Introduction à l'optimisation stochastique

A. Godichon-Baggioni

I. Cadre

Cadre

Objectif: minimiser la fonction $G : \mathbb{R}^d \longrightarrow \mathbb{R}$ définie par

$$G(h) := \mathbb{E}\left[g(X, h)\right]$$

où X est une variable aléatoire à valeurs dans un espace $\mathcal{X}$.

Cadre: On supposera que G est convexe et on considèrera deux cas

- G est fortement convexe
- G est strictement convexe

FONCTIONS FORTEMENT CONVEXES

Moyenne d'une variable aléatoire: Soit X un vecteur aléatoire de $\mathbb{R}^d$. Sa moyenne minimise la fonction G définie par

$$G(h) = \frac{1}{2}\mathbb{E}\left[\|X-h\|^2 - \|X\|^2\right].$$

Modèle linéaire: On considère le modèle

$$Y = \theta^T X + \epsilon$$

avec X, ϵ des vecteurs aléatoires de $\mathbb{R}^d$ indépendants et $\mathbb{E}[\epsilon] = 0$. Le paramètre θ est un minimiseur de la fonction

$$G(h) = \frac{1}{2}\mathbb{E}\left[\left(Y - h^T X\right)^2\right].$$

FONCTIONS STRICTEMENT CONVEXES

Régression logistique: Soit (X, Y) vérifiant

$$Y|X \sim \mathbb{B}\left(\pi\left(\theta^T X\right)\right)$$

avec $\pi(x) = \frac{\exp(x)}{1+\exp(x)}$. Alors θ est un minimiseur de la fonction

$$G(h) = \mathbb{E}\left[\log\left(1 + \exp\left(h^T X\right)\right) - h^T XY\right]$$

Médiane d'une variable aléatoire: Soit X à valeurs dans $\mathbb{R}$. Sa médiane minimise la fonction

$$G(h) = \mathbb{E}\left[|X - h|\right]$$

Médiane géométrique: Soit X à valeurs dans $\mathbb{R}^d$. Sa médiane géométrique est un minimum de la fonction

$$G(h) = \mathbb{E}\left[\|X - h\| - \|X\|\right]$$

II. M-estimateurs

DÉFINITION

Soient $X_1, \ldots, X_n$ des variables aléatoires i.i.d de même loi que X. On considère la fonction empirique

$$G_n(h) = \frac{1}{n} \sum_{k=1}^{n} g\left(X_k, h\right)$$

Un M-estimateur est un minimiseur $\hat{m}_n$ de G_n.

Médiane d'une variable aléatoire: On a

$$G_n(h) = \frac{1}{n} \sum_{i=1}^{n} |X_i - h|$$

on retrouve la médiane empirique $\hat{m}_n = X_{\left(\lceil \frac{n}{2} \rceil\right)}$.

Modèle linéaire: On a

$$G_n(h) = \frac{1}{2n} \sum_{i=1}^{n} \left(Y_i - X_i^T h\right)^2$$

on obtient l'estimateur des moindres carrés $\hat{\theta}_n = \left(\mathbf{X}\mathbf{X}^T\right)^{-1} \mathbf{X}^T \mathbf{Y}$.

"CONTRE-EXEMPLES"

Régression logistique: On a

$$G_n(h) = \frac{1}{n}\sum_{i=1}^{n} \log\left(1 + \exp\left(h^T X_i\right)\right) - h^T X_i Y_i.$$

- Algorithme de gradient

Médiane géométrique: On a

$$G_n(h) = \frac{1}{n}\sum_{i=1}^{n} \|X_i - h\|$$

- Algorithme de Weiszfeld

ESTIMATION EN LIGNE

On considères des variables aléatoires $X_1, \ldots, X_n, X_{n+1}, \ldots$ arrivant de manière séquentielle.

Objectifs: Mettre à jours les estimateurs avec le moins de calculs possibles.

Estimation de la moyenne: On a

$$\overline{X}_{n+1} = \overline{X}_n + \frac{1}{n+1}\left(X_{n+1} - \overline{X}_n\right)$$

Estimateur de la variance: On a

$$\Sigma^2_{n+1} = \Sigma^2_n + \frac{1}{n+1}\left(X_{n+1}X^T_{n+1} - \Sigma^2_n\right)$$
$$S^2_{n+1} = \Sigma^2_{n+1} - \overline{X}_{n+1}\overline{X}^T_{n+1}.$$

