamma_{n+1}\nabla\ell(\theta_n))) - F(\theta_n) \\
&= F(\mathrm{Prox}_{\gamma_{n+1}}(\theta_n + \gamma_{n+1}H_{n+1})) - F(\mathrm{Prox}_{\gamma_{n+1}}(\theta_n + \gamma_{n+1}\nabla\ell(\theta_n))) + \underbrace{\cdots}_{\text{quantité négative}}
\end{aligned}
$$

- Néanmoins, on sait montrer que sous A0-A1, pour tout compact $\mathcal{K}$

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

dès que $\lim_n \{H_{n+1} - \nabla\ell(\theta_n)\}\mathbb{1}_{\theta_n\in\mathcal{K}} = 0$.

Convergence de l'algorithme (2/2)

Résultat de convergence, cas général

Théorème (AFM,14)

Sous A0-A1, pour $\gamma \in (0,1/L]$.
Si $\mathcal{L}$ *est non vide,* $\limsup_n \|\theta_n\|$ *est finie et* $\lim_n \{H_{n+1} - \nabla \ell(\theta_n)\} = 0$, *alors la suite* $\{F(\theta_n), n \geq 0\}$ *converge vers une composante connexe de* $F(\mathcal{L})$.
Si de plus $F(\mathcal{L})$ *est d'intérieur vide, il existe* $\theta_\star \in \mathcal{L}$ *tq*

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

(b) *la suite* $\{\theta_n, n \geq 0\}$ *converge vers* $\mathcal{L} \cap \{\theta : F(\theta) = F(\theta_\star)\}$.

- Quand $-\ell$ est convexe, $F(\mathcal{L})$ est d'intérieur vide. Si $-\ell$ fortement convexe, $\mathcal{L}$ est un singleton.

- Dans le cas non convexe: résultat nouveau.
- Dans le cas convexe : conditions plus faibles que travaux antérieurs.

Vitesses de convergence (1/3)

Quelle stratégie ? (a)

- Les méthodes Gdt-Prox proches de l'approximation stochastique

$$\begin{aligned} U_{n+1} &= U_n + \gamma_{n+1} H_{n+1} \\ &= U_n + \gamma_{n+1} \nabla h(U_n) + \gamma_{n+1} \left(H_{n+1} - \nabla h(U_n) \right) . \end{aligned}$$

$\hookrightarrow$ perturbation non asympt. nulle et pas γ_n décroissant; techniques d'averaging améliorent la vitesse : $1/\sqrt{n}$ au lieu de $\sqrt{\gamma_n}$

Vitesses de convergence (1/3)

Quelle stratégie ? (a)

- Les méthodes Gdt-Prox proches de l'approximation stochastique

$$\begin{aligned} U_{n+1} &= U_n + \gamma_{n+1} H_{n+1} \\ &= U_n + \gamma_{n+1} \nabla h(U_n) + \gamma_{n+1} \left(H_{n+1} - \nabla h(U_n) \right) . \end{aligned}$$

$\hookrightarrow$ perturbation non asympt. nulle et pas γ_n décroissant; techniques d'averaging améliorent la vitesse : $1/\sqrt{n}$ au lieu de $\sqrt{\gamma_n}$

- Averaging pour Gdt-Prox : en parallèle, calculer récursivement

$$\bar{\theta}_n = \frac{\sum_{k=1}^n a_k \theta_k}{\sum_{k=1}^n a_k}$$

où $\{a_n, n \geq 0\}$ suite positive.
$\hookrightarrow$ Quelle suite $\{a_n, n \geq 0\}$ et a-t-on le même gain en vitesse que pour l'approx. sto. ?

