Algorithmes stochastiques

Vitesses de convergence des algorithmes de gradient stochastiques

A. Godichon-Baggioni

Convergence presque sûre

Approche directe

Théorème

On suppose que la fonction G est strictement convexe, i.e pour tout $h \neq m$,

$$\langle \nabla G(h), h - m \rangle > 0$$

et que l'hypothèse ***(PS0)*** *est vérifiée, i.e pour tout h*

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

Alors

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

Approche via le développement de Taylor

Théorème

On suppose que m est l'unique zéro du gradient et l'unique minimiseur de G . On suppose également qu'il existe C, C' tels que pour tout h,

$$\left\|\nabla^2 G(h)\right\|_{op} \leq C \qquad et \qquad \mathbb{E}\left[\left\|\nabla_h g\left(X,h\right)\right\|^2\right] \leq C'\left(1 + G(h) - G(m)\right).$$

Alors

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

Approche Lyapunov

Théorème

On suppose qu'il existe une fonction $V : \mathbb{R}^d \longrightarrow \mathbb{R}$ *vérifiant :*

- $V(m) = 0$ *et* $\forall h \neq m, V(h) \neq 0$
- *Il existe une constante C telle que pour tout h*

$$\mathbb{E}\left[\|\nabla_h g(X,h)\|^2\right] \leq C(1 + V(h))$$

- *Il existe* $\alpha > 0$ *tel que pour tout h*

$$\langle \nabla G(h), \nabla V(h) \rangle \geq \alpha V(h)$$

Alors

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

Application au modèle linéaire

Théorème

Si X admet un moment d'ordre 4,ϵ admet un moment d'ordre 2 et si $\mathbb{E}\left[XX^T\right]$ est définie positive, alors

$$\theta_n \xrightarrow[n\to+\infty]{p.s} \theta.$$

Application à la régression logistique

Théorème

On suppose que X admet un moment d'ordre 2 et que la Hessienne de G en θ est positive. Alors

$$\theta_n \xrightarrow[n\to+\infty]{p.s} \theta.$$

Vitesses de convergence presque sûre

CADRE

On considère une suite de pas de la forme $\gamma_n = c_\gamma n^{-\alpha}$ avec $c_\gamma > 0$ et $\alpha \in (1/2, 1)$. On suppose que les hypothèses suivantes sont vérifiées :

(PS1) Il existe $\eta > \frac{1}{\alpha} - 1$ et C_η tels que pour tout h,

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

(PS2) La fonction G est deux fois continument différentiable sur une voisinage de m et

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

Vitesse de convergence

Théorème

*On suppose que les hypothèses **(PS1)** et **(PS2)** sont vérifées. Alors*

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

Application au modèle linéaire

Théorème

Soit $\eta > 0$. On suppose que X admet un moment d'ordre $4 + 4\eta$ et que ϵ admet un moment d'ordre $2 + 2\eta$. De plus on suppose que $\mathbb{E}\left[XX^T\right]$ est positive. Alors

$$\|\theta_n - \theta\|^2 = O\left(\frac{\ln n}{n^\alpha}\right) \quad p.s.$$

Application au modèle linéaire

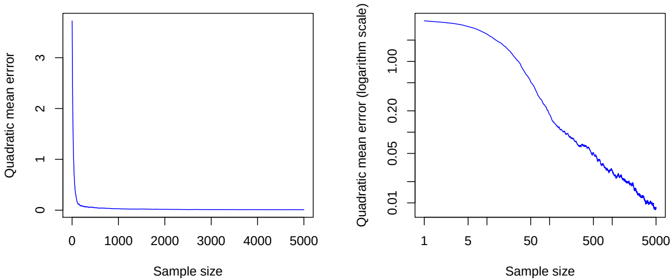

FIGURE – Evolution de l'erreur quadratique moyenne de θ_n en fonction de la taille d'échantillon n dans le cadre de la régression linéaire.