III. Algorithmes de gradient stochastiques

ALGORITHME DE GRADIENT

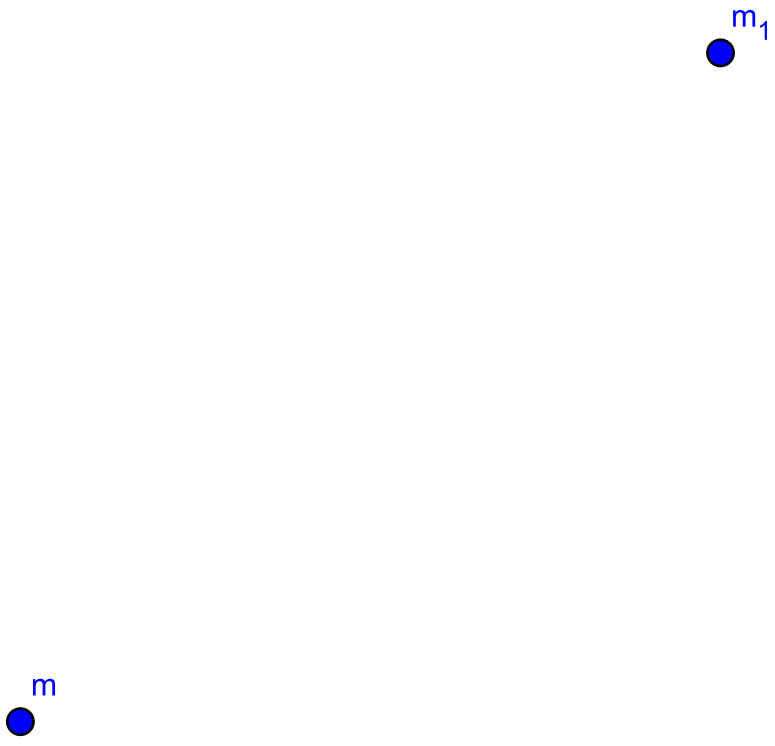
On cherche à minimiser la fonction convexe G définie par

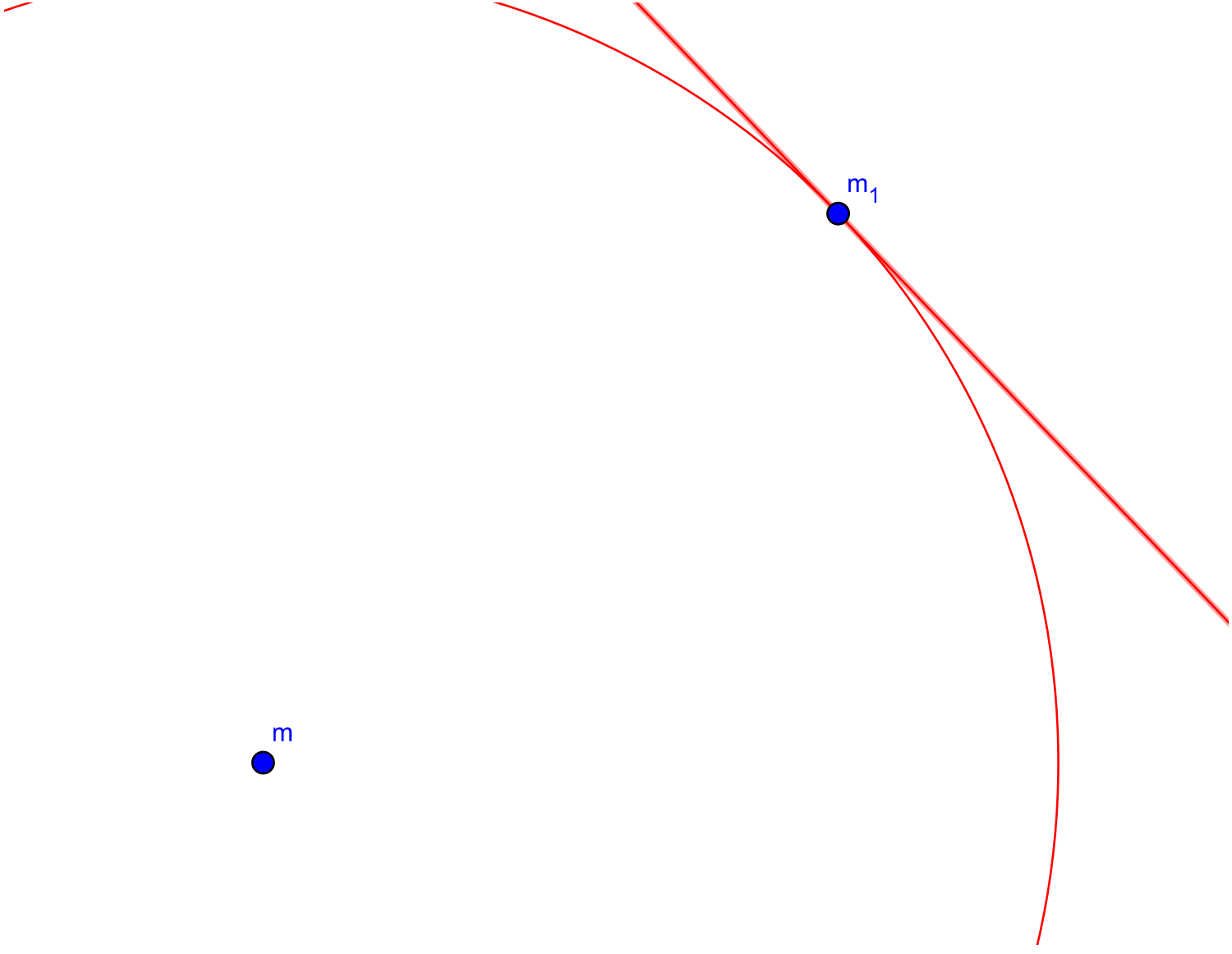
$$G(h) = \mathbb{E}\left[g\left(X, h\right)\right].$$

