Algorithmes de gradient stochastiques moyennés

A. Godichon-Baggioni

Algorithme de gradient stochastique moyenné

Définition

Algorithme moyenné :

$$\overline{m}_n = \frac{1}{n+1} \sum_{k=0}^{n} m_k$$

où les m_k sont les estimateurs de gradient stochastique.

Ecriture récursive :

$$m_{n+1} = m_n - \gamma_{n+1} \nabla_h g\left(X_{n+1}, m_n\right)$$

$$\overline{m}_{n+1} = \overline{m}_n + \frac{1}{n+2}\left(m_{n+1} - \overline{m}_n\right).$$

avec $\gamma_n = c_\gamma n^{-\alpha}$ et $\alpha \in (1/2, 1)$.

Lemme de Toeplitz

Lemme

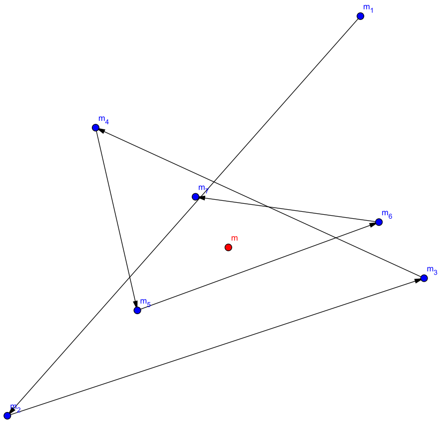
Soit (a_n) positive telle que $\sum_{n\geq 0} a_n = +\infty$ et X_n une suite de variables aléatoires convergeant presque sûrement vers X. Alors

$$\frac{1}{\sum_{k=0}^{n} a_k} \sum_{k=0}^{n} a_k X_k \xrightarrow[n\to+\infty]{p.s} X.$$

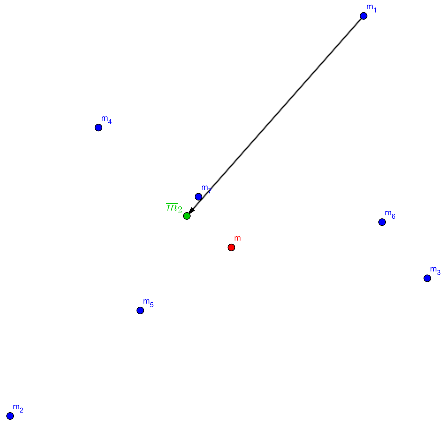
Application :

$$m_n \xrightarrow[n\to+\infty]{p.s} m \quad \Longrightarrow \quad \overline{m}_n \xrightarrow[n\to+\infty]{p.s} m$$

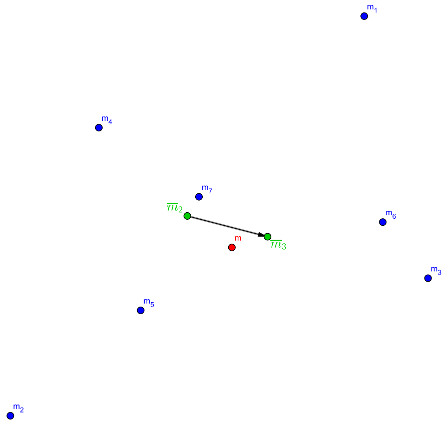
Comment ça marche ?



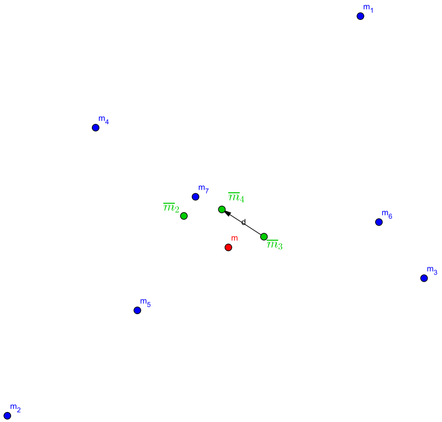
Comment ça marche ?