Vitesses de convergence (2/3)

Quelle stratégie ? (b)

- Accélération " à la Nesterov"

$$\theta_{n+1} = \mathrm{Prox}_{\gamma_{n+1}}(\vartheta_n + \gamma_{n+1}\nabla\ell(\vartheta_n))$$

où

$$\vartheta_n = \theta_n + \frac{(t_{n-1}-1)^2}{t_n}(\theta_n - \theta_{n-1})$$

Vitesses de convergence (2/3)

Quelle stratégie? (b)

- Accélération " à la Nesterov"

$$\theta_{n+1} = \mathrm{Prox}_{\gamma_{n+1}}(\vartheta_n + \gamma_{n+1}\nabla\ell(\vartheta_n))$$

où

$$\vartheta_n = \theta_n + \frac{(t_{n-1}-1)^2}{t_n}(\theta_n - \theta_{n-1})$$

$\hookrightarrow$ pour l'algorithme exact, cas convexe : vitesse en $1/n^2$ au lieu de $1/n$

- Nesterov pour Gdt-Prox "perturbé"

$$\theta_{n+1} = \mathrm{Prox}_{\gamma_{n+1}}(\vartheta_n + \gamma_{n+1}H_{n+1})$$

où H_{n+1} est une approximation de $\nabla\ell(\vartheta_n)$.
$\hookrightarrow$ A-t-on le même gain en vitesse que pour les algos Gdt-Prox non perturbés?

Vitesses de convergence (3/3)

Vitesse de convergence pour la suite averagée, cas convexe

Proposition (AFM,14)

Sous A0-A1 et $-\ell$ convexe, pour $\gamma_n \in (0,1/L]$. Pour tout minimiseur $\theta_\star$ de F,

$$\left(\sum_{j=1}^{n} a_j\right)\left(F(\bar{\theta}_n) - F(\theta_\star)\right)$$

$$\leq \frac{1}{2}\sum_{j=2}^{n}\left(\frac{a_j}{\gamma_j} - \frac{a_{j-1}}{\gamma_{j-1}}\right)\|\theta_{j-1} - \theta_\star\|^2 + \frac{a_1}{2\gamma_1}\|\theta_1 - \theta_\star\|^2$$

$$+ \sum_{j=1}^{n} a_j \left\langle \mathrm{Prox}_{\gamma_j}(\theta_{j-1} + \gamma_j \nabla\ell(\theta_{j-1})) - \theta_\star, \eta_j \right\rangle$$

$$+ \sum_{j=1}^{n} a_j \gamma_j \|\eta_j\|^2.$$

où $\eta_j = H_j - \nabla\ell(\theta_j)$.

Majoration analogue pour l'algo accéléré "à la Nesterov" voir AFM,14.

Plan

Approximation Monte Carlo (1/2)

Dans la suite, on considère le cas (plus général que l'apprentissage en ligne)

(A2)

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

Approximations possibles

- Intégration numérique - OK si espace de petite dimension
- Echantillonnage d'importance

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

- MCMC

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

où $\{X_{n,k}, k \geq 0\}$ est une chaîne de Markov de loi stationnaire $\pi_{\theta_n}(\mathrm{d}x)$.

Approximation Monte Carlo (2/2)

Dans cette partie, on considère l'approximation Monte Carlo

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

- Approximation biaisée

$$\mathbb{E}\left[H_{n+1} | \mathcal{F}_n\right] \neq \nabla \ell(\theta_n)$$

- mais pour certains échantillonneurs

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

$\hookrightarrow$ Quels choix pour le pas γ_n, l'effort computationnel m_n : régime *approx. sto.* ou régime *mini-batch*?
Les versions stochastiques des algorithmes averagés et à la Nesterov sont-elles aussi efficaces qu'en déterministe?

Convergence du Monte Carlo Gradient-Proximal

Les résultats

- de convergence
- de vitesse de convergence

énoncés précédemment restent valables : si les hypothèses sont vérifiées "presque-sûrement", les conclusions sont vraies "presque-sûrement".

En particulier, il existe des résultats dans la littérature sur la convergence des chaînes de Markov pour établir des conditions de la forme

$$\sum_n w_n \, (H_{n+1} - \nabla \ell(\theta_n)) \, \mathbb{1}_{\theta_n \in \mathcal{K}} < \infty \qquad \text{p.s.}$$

pour tout compact $\mathcal{K}$ et une suite déterministe $w_n \geq 0$.