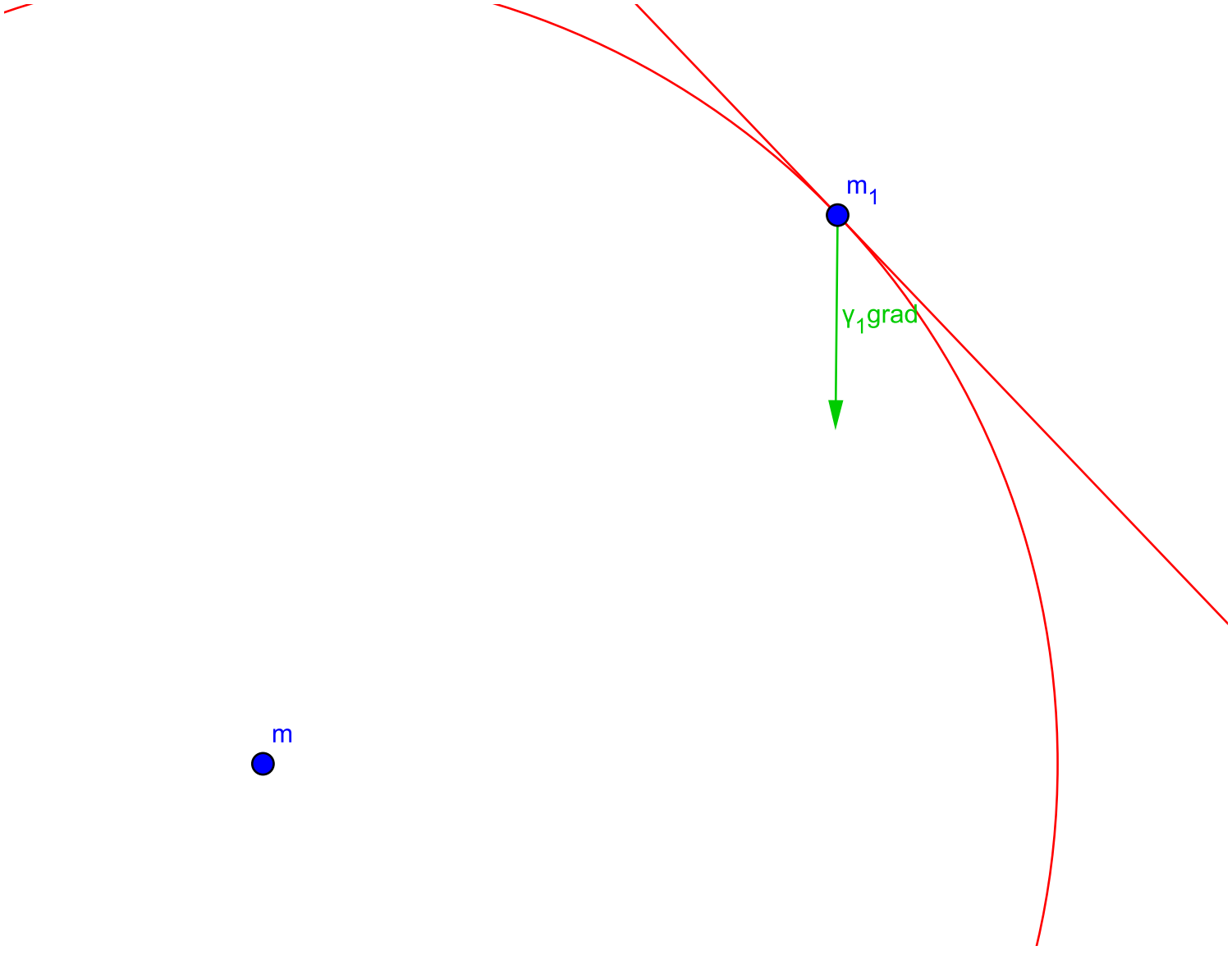
L'algorithme de gradient est défini de manière itérative par