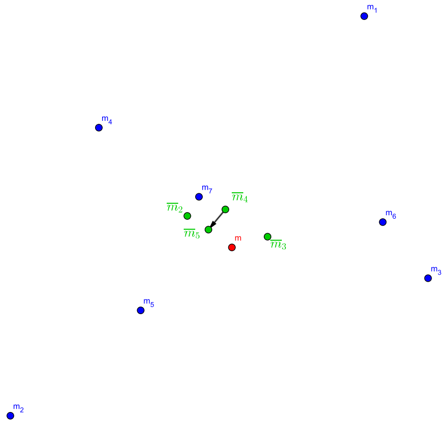
Comment ça marche ?



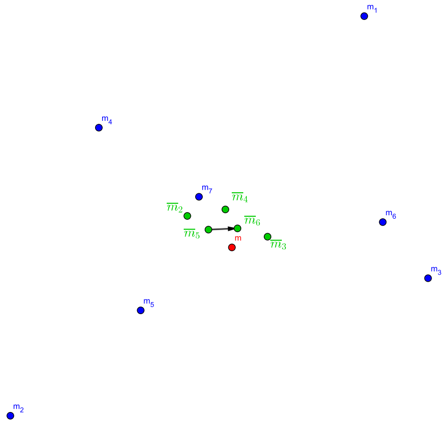
COMMENT ÇA MARCHE ?



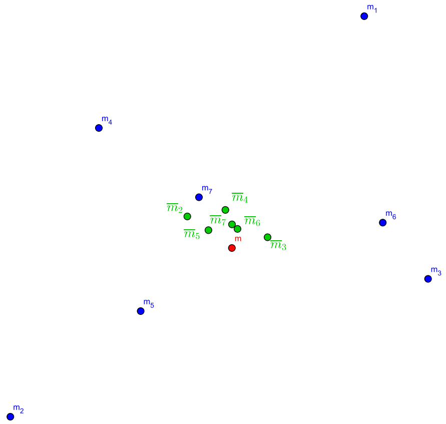
Comment ça marche ?



Comment ça marche ?



COMMENT ÇA MARCHE ?



Vitesse de convergence

Un corollaire du lemme de Toeplitz

Corollaire

Soit (X_n) une suite de variables aléatoires positives et (a_n) une suite positive telles que

$$X_n = o\left(a_n\right) \quad p.s.$$

Alors

$$\sum_{k=1}^{n} X_k = O\left(\sum_{k=1}^{n} a_k\right) \quad a.s.$$

Cadre (gradient stochastique)

(PS1) Il existe $\eta > \frac{1}{\alpha} - 1$ et $C_\eta \geq 0$ tels que

$$\mathbb{E}\left[\left\|\nabla_h g\left(X,h\right)\right\|^{2+2\eta}\right] \leq C_\eta\left(1 + \left\|h - m\right\|^{2+2\eta}\right)$$

(PS2) G est deux fois continûment différentiable et

$$\lambda_{\min} := \lambda_{\min}\left(\nabla^2 G(m)\right) > 0.$$

$$\textbf{(PS1) et (PS2)} \quad \Longrightarrow \quad \left\|m_n - m\right\|^2 = O\left(\frac{\ln n}{n^\alpha}\right) \quad p.s.$$

Cadre

(PS3) Il existe $\eta > 0$ et $C_\eta \geq 0$ t.q pour tout $h \in \mathcal{B}_\eta := \mathcal{B}(m, \eta)$,

$$\left\| \nabla G(h) - \nabla^2 G(m)(h-m) \right\| \leq C_\eta \left\| h - m \right\|^2$$

L'hypothèse **(PS3)** est vérifiée si $\nabla^2 G(.)$ est Lipschitz sur $\mathcal{B}_\eta$.