Vitesse de convergence, cas convexe : algorithmes averagés (1/2)

Algorithme averagé

$$\bar{\theta}_n = \frac{\sum_{k=1}^n a_k \theta_k}{\sum_{k=1}^n a_k}$$

Proposition (AFM14)

Sous A0-A1-A2, $-\ell$ convexe et

$$\left|\mathbb{E}\left[H_{n+1} - \nabla\ell(\theta_n)|\mathcal{F}_n\right]\right| \leq \frac{C_1}{m_{n+1}} \qquad \mathbb{E}\left[\left\|H_{n+1} - \nabla\ell(\theta_n)\right\|^2\right] \leq \frac{C_2}{m_{n+1}}$$

$$\sup_n \|\theta_n - \theta_\star\| \leq B \text{ p.s.}$$

Dans le cas $a_n = C_{\mathsf{a}} n^{\mathsf{a}}$ $\mathsf{a} > -1$, $m_n = C_{\mathsf{b}} n^{\mathsf{b}}$ $\mathsf{b} \geq 0$, $\gamma_n = C_{\mathsf{c}} n^{-\mathsf{c}}$ $\mathsf{c} \geq 0$, *pour tout* $\theta_\star \in \mathcal{L}$,

$$\frac{\mathbb{E}\left[F(\bar{\theta}_n)\right] - F(\theta_\star)}{1+\mathsf{a}} \leq \frac{B^2}{2C_{\mathsf{c}}} \frac{1}{n^{1-\mathsf{c}}} + \frac{BC_1}{C_{\mathsf{b}}} \frac{1}{n^{\mathsf{a}+1}} \sum_{j=1}^n \frac{1}{j^{\mathsf{b}-\mathsf{a}}} + \frac{C_2 C_{\mathsf{c}}}{C_{\mathsf{b}}} \frac{1}{n^{\mathsf{a}+1}} \sum_{j=1}^n \frac{1}{j^{\mathsf{b}-\mathsf{a}+\mathsf{c}}}.$$

$\hookrightarrow$ quelle est la stratégie optimale pour (a,b,c) ?

Vitesse de convergence, cas convexe : algorithmes averagés (2/2)

	c	a	b	Rate $1/n^{\bullet}$	MC $1/\delta^{\bullet}$
Sans biais $(C_1 = 0)$	0	$(0,\infty)$	1	1	2
	$[0,1/2]$	$(-\mathsf{c}, +\infty)$	$1 - 2\mathsf{c}$	$1 - \mathsf{c}$	2
	$[0,1)$	$-\mathsf{c}$	$(1 - 2\mathsf{c},\infty) \cap [0,\infty)$	$1 - \mathsf{c}$	$\frac{1+\mathsf{b}}{1-\mathsf{c}}$
Avec biais $(C_1 > 0)$	0	$(0,\infty)$	1	1	2
	$[0,1)$	$(-\mathsf{c},\infty)$	$1 - \mathsf{c}$	$1 - \mathsf{c}$	$\frac{2-\mathsf{c}}{1-\mathsf{c}}$
	$[0,1)$	$-\mathsf{c}$	$(1 - \mathsf{c},\infty)$	$1 - \mathsf{c}$	$\frac{1+\mathsf{b}}{1-\mathsf{c}}$
$C_1 = C_2 = 0$	$[0,1)$	$(-1,\infty)$	-	$1 - \mathsf{c}$	-

Valeurs de (a,b,c) pour atteindre la vitesse de convergence `Rate`. La colonne `MC` donne le nombre d'échantillons MC nécessaires pour avoir $\mathbb{E}[F(\bar{\theta}_n)] - F(\theta_\star) = O(\delta)$. Comme référence, la dernière ligne donne la vitesse de l'algorithme exact (i.e. $H_{n+1} = \nabla \ell(\theta_n)$).