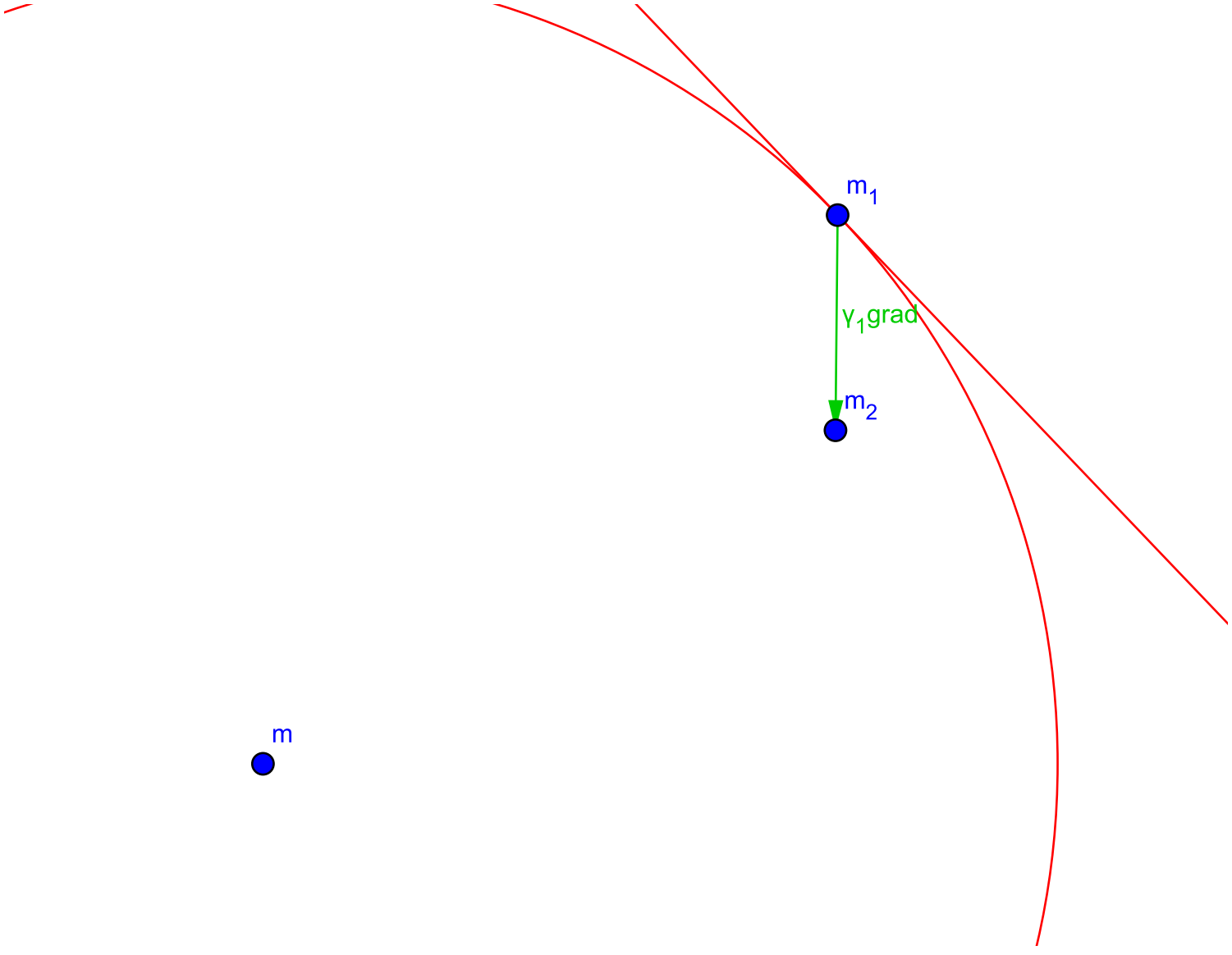
$$m_{t+1} = m_t - \gamma_t \nabla G\left(m_t\right)$$