Vitesse de convergence

Théorème

*On suppose que les hypothèses **(PS1)** à **(PS3)** sont vérifiées. Alors, pour tout $\delta > 0$,*

$$\left\|\overline{m}_n - m\right\|^2 = o\left(\frac{(\ln n)^{1+\delta}}{n}\right) \quad p.s.$$

Preuve

La preuve repose sur le résultat suivant :

Théorème

Soit (ξ_k) *une suite de différences de martingale telle que* $\mathbb{E}\left[\left\|\xi_k\right\|^2 | \mathcal{F}_{k-1}\right] \leq C$*. Alors, pour tout* $\delta > 0$*,*

$$\left\|\sum_{k=1}^{n} \xi_k\right\|^2 = o\left(n(\ln n)^{1+\delta}\right) \quad p.s.$$

Efficacité asymptotique

(PS4) La fonction $\Sigma : \mathbb{R}^d \rightarrow \mathcal{M}_d(\mathbb{R})$ définie par

$$\Sigma(h) = \mathbb{E}\left[\nabla_h g(X,h) \nabla_h g(X,h)^T\right]$$

est continue en m.

Théorème

On suppose que les hypothèses ***(PS1)*** *à* ***(PS4)*** *sont vérifiées. Alors*

$$\sqrt{n}\left(\overline{m}_n - m\right) \xrightarrow[n\rightarrow+\infty]{\mathcal{L}} \mathcal{N}\left(0, H^{-1}\Sigma H^{-1}\right)$$

avec $H = \nabla^2 G(m)$ *et* $\Sigma = \Sigma(m)$*.*

Régression linéaire

L'algorithme

Algorithme moyenné :

$$\theta_{n+1} = \theta_n + \gamma_{n+1}\left(Y_{n+1} - X_{n+1}^T\theta_n\right)X_{n+1}$$
$$\overline{\theta}_{n+1} = \overline{\theta}_n + \frac{1}{n+2}\left(\theta_{n+1} - \overline{\theta}_n\right)$$

avec $\overline{\theta}_0 = \theta_0$.

Vitesse de convergence

Théorème

On suppose qu'il existe $\eta > \frac{1}{\alpha} - 1$ tel que X et ϵ admettent des moments d'ordre $4 + 4\eta$ et $2 + 2\eta$. Alors pour tout $\delta > 0$,

$$\left\| \overline{\theta}_n - \theta \right\|^2 = o\left(\frac{(\ln n)^{1+\delta}}{n} \right) p.s. \quad et \quad \sqrt{n}\left(\overline{\theta}_n - \theta \right) \xrightarrow[n \to +\infty]{\mathcal{L}} \mathcal{N}\left(0, \sigma^2 H^{-1}\right)$$

SIMULATIONS

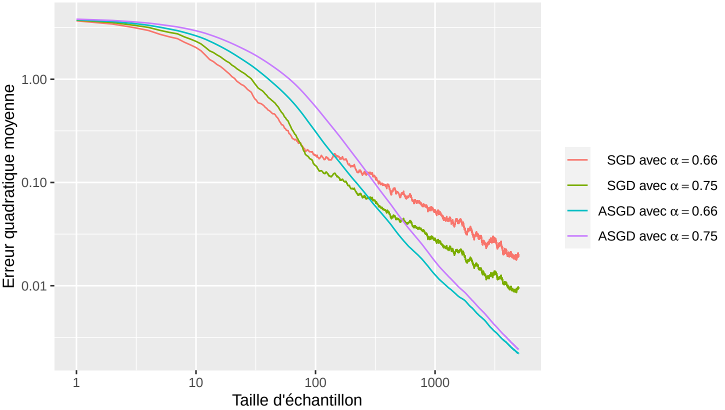


FIGURE – Evolution de l'erreur quadratique moyenne de l'estimateur de gradient θ_n (SGD) et de sa version moyennée $\overline{\theta}_n$ (ASGD) en fonction de la taille d'échantillon n dans le cadre de la régression linéaire.