Vitesse de convergence, cas convexe : algorithmes averagés (2/2)

	c	a	b	Rate $1/n^\bullet$	MC $1/\delta^\bullet$
Sans biais $(C_1 = 0)$	0	$(0,\infty)$	1	1	2
	$[0,1/2]$	$(-\mathsf{c}, +\infty)$	$1-2\mathsf{c}$	$1-\mathsf{c}$	2
	$[0,1)$	$-\mathsf{c}$	$(1-2\mathsf{c},\infty)\cap[0,\infty)$	$1-\mathsf{c}$	$\frac{1+\mathsf{b}}{1-\mathsf{c}}$
Avec biais $(C_1 > 0)$	0	$(0,\infty)$	1	1	2
	$[0,1)$	$(-\mathsf{c},\infty)$	$1-\mathsf{c}$	$1-\mathsf{c}$	$\frac{2-\mathsf{c}}{1-\mathsf{c}}$
	$[0,1)$	$-\mathsf{c}$	$(1-\mathsf{c},\infty)$	$1-\mathsf{c}$	$\frac{1+\mathsf{b}}{1-\mathsf{c}}$
$C_1 = C_2 = 0$	$[0,1)$	$(-1,\infty)$	-	$1-\mathsf{c}$	-

Valeurs de (a,b,c) pour atteindre la vitesse de convergence `Rate`. La colonne `MC` donne le nombre d'échantillons MC nécessaires pour avoir $\mathbb{E}[F(\bar{\theta}_n)] - F(\theta_\star) = O(\delta)$. Comme référence, la dernière ligne donne la vitesse de l'algorithme exact (i.e. $H_{n+1} = \nabla\ell(\theta_n)$).

Meilleure stratégie : en fonction du nombre d'itérations, en fonction du coût de calcul :

pas fixe $\gamma_n = \gamma$ croissance linéaire MC : $m_n \sim n$

Vitesse de convergence, cas convexe : acceleration à la Nesterov (1/2)

Algorithme accéléré "à la Nesterov"

Proposition (AFM14)

Sous A0-A1-A2, $-\ell$ convexe et

$$|\mathbb{E}\left[H_{n+1} - \nabla\ell(\theta_n)|\mathcal{F}_n\right]| \leq \frac{C_1}{m_{n+1}} \qquad \mathbb{E}\left[\|H_{n+1} - \nabla\ell(\theta_n)\|^2\right] \leq \frac{C_2}{m_{n+1}}$$

$$\sup_n \|\theta_n - \theta_\star\| \leq B \text{ p.s.}$$

Dans le cas $m_n = C_{\mathsf{b}} n^{\mathsf{b}}$ $\mathsf{b} \geq 0$, $\gamma_n = C_{\mathsf{c}} n^{-\mathsf{c}}$ $\mathsf{c} \geq 0$, *pour tout* $\theta_\star \in \mathcal{L}$,

$$\mathbb{E}\left[F(\theta_{n+1})\right] - F(\theta_\star) \leq \frac{(n+1)^{\mathsf{c}}}{t_n^2}\left(\mathbb{E}\left[F(\theta_1)\right] - F(\theta_\star) + \frac{B^2}{2C_{\mathsf{c}}}\right)$$
$$+ 4BC_1C_{\mathsf{b}}\frac{(n+1)^{\mathsf{c}}}{t_n^2}\sum_{k=1}^{n}\frac{t_k^2}{(k+1)^{\mathsf{c}+\mathsf{b}}} + \frac{C_2(n+1)^{\mathsf{c}}}{t_n^2}\sum_{k=1}^{n}\frac{t_k^2}{(k+1)^{2\mathsf{c}+\mathsf{b}}}.$$

$\hookrightarrow$ quelle est la stratégie optimale pour (b,c) ?

