

Algorithmes de gradient stochastiques moyennés

A. Godichon-Baggioni

Algorithme de gradient stochastique moyenné

DÉFINITION

Algorithme moyenné :

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

où les m_k sont les estimateurs de gradient stochastique.

Ecriture récursive :

$$\begin{aligned} m_{n+1} &= m_n - \gamma_{n+1} \nabla_h g(X_{n+1}, m_n) \\ \bar{m}_{n+1} &= \bar{m}_n + \frac{1}{n+2} (m_{n+1} - \bar{m}_n). \end{aligned}$$

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

LEMME DE TOEPLITZ

Lemme

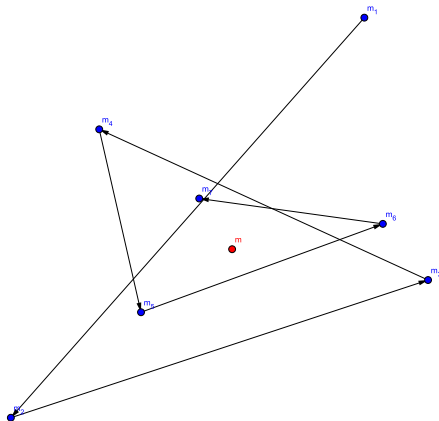
Soit (a_n) positive telle que $\sum_{n \geq 0} a_n = +\infty$ et X_n une suite de variables aléatoires convergeant presque sûrement vers X . Alors

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

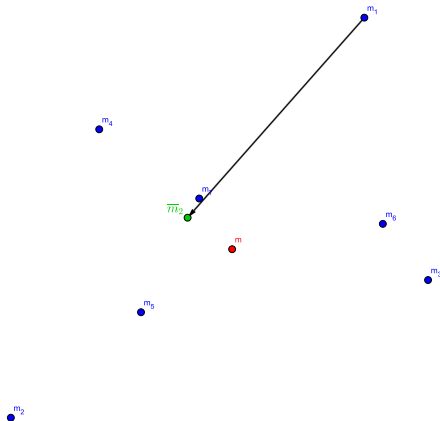
Application :

$$m_n \xrightarrow[n \rightarrow +\infty]{p.s.} m \implies \bar{m}_n \xrightarrow[n \rightarrow +\infty]{p.s.} m$$

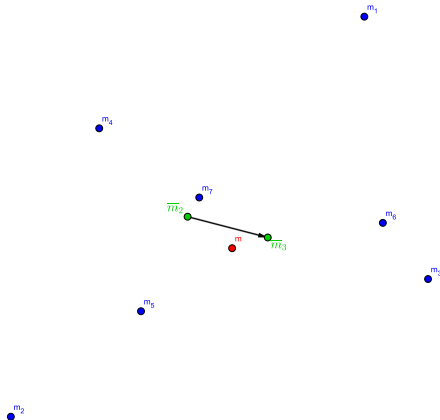
COMMENT ÇA MARCHE ?



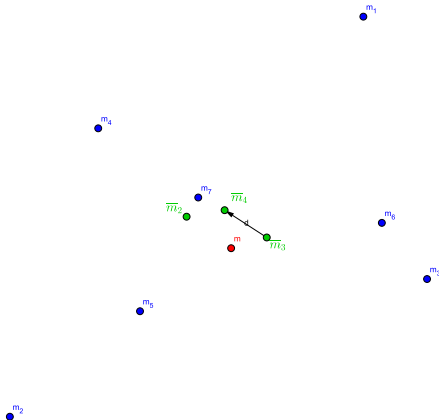
COMMENT ÇA MARCHE ?



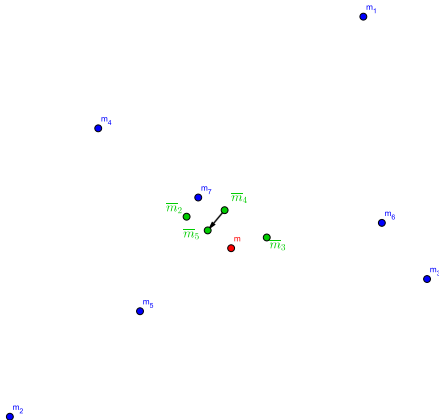
COMMENT ÇA MARCHE ?



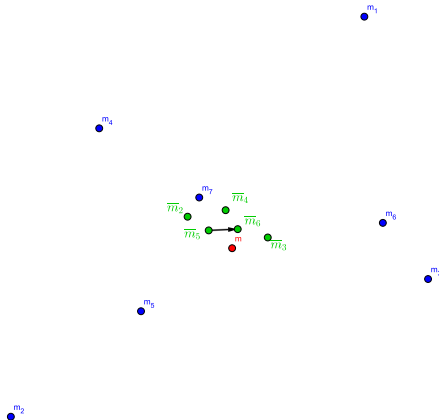
COMMENT ÇA MARCHE ?