Application au modèle linéaire

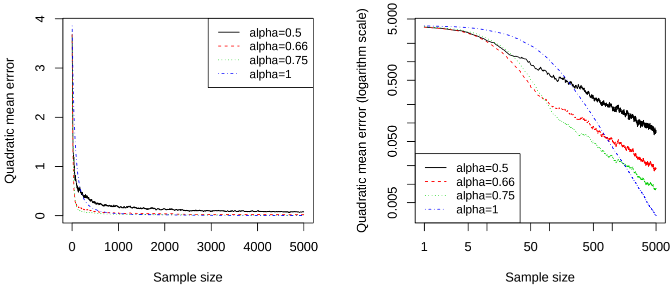

FIGURE – Evolution de l'erreur quadratique moyenne de θ_n en fonction de la taille d'échantillon n et du choix de α dans le cadre de la régression linéaire.

Application à la régression logistique

Théorème

On suppose qu'il existe $\eta > 0$ tel que X admette un moment d'ordre $2 + 2\eta$. On suppose également que $\nabla^2 G(\theta)$ est est positive. Alors

$$\|\theta_n - \theta\|^2 = O\left(\frac{\ln n}{n^\alpha}\right) \quad p.s.$$

Application à la régression logistique

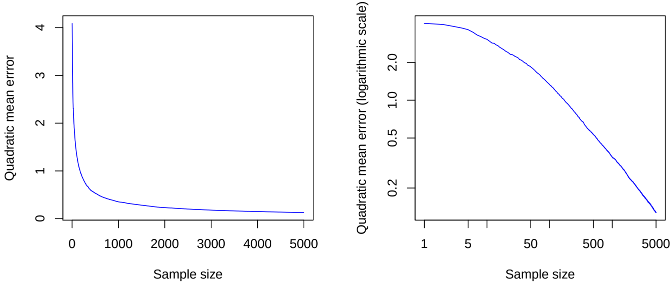

FIGURE – Evolution de l'erreur quadratique moyenne de θ_n en fonction de la taille de l'échantillon n dans le cadre de la régression logistique.

Application à la régression logistique

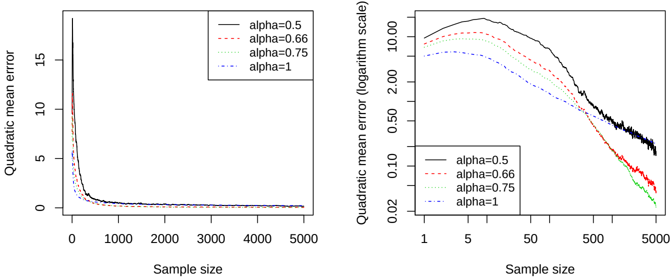

Figure – Evolution de l'erreur quadratique moyenne de θ_n en fonction de la taille de l'échantillon n et du choix du paramètre α dans le cadre de la régression logistique.

Remarques

En prenant $\gamma_n = c_\gamma n^{-1}$ et $c_\gamma > \frac{1}{2\lambda_{\min}}$, on peut montrer

$$\sqrt{n}\left(m_n - m\right) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}\left(0, \Sigma_{RM}\right)$$

avec

$$\Sigma_{RM} = \int_0^{+\infty} e^{-s\left(H - \frac{1}{2c_\gamma} I_d\right)} \Sigma e^{-s\left(H - \frac{1}{2c_\gamma} I_d\right)} ds$$

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

EXERCICE

Dans ce qui suit, on considère le modèle linéaire $Y = X^T\theta + \epsilon$ et $\theta = (-2, -1, 0, 1, 2)^T$, $X \sim \mathcal{N}(0, I_5)$ et $\epsilon \sim \mathcal{N}(0, 1)$.

- Générer un échantillon de taille $n = 5000$
- Ecrire un programme permettant d'obtenir l'estimateur du gradient stochastique
- Sur un même graphique, tracer l'évolution de l'erreur quadratique moyenne pour différents choix de α (pour cela, on pourra générer 50 échantillons).
- Comparer avec l'estimateur des moindres carrés.
- Prendre $\alpha = 1$ et regarder ce qu'il se passe pour $c_\gamma = 0.1, 0.5, 1, 2, 5$.