Vitesse de convergence, cas convexe : accélération à la Nesterov (2/2)

	c	b	Rate $1/n^{\bullet}$	MC $1/\delta^{\bullet}$
Sans biais $(C_1 = 0)$	0	$(3,\infty)$	2	$(\mathsf{b}+1)/2$
	$[0,2)$	$(3-2\mathsf{c},+\infty)\cap[0,\infty)$	$2-\mathsf{c}$	$\frac{\mathsf{b}+1}{2-\mathsf{c}}$
Avec biais $(C_1 > 0)$	0	$(3,\infty)$	2	$(\mathsf{b}+1)/2$
	$[0,2)$	$(3-\mathsf{c},\infty)$	$2-\mathsf{c}$	$\frac{\mathsf{b}+1}{2-\mathsf{c}}$
$C_1 = C_2 = 0$	$[0,2)$	-	$2-\mathsf{c}$	-

Valeurs (b,c) pour atteinde la vitesse de convergence `Rate` quand $t_n = O(n^2)$. La colonne`MC` donne le nombre d'échantillons MC pour atteindre la précision $\mathbb{E}[F(\theta_n)] - F(\theta_\star) = O(\delta)$. La dernière ligne donne, pour référence, la vitesse de l'algorithme exact (i.e. $H_{n+1} = \nabla\ell(\theta_n)$).

Vitesse de convergence, cas convexe : accélération à la Nesterov (2/2)

	c	b	Rate $1/n^{\bullet}$	MC $1/\delta^{\bullet}$
Sans biais $(C_1 = 0)$	0	$(3,\infty)$	2	$(\mathsf{b}+1)/2$
	$[0,2)$	$(3-2\mathsf{c},+\infty)\cap[0,\infty)$	$2-\mathsf{c}$	$\frac{\mathsf{b}+1}{2-\mathsf{c}}$
Avec biais $(C_1 > 0)$	0	$(3,\infty)$	2	$(\mathsf{b}+1)/2$
	$[0,2)$	$(3-\mathsf{c},\infty)$	$2-\mathsf{c}$	$\frac{\mathsf{b}+1}{2-\mathsf{c}}$
$C_1 = C_2 = 0$	$[0,2)$	-	$2-\mathsf{c}$	-

Valeurs (b,c) pour atteinde la vitesse de convergence `Rate` quand $t_n = O(n^2)$. La colonne`MC` donne le nombre d'échantillons MC pour atteindre la précision $\mathbb{E}[F(\theta_n)] - F(\theta_\star) = O(\delta)$. La dernière ligne donne, pour référence, la vitesse de l'algorithme exact (i.e. $H_{n+1} = \nabla\ell(\theta_n)$).

Meilleure stratégie : en fonction du nombre d'itérations, en fonction du coût de calcul :

pas fixe $\gamma_n = \gamma$ croissance MC : $m_n \sim n^{3+\iota}$

Vitesse de convergence, cas convexe : averaging ou Nesterov ?

- On peut choisir (a,b,c) pour que les versions stochastiques atteignent les mêmes vitesses que les algorithmes exacts.
- En particulier, vitesse maximale atteinte par

  Averaging : $1/n$
  Nesterov : $1/n^2$

  que l'approximation soit avec ou sans biais.
- Mais ces stratégies optimales en vitesse n'ont pas le même coût de calcul:

  Averaging : $1/\delta^2$
  Nesterov : $1/\delta^{(b+1)/2}$ avec $b > 3$.
- Pour qu'il y ait convergence, il suffit

  Averaging : $0 \leq c < 1 \qquad 0 < b \leq b + c \leq 1 + a$
  Nesterov : $0 \leq c < 2 \qquad 0 < b \leq b + c < 3 - c$

Plan

Rappels

- Minimisation de la log-vraisemblance négative pénalisée

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

- lorsque g est convexe, sci; $-\ell$ est non convexe
- $\nabla\ell$ est Lipschitz et $\nabla\ell(\theta) = \int H_\theta(\mathbf{u})\pi_\theta(\mathbf{u})\mathrm{d}\mathbf{u}$ avec

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

Approximation Monte Carlo

MCMC de type Gibbs, basé sur une approche "augmentation de données"