TESTER H0 : $\theta = \theta_0$ "EN LIGNE"

Réécriture du TLC : Sous H0,

$$\sqrt{n}\frac{\left(\overline{\theta}_n - \theta_0\right)^T H \left(\overline{\theta}_n - \theta_0\right)}{\sigma^2} \xrightarrow[n\to+\infty]{\mathcal{L}} \chi^2_d$$

Application : Soit $\overline{H}_n$ et $\hat{\sigma}_n^2$ des estimateurs consistants. Alors

$$C_n := \sqrt{n}\frac{\left(\overline{\theta}_n - \theta_0\right)^T \overline{H}_n \left(\overline{\theta}_n - \theta_0\right)}{\hat{\sigma}_n^2} \xrightarrow[n\to+\infty]{\mathcal{L}} \chi^2_d$$

SIMULATIONS

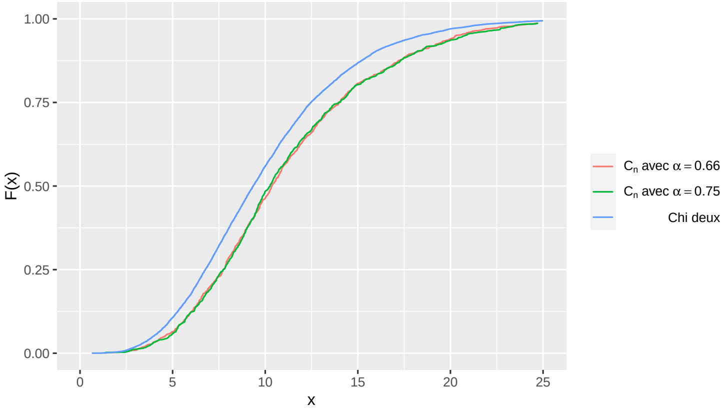


FIGURE – Comparaison de la fonction de répartition de C_n avec $n = 5000$, pour $\alpha = 0.66$ et $\alpha = 0.75$, et de celle d'une Chi 2 à 10 degrés de liberté dans le cadre du modèle linéaire.

Régression logistique

L'ALGORITHME

Algorithme moyenné :

$$\theta_{n+1} = \theta_n + \gamma_{n+1}\left(Y_{n+1} - \pi\left(X_{n+1}^T\theta_n\right)\right)X_{n+1}$$
$$\overline{\theta}_{n+1} = \overline{\theta}_n + \frac{1}{n+2}\left(\theta_{n+1} - \overline{\theta}_n\right)$$

avec $\overline{\theta}_0 = \theta_0$ et $\pi(x) = \frac{e^x}{1+e^x}$.

Vitesse de convergence

Théorème

On suppose que X admet un moment d'ordre 4*. Alors pour tout* $\delta > 0$*,*

$$\left\| \overline{\theta}_n - \theta \right\|^2 = o\left(\frac{(\ln n)^{1+\delta}}{n} \right) p.s. \quad et \quad \sqrt{n}\left(\overline{\theta}_n - \theta \right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, H^{-1}\right)$$

SIMULATIONS

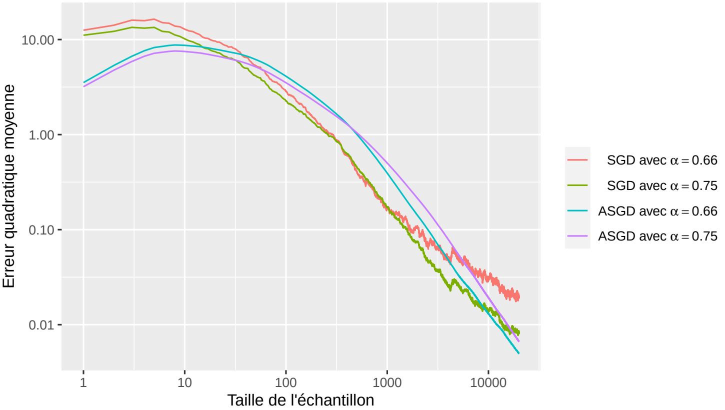


FIGURE – Evolution de l'erreur quadratique moyenne par rapport à la taille de l'échantillon des estimateurs de gradients θ_n (SGD) et de leurs versions moyennées $\overline{\theta}_n$ (ASGD) dans le cadre de la régression logistique.