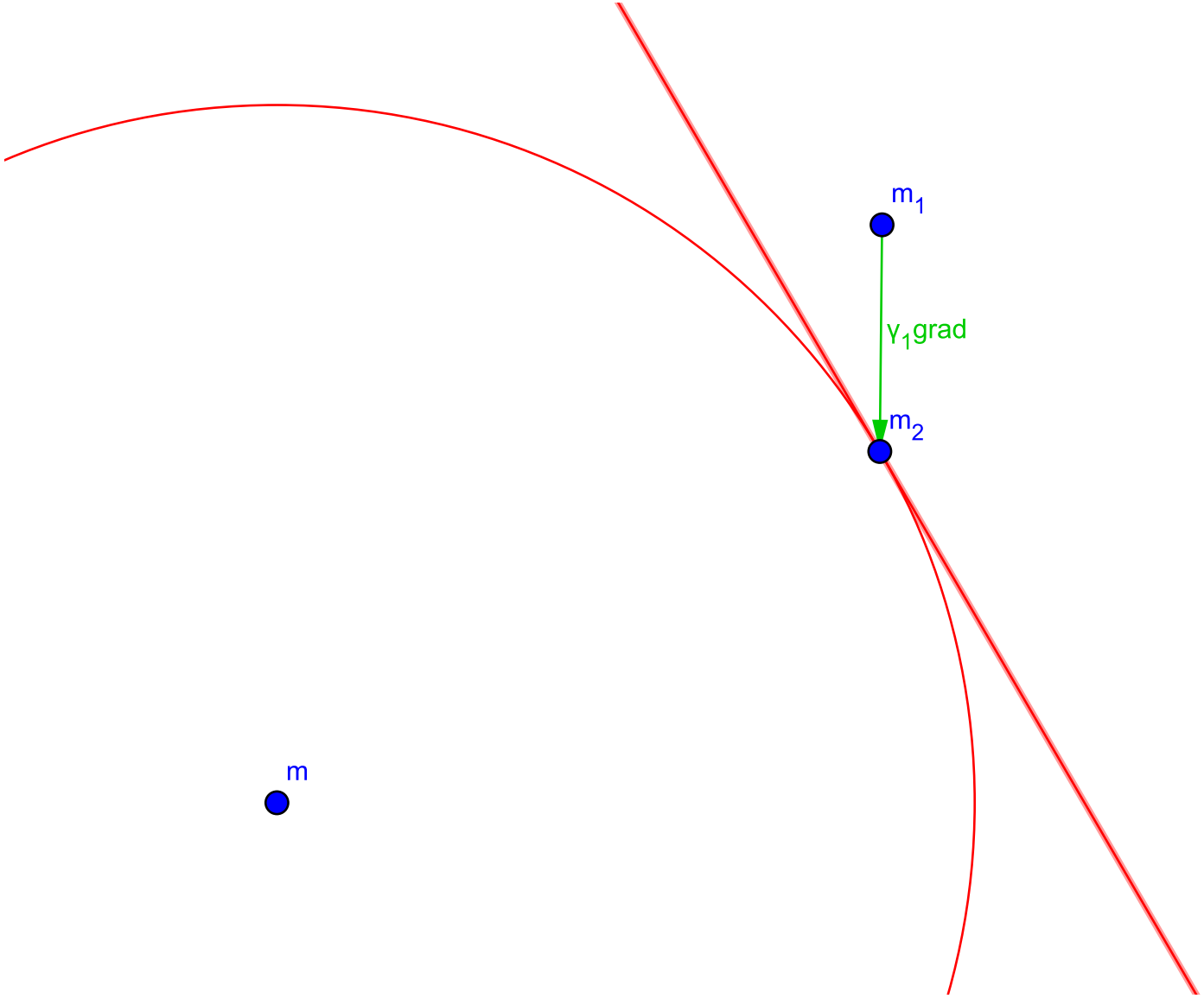
où η_t est une suite de pas positifs.