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

en ayant posé

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

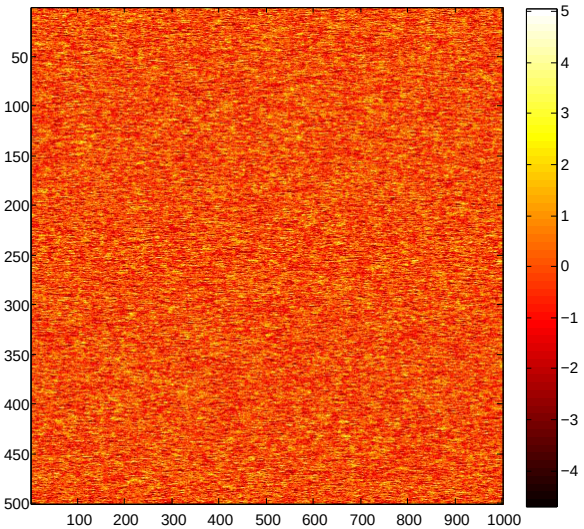
où $\bar{\pi}_{\mathrm{PG}}$ désigne la densité d'une loi Polya-Gamma.

- Algorithme de Polson et al. (2012) Etant donnée la valeur courante $(\mathbf{u}^t,\mathbf{w}^t)$,
  (i) simuler $\mathbf{u}^{t+1} \sim \mathcal{N}_q \left(\mu_\theta(\mathbf{w}^t) ; \Gamma_\theta(\mathbf{w}^t) \right)$
  (ii) simuler $\mathbf{w}^{t+1} \sim \prod_{i=1}^N \bar{\pi}_{\mathrm{PG}}(w_i ; |\mathbf{X}_{i\cdot}\beta + \sigma(\mathbf{Z}\mathbf{u}^{t+1})_i|)$

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

Données simulées (1/2)

- $N = 500$ observations; $p = 1000$ régresseurs.
- Matrice de design $\mathbf{X}$ construite par colonne selon un AR $\mathbf{X}_{\cdot,n+1} = 0.8\mathbf{X}_{\cdot,n} + \sqrt{1-0.8^2}\mathcal{N}_q(0,I)$.
- Coefficients de régression β_{true} : 2% de coeff non nuls.



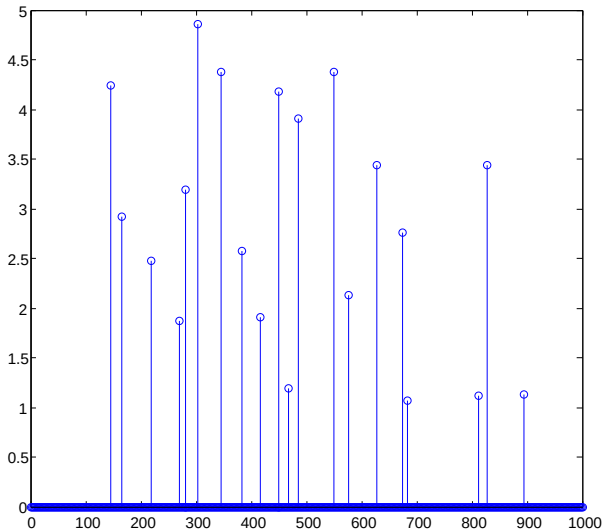


FIG.: (gauche) Matrice de design $\mathbf{X}$ de taille 500×1000. (droite) Coeff. de regression

Données simulées (2/2)

- $q = 5$ effets aléatoires.
- $\mathbf{U} \sim \mathcal{N}_q(0, I)$.
- Matrice $\mathbf{Z}$ de forme simple - pour rendre qq calculs explicites dans cette illustration numérique.