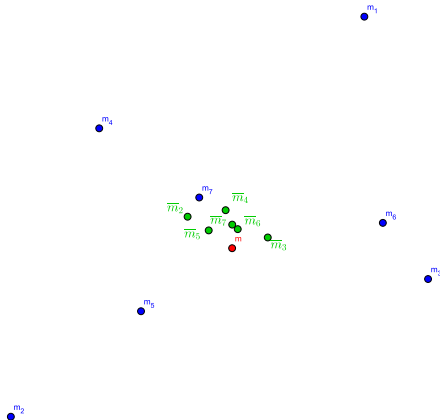
COMMENT ÇA MARCHE ?



COMMENT ÇA MARCHE ?



COMMENT ÇA MARCHE ?



Vitesse de convergence

CADRE (GRADIENT STOCHASTIQUE)

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

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

(PS2) G est deux fois continûment différentiable et

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

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

CADRE

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

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

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

VITESSE DE CONVERGENCE

Théorème

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

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

PREUVE

La preuve repose sur le résultat suivant :

Théorème

Soit (ξ_k) une suite de différences de martingale telle que

$$\mathbb{E} \left[\|\xi_k\|^2 \mathcal{F}_{k-1} \right] \leq C. \text{ Alors, pour tout } \delta > 0,$$

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

EFFICACITÉ ASYMPTOTIQUE

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

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

est continue en m .

Théorème

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

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

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

Régression linéaire

L'ALGORITHME

Algorithme moyenné :

$$\theta_{n+1} = \theta_n + \gamma_{n+1} (Y_{n+1} - X_{n+1}^T \theta_n) X_{n+1}$$

$$\bar{\theta}_{n+1} = \bar{\theta}_n + \frac{1}{n+2} (\theta_{n+1} - \bar{\theta}_n)$$

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

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 $\delta > 0$,

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

SIMULATIONS

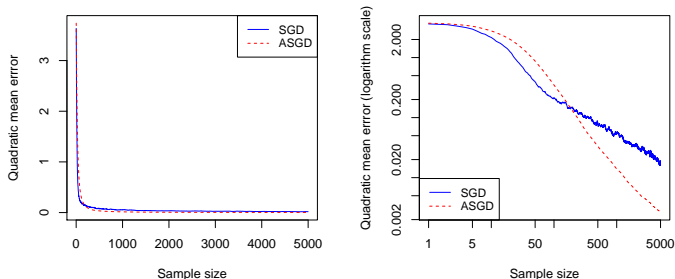


FIGURE – Evolution de l'erreur quadratique moyenne de l'estimateur de gradient θ_n (SGD) et de sa version moyennée $\bar{\theta}_n$ (ASGD) en fonction de la taille d'échantillon n dans le cadre de la régression linéaire.

TESTER $H_0 : \theta = \theta_0$ "EN LIGNE"

Réécriture du TLC : Sous H_0 ,

$$\sqrt{n} \frac{(\bar{\theta}_n - \theta_0)^T H (\bar{\theta}_n - \theta_0)}{\sigma^2} \xrightarrow[n \rightarrow +\infty]{\mathcal{L}} \chi_d^2$$

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

$$C_n := \sqrt{n} \frac{(\bar{\theta}_n - \theta_0)^T \bar{H}_n (\bar{\theta}_n - \theta_0)}{\hat{\sigma}_n^2} \xrightarrow[n \rightarrow +\infty]{\mathcal{L}} \chi_d^2$$

EXERCICE

1. Proposer un estimateur récursif de H .
2. Montrer sa consistance.
3. Proposer un estimateur récursif de σ^2 .
4. Montrer sa consistance et donner sa vitesse de convergence.
5. Donner sa normalité asymptotique.

CONSTRUCTION DE \bar{H}_n ET $\hat{\sigma}_n^2$

Ecriture directe :

$$\bar{H}_n = \frac{1}{n+1} \left(H_0 + \sum_{k=1}^n X_k X_k^T \right)$$
$$\hat{\sigma}_n^2 = \frac{1}{n+1} \sum_{k=1}^n (Y_k - X_k^T \bar{\theta}_{k-1})^2$$

Ecriture récursive :

$$\bar{H}_{n+1} = \bar{H}_n + \frac{1}{n+2} (X_{n+1} X_{n+1}^T - \bar{H}_n)$$
$$\hat{\sigma}_{n+1}^2 = \hat{\sigma}_n^2 + \frac{1}{n+2} \left((Y_{n+1} - X_{n+1}^T \bar{\theta}_n)^2 - \hat{\sigma}_n^2 \right)$$

SIMULATIONS

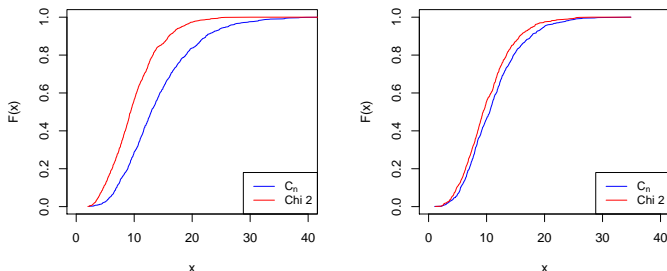


FIGURE – Comparaison de la fonction de répartition de C_n , pour $n = 1000$ (à gauche) et $n = 5000$ (à droite), et de la fonction de répartition d'une Chi 2 à 10 degrés de liberté dans le cadre du modèle linéaire.