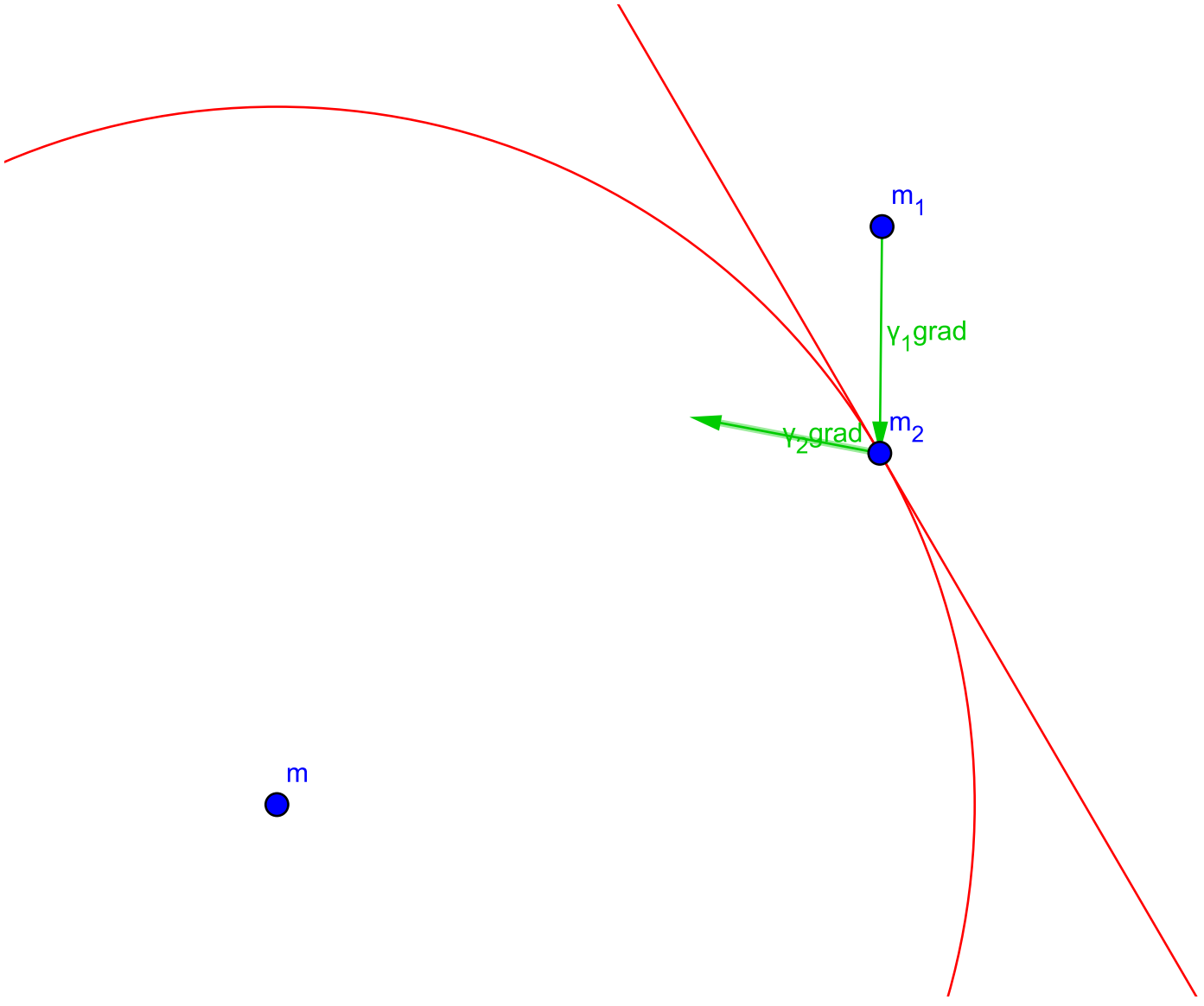

m1
m

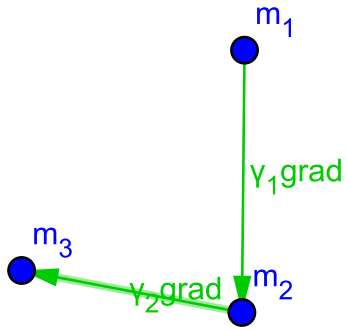

m1
m


m1
γ1grad
m

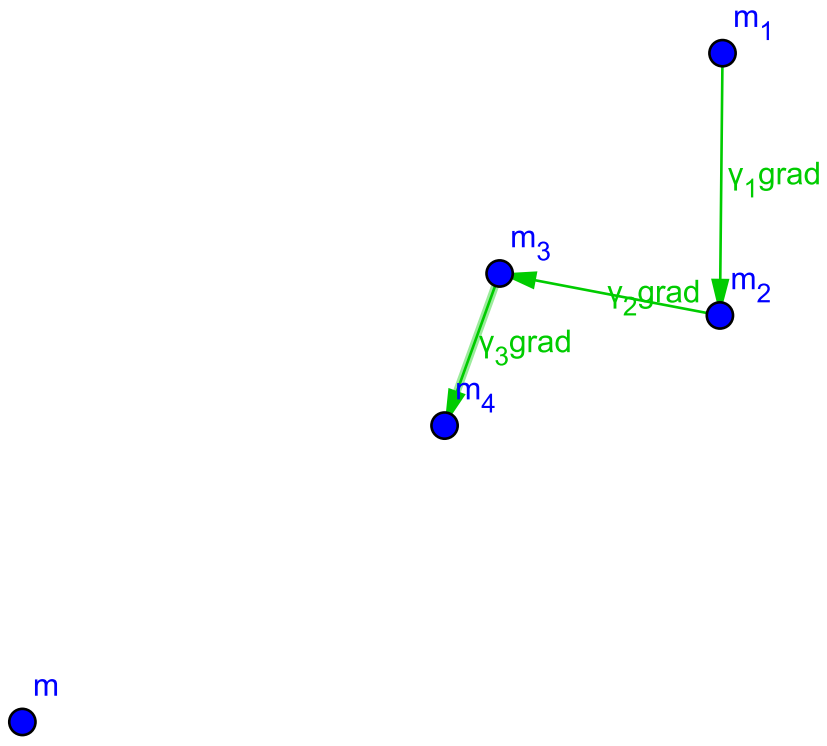

m1
γ1grad
m2
m

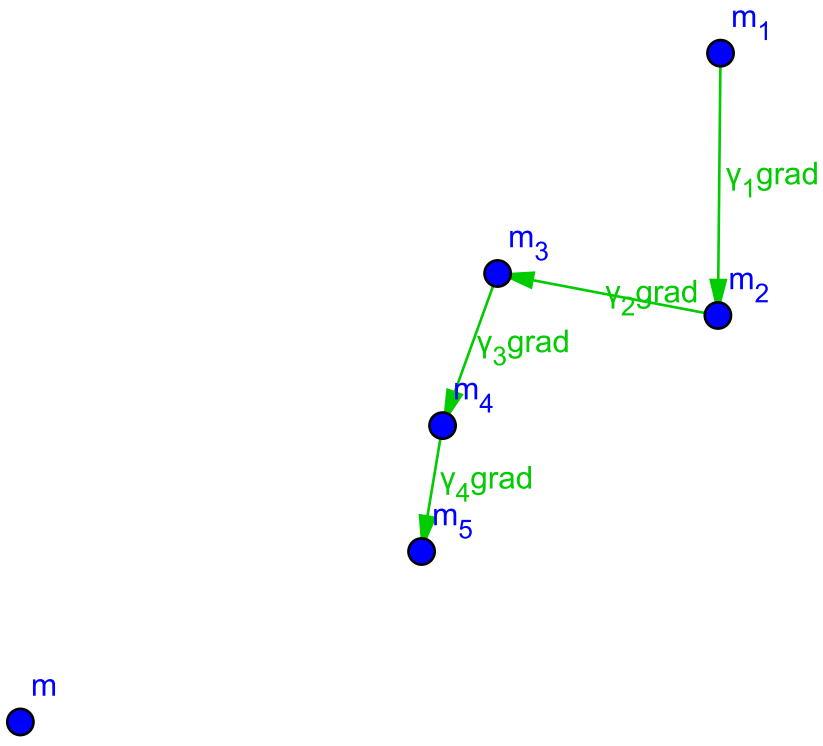

m1
γ1grad
m2
m

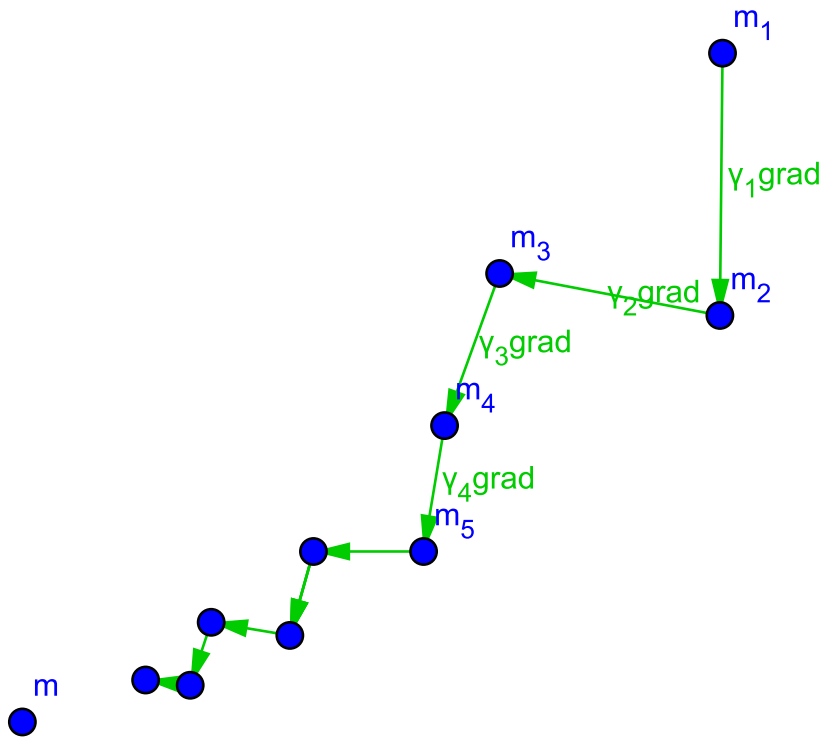

m1
γ1grad
γ2grad
m2
m


m1
γ1grad
m3
γ2grad
m2


m


m1
γ1grad
m3
m2
γ2grad
γ3grad
m4
m


m1
γ1grad
m3
γ2grad
m2
γ3grad
m4
γ4grad
m5
m


m1
γ1grad
m3
m2
γ2grad
γ3grad
m4
γ4grad
m5
m

ALGORITHMES DE GRADIENT STOCHASTIQUES

Fonction à minimiser :

$$G(h) = \mathbb{E}\left[g(X,h)\right]$$

Algorithme de gradient stochastique : Pour tout $n \geq 0$

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

avec m_0 borné et (γ_n) une suite de pas positifs vérifiant

$$\sum_{n\geq 1} \gamma_n = +\infty \qquad \text{et} \qquad \sum_{n\geq 1} \gamma_n^2 < +\infty.$$

Médiane géométrique: On a

$$m_{n+1} = m_n + \gamma_{n+1} \frac{X_{n+1} - m_n}{\|X_{n+1} - m_n\|}.$$

Régression linéaire: On a