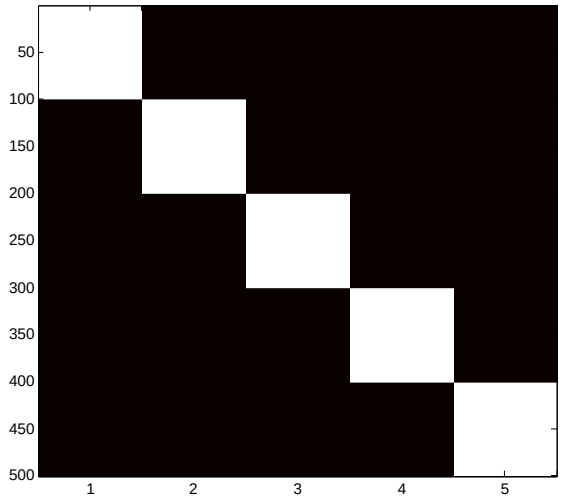


FIG.: Matrice $\mathbf{Z}$ binaire, de taille 500×5: 1 (en blanc) et 0 (en noir)

Convergence de β_n (1/2)

- On itère l'algorithme Gradient Proximal Stoch sur 150 itérations dans le cas $\gamma_n = 0.005$ et $m_n = 200 + n$. On relève la valeur limite β_∞.
- On trace le support de β_∞ et celui de β_{true} pourr comparaison (attention, $\beta_{\text{true}} \neq \text{argmin}(-\ell + g)$).

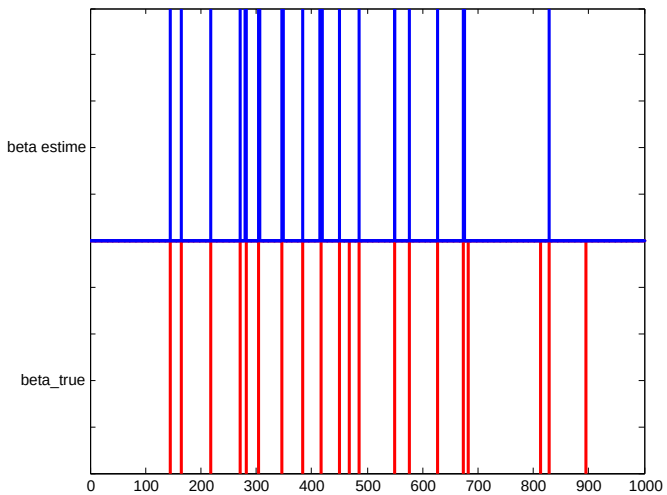


FIG.: Comparaison des supports : β_∞ (haut) et β_{true} (bas)

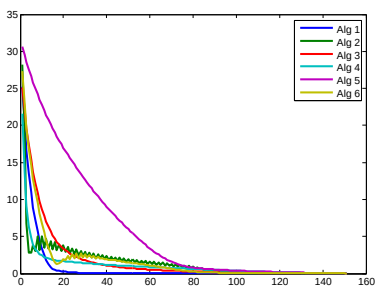
Convergence de β_n (2/2)

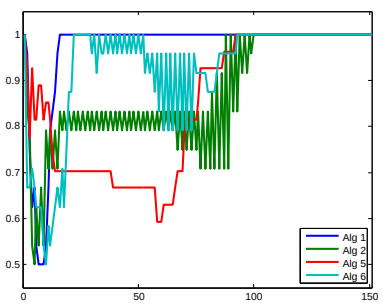
- On itère plusieurs algorithmes sur 150 itérations. On relève la valeur limite β_∞.

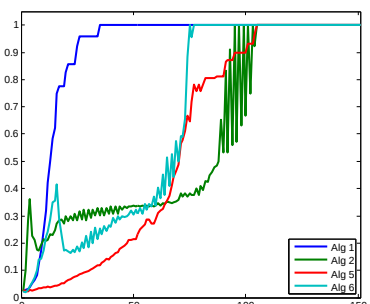
	Alg	γ_n	m_n	autres param.
1	GP-stoch	0.005	$200+n$	
2	GP-stoch	$0.05/\sqrt{n}$	$290+\sqrt{n}$	
3	Averagé	0.005	$200+n$	$a_n=\sqrt{n}$
4	Averagé	$0.05/\sqrt{n}$	$290+\sqrt{n}$	$a_n=\sqrt{n}$
5	Nesterov	0.001	$45+n^{3.1}/6000$	$t_n \sim n^2$
6	Nesterov	$0.005 \wedge 0.1/n$	$155+n^{2.1}/100$	$t_n \sim n^2$