TESTER H0 : $\theta = \theta_0$ "EN LIGNE"

Réécriture du TLC : Sous H0,

$$\sqrt{n}\left(\overline{\theta}_n - \theta_0\right)^T H \left(\overline{\theta}_n - \theta_0\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \chi^2_d$$

Application : Soit $\overline{H}_n$ un estimateur consistant de H. Alors

$$C_n := \sqrt{n}\left(\overline{\theta}_n - \theta_0\right)^T \overline{H}_n \left(\overline{\theta}_n - \theta_0\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \chi^2_d$$

SIMULATIONS

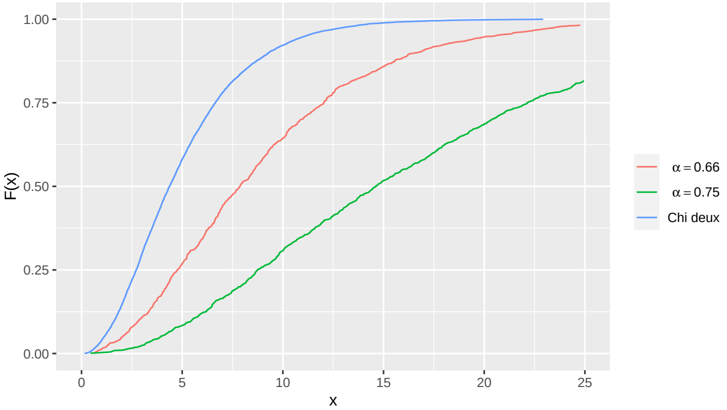


FIGURE – Comparaison de la fonction de répartition de C_n, avec $n = 20000$ et $\alpha = 0.66$ ou $\alpha = 0.75$, et de la fonction de répartition d'une Chi deux à 5 degrés de liberté dans le cadre de la régression logistique.

MOYENNÉ PONDÉRÉ

$$\overline{m}_n = \frac{1}{\sum_{k=1}^{n} \log(k+1)^w} \sum_{k=1}^{n} \log(k+1)^w m_k$$

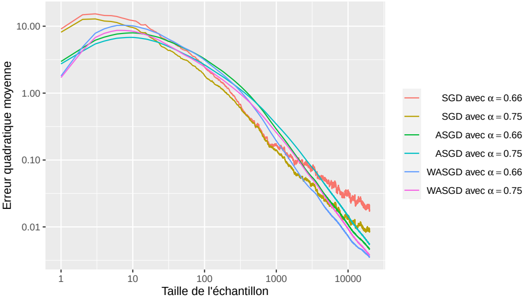


FIGURE – Evolution de l'erreur quadratique moyenne par rapport à la taille de l'échantillon des estimateurs de gradients θ_n (SGD) et de leurs versions moyennées $\overline{\theta}_n$ (ASGD) dans le cadre de la régression logistique.

EXERCICE

- Modèle linéaire :

  On prend $\theta = (-2, -1, 0, 1, 2)^T$, $X_i \sim \mathcal{N}(0, I_5)$ et $\epsilon_i \sim \mathcal{N}(0, 1)$.

  1. Générer un échantillon $(X_i, Y_i)_{i=1,\ldots,n}$ avec $n = 10000$.
  2. Ecrire une fonction qui ressorte l'ensemble de tous les estimateurs θ_i, $i = 0, \ldots, n$ et de leurs versions moyennées $\overline{\theta}_i$.
  3. Tracer les courbes des évolutions des erreurs quadratiques.
  4. Générer maintenant 50 échantillons et tracer les courbes des évolutions de des erreurs quadratiques moyennes.

- Faire de même pour la régression logistique.
- Revenir à l'exemple de la régression linéaire mais en prenant $X \sim \mathcal{N}(0, D)$ avec $D = \text{diag}\left(10^{-2}, 10^{-1}, 1, 10, 10^2\right)$. Regarder les évolutions des erreurs quadratiques moyennes pour les estimateurs de gradient stochastique et leurs versions moyennées.