$$m_{n+1} = m_n + \gamma_{n+1} \left(Y_{n+1} - m_n^T X_{n+1}\right).$$

APPROCHE NON ASYMPTOTIQUE

Pas : On prend

$$\gamma_n = c_\gamma n^{-\alpha}, \quad \text{avec} \quad c_\gamma > 0, \quad \text{et} \quad \alpha \in (1/2, 1)$$

Hypothèses :

1. Il existe un minimiseur m de la fonction G.
2. La fonction G est μ-fortement convexe: pour tout $h \in \mathbb{R}^d$,

$$\langle \nabla G(h), h - m \rangle \geq \mu \|h - m\|^2 .$$

3. Il existe $C \geq 0$ tel que pour tout $h \in \mathbb{R}^d$,

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

Vitesse de convergence : Il existe une constant C' telle que pour tout $n \geq 1$,

$$\mathbb{E}\left[\|m_n - m\|^2\right] \leq C' \gamma_n .$$

PREUVE

Lemme

Soit (δ_n) une suite positive vérifiant

$$\delta_{n+1} \leq \left(1 - a\gamma_{n+1} + b\gamma_{n+1}^2\right) \delta_n + c\gamma_{n+1}^2$$

avec $\gamma_n = c_\gamma n^{-\alpha}$, $c_\gamma, a, b, c > 0$, et $\alpha \in (1/2, 1)$. Alors il existe une constante C' telle que pour tout $n \geq 1$,