- On trace *l'erreur relative*, *la sensitivity* et la *précision*

$$\mathcal{E}_n = \frac{\|\beta_n - \beta_\infty\|}{\|\beta_\infty\|}, \qquad \mathrm{SEN}_n = \frac{\sum_i 1_{|\beta_{n,i}|>0} 1_{|\beta_{\infty,i}|>0}}{\sum_i 1_{|\beta_{\infty,i}|>0}} \qquad \mathrm{PRE}_n = \frac{\sum_i 1_{|\beta_{n,i}|>0} 1_{|\beta_{\infty,i}|>0}}{\sum_i 1_{|\beta_{n,i}|>0}}$$



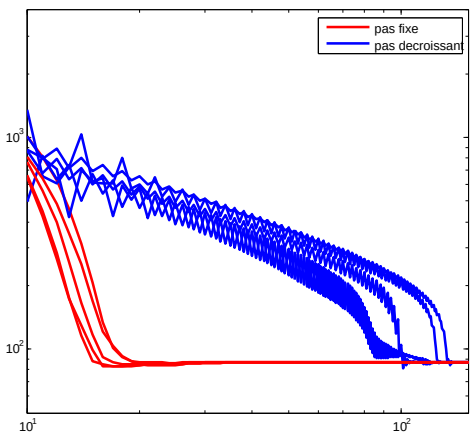




Convergence de $F(\theta_n)$

Comparaison de stratégies d'averaging. On itère plusieurs algorithmes sur 150 itérations. On trace une approximation de $n \mapsto \mathbb{E}[F(\bar{\theta}_n)] - F(\theta_\infty)$, calculée par Monte Carlo sur 50 runs indépendants. (on n'observe pas de valeurs significativement différentes pour $F(\theta_\infty)$ sur les 50 runs)

Alg	γ_n	m_n	autres param.
1	0.005	$200+n$	$a_n = \sqrt{n}$
2	0.005	$200+n$	$a_n = 1$
3	0.005	$200+n$	$a_n = n^{-0.1}$
4	$0.05/\sqrt{n}$	$290+\sqrt{n}$	$a_n = \sqrt{n}$
5	$0.05/\sqrt{n}$	$290+\sqrt{n}$	$a_n = 1$
6	$0.05/\sqrt{n}$	$290+\sqrt{n}$	$a_n = n^{-0.49}$



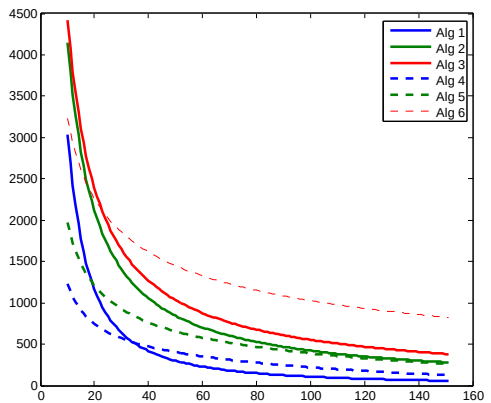


FIG.: (gauche) Trajectoires de l'algorithme Gdt Prox: à pas fixe et à pas décroissant. (droite) Estimation de $\mathbb{E}[F(\bar{\theta}_n)] - F(\theta_\star)$.

Plan

Conclusion de la présentation

- On dipose d'outils numériques pour calculer l'estimateur du maximum de vraisemblance pénalisée dans des modèles à données latentes, dans le cas de pénalités convexes engendrant des opérateurs proximaux calculables.
- Pour les algorithmes Gradient-Proximaux stochastiques

  - la stratégie "pas fixe + mini-batch" conduit aux meilleures vitesses de convergence : $1/n$ pour des stratégies d'averaging et $1/n^2$ pour des stratégies d'accélération à la Nesterov.

  - le coût computationnel est moindre pour les algorithmes averagés : $1/\delta^2$ pour atteindre une précision de δ dans l'estimation de $\min F$.