Régression logistique

L'ALGORITHME

Algorithme moyenné :

$$\theta_{n+1} = \theta_n + \gamma_{n+1} (Y_{n+1} - \pi(X_{n+1}^T \theta_n)) X_{n+1}$$

$$\bar{\theta}_{n+1} = \bar{\theta}_n + \frac{1}{n+2} (\theta_{n+1} - \bar{\theta}_n)$$

avec $\bar{\theta}_0 = \theta_0$ 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 $\delta > 0$,

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

SIMULATIONS

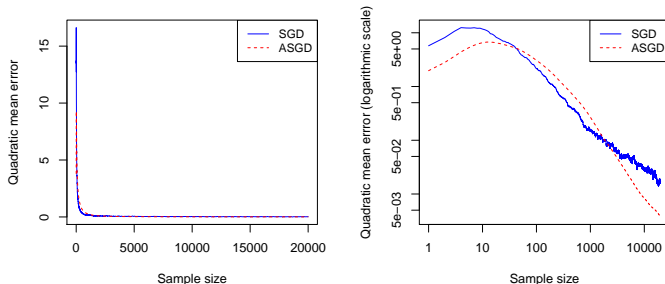


FIGURE – Evolution de l'erreur quadratique moyenne de l'estimateur de gradient θ_n (SGD) et de sa version moyennée $\bar{\theta}_n$ (ASGD) en fonction de la taille d'échantillon n dans le cadre de la régression logistique.

TESTER $H_0 : \theta = \theta_0$ "EN LIGNE"

Réécriture du TLC : Sous H_0 ,

$$\sqrt{n} (\bar{\theta}_n - \theta_0)^T H (\bar{\theta}_n - \theta_0) \xrightarrow[n \rightarrow +\infty]{\mathcal{L}} \chi_d^2$$

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

$$C_n := \sqrt{n} (\bar{\theta}_n - \theta_0)^T \bar{H}_n (\bar{\theta}_n - \theta_0) \xrightarrow[n \rightarrow +\infty]{\mathcal{L}} \chi_d^2$$

EXERCICE

1. Proposer un estimateur récursif de H .
2. Donner sa vitesse de convergence.

CONSTRUCTION DE \bar{H}_n

Rappel :

$$\nabla^2 G(\theta) = \mathbb{E} \left[\pi \left(X^T \theta \right) \left(1 - \pi \left(X^T \theta \right) \right) X X^T \right].$$

Ecriture directe :

$$\bar{H}_n = \frac{1}{n+1} \left(H_0 + \sum_{k=1}^n \pi \left(X_k^T \bar{\theta}_{k-1} \right) \left(1 - \pi \left(X_k^T \bar{\theta}_{k-1} \right) \right) X_k X_k^T \right)$$

Ecriture récursive :

$$\bar{H}_{n+1} = \bar{H}_n + \frac{1}{n+2} \left(\pi \left(X_{n+1}^T \bar{\theta}_n \right) \left(1 - \pi \left(X_{n+1}^T \bar{\theta}_n \right) \right) X_{n+1} X_{n+1}^T - \bar{H}_n \right)$$

SIMULATIONS

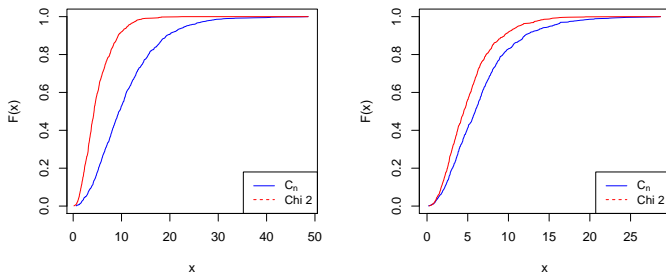


FIGURE – Comparaison de la fonction de répartition de C_n , pour $n = 1000$ (à gauche) et $n = 20000$ (à droite), et de la fonction de répartition d'une Chi 2 à 10 degrés de liberté dans le cadre de la régression logistique.

EXERCICE

- ▶ Sur un même graphique, tracer l'évolution de l'erreur quadratique moyenne de l'algorithme de gradient et de sa version moyennée (pour cela, on pourra générer 50 échantillons).
- ▶ Faire un tableau pour comparer les erreurs quadratiques moyennes pour $c_\gamma = 10^{-2}, 0.1, 1, 5, 100$ et $\alpha = 0.5, 0.66, 0.75, 1$.
- ▶ Faire de même pour la régression logistique avec $\theta = (-2, -1, 0, 1, 2)$ et $X \sim U[0, 1]$.
- ▶ Revenir à l'exemple de la régression linéaire mais en prenant $X \sim \mathcal{N}(0, D)$ avec $D = \text{diag}(10^{-2}, 10^{-1}, 1, 10, 10^2)$. Regarder les évolutions des erreurs quadratiques moyennes pour les estimateurs de gradient stochastique et leurs versions moyennées.