$$\delta_n \leq C' \gamma_n$$

Application à la régression linéaire

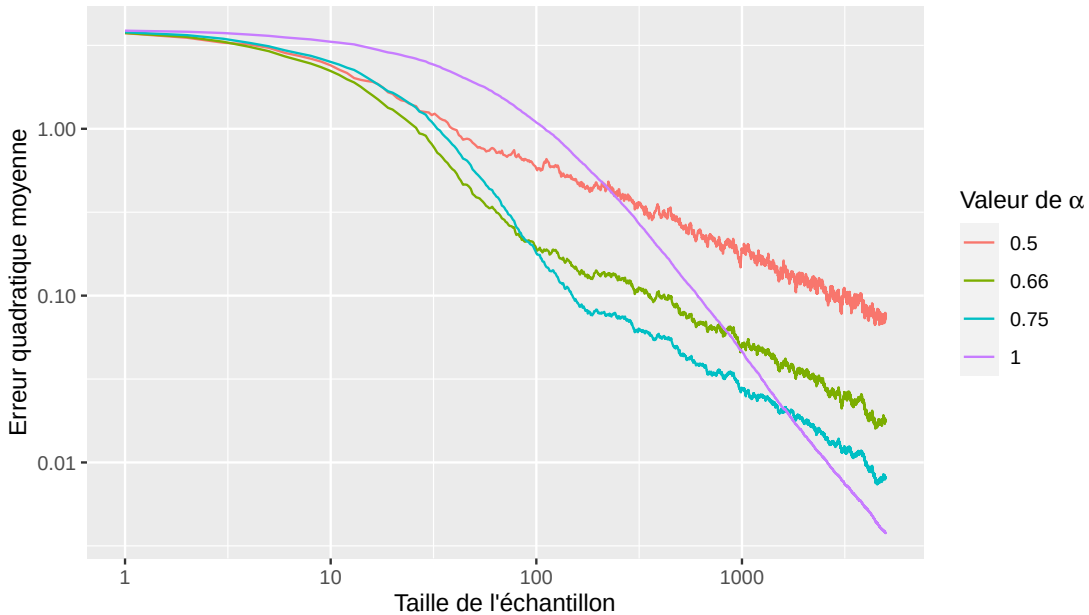


Figure: Evolution de l'erreur quadratique moyenne de θ_n en fonction de la taille d'échantillon n dans le cadre de la régression linéaire.

IV. Théorème de Robbins-Siegmund et convergence presque sûre

Théorème de Robbins-Siegmund

Théorème (Robbins-Siegmund)

Soit (V_n), (A_n), (B_n), (C_n) trois suites de variables aléatoires positives adaptées à une filtration $(\mathcal{F}_n)$. On suppose que

$$\mathbb{E}\left[V_{n+1}|\mathcal{F}_n\right] \leq (1 + A_n)\, V_n + B_n - C_n$$

et que les suites (A_n) et (B_n) vérifient

$$\sum_{n\geq 0} A_n < +\infty \quad p.s \qquad et \qquad \sum_{n\geq 0} B_n < +\infty \quad p.s.$$

Alors V_n converge presque sûrement vers une variable aléatoire finie et

$$\sum_{n\geq 0} C_n < +\infty \quad p.s.$$

ESTIMATION EN LIGNE DES QUANTILES

On considère x_p le quantile d'ordre $p \in (0,1)$ de X. On peut construire l'estimateur en ligne

$$m_{n+1} = m_n - \gamma_{n+1}\left(\mathbf{1}_{X_{n+1} \leq m_n} - p\right)$$

avec

$$\sum_{n \geq 1} \gamma_n = +\infty \qquad \text{et} \qquad \sum_{n \geq 1} \gamma_n^2 < +\infty.$$

Si la fonction de répartition F_X est strictement croissante au voisinage de x_p, alors

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

ESTIMATION EN LIGNE DES QUANTILES

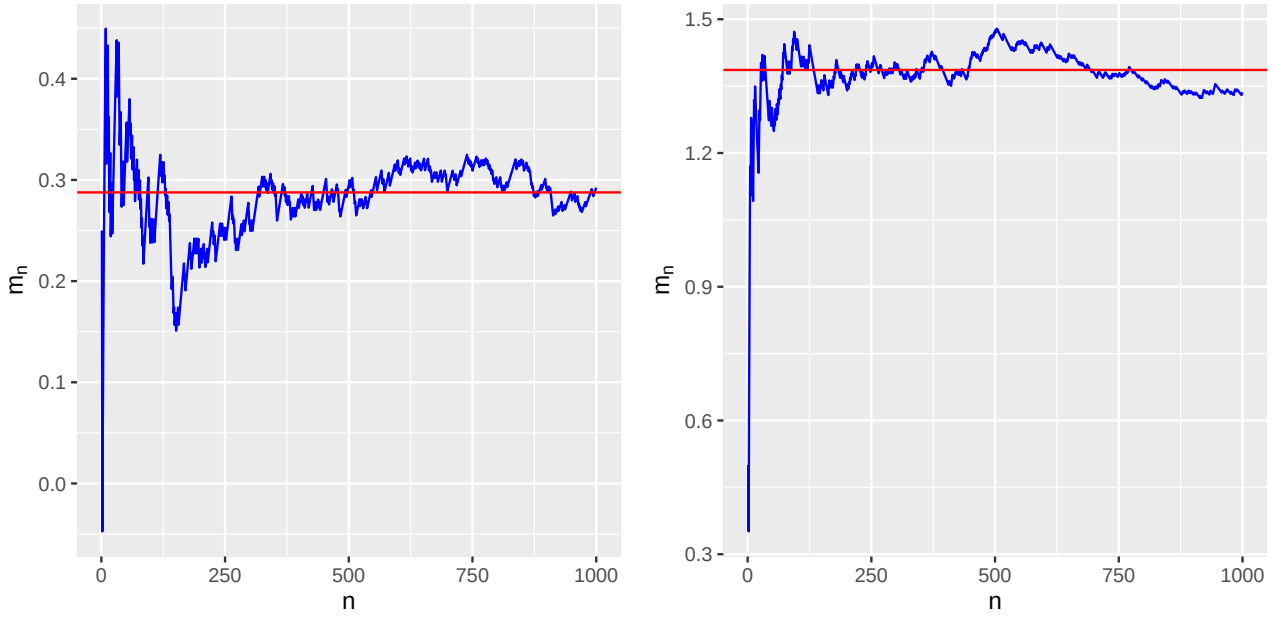


Figure: Evolution de l'estimation des quantiles d'ordre 0.25 (à gauche) et 0.75 à droite pour la loi exponentielle de paramètre 1.

Convergence presque sûre

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 qu'il existe $C > 0$ tel que pour tout h

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

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,

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

Alors

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

Optimalité ?

Remarque : 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

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

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