Algorithmes stochastiques

Introduction

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

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

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.

Exemples

Moyenne d'un vecteur aléatoire : On a

$$G_n(h) = \frac{1}{2n} \sum_{i=1}^{n} \|X_i - h\|^2 - \|X_i\|^2$$

et on retrouve $\hat{m}_n = \overline{X}_n$.

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 $\hat{\theta}_n = \left(\mathbf{X}\mathbf{X}^T\right)^{-1} \mathbf{X}^T \mathbf{Y}$.

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 $\hat{m}_n = X_{(\lceil \frac{n}{2} \rceil)}$.

"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

Un résultat de convergence

Théorème

On suppose que les hypothèses suivantes sont vérifiées :

- $\hat{m}_n$ *converge en probabilité vers* m.
- *Pour presque tout* x, *la fonction* $g(x, .)$ *est convexe et deux fois continument différentiable.*
- *Pour presque tout* x, *la fonction* $\nabla_h^2 g(x, .)$ *est* $L(x)$*-lipschitz, i.e*

$$\forall h, h', \quad \left\| \nabla^2 g\left(x, h\right) - \nabla^2 g\left(X, h'\right) \right\|_{op} \leq L(x) \left\| h - h' \right\| .$$

- ▶ $L(X)$ *et* $\nabla^2 g(X, m)$ *admettent des moments d'ordre* 2.
- ▶ $H := \nabla^2 G(m)$ *est inversible.*

Alors, en notant $\Sigma = \mathbb{E}\left[\nabla_h g\left(X, m\right) \nabla_h g\left(X, m\right)^T\right]$, *on a*

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

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 sans biais de la variance : On a

$$\Sigma_{n+1}^2 = \Sigma_n^2 + \frac{1}{n+1}\left(X_{n+1}X_{n+1}^T - \Sigma_n^2\right)$$
$$S_{n+1}^2 = \frac{n+1}{n}\Sigma_{n+1}^2 - \frac{n+1}{n}\overline{X}_{n+1}\overline{X}_{n+1}^T.$$

III. Algorithmes de gradient stochastiques

Algorithme de gradient

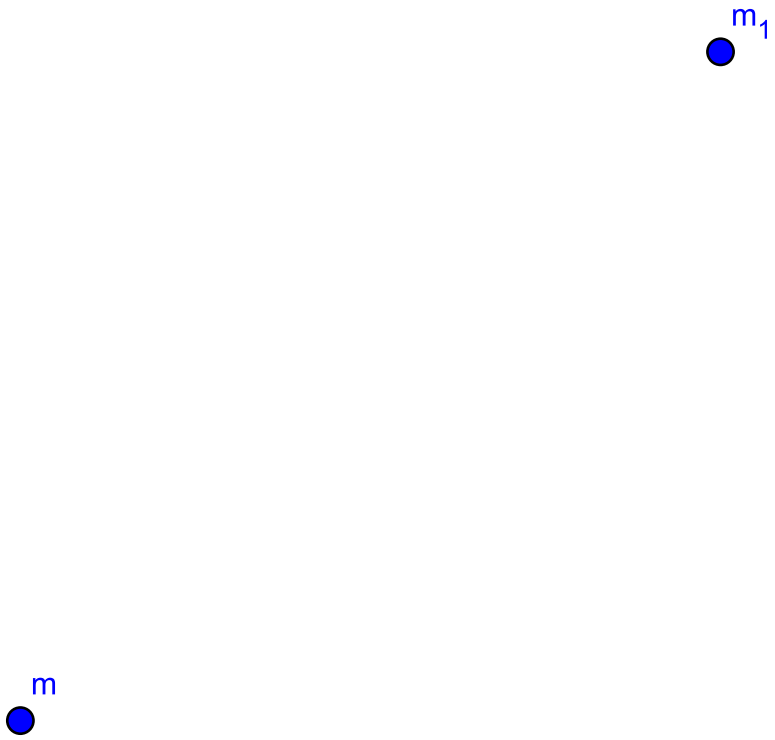
On cherche à minimiser la fonction convexe G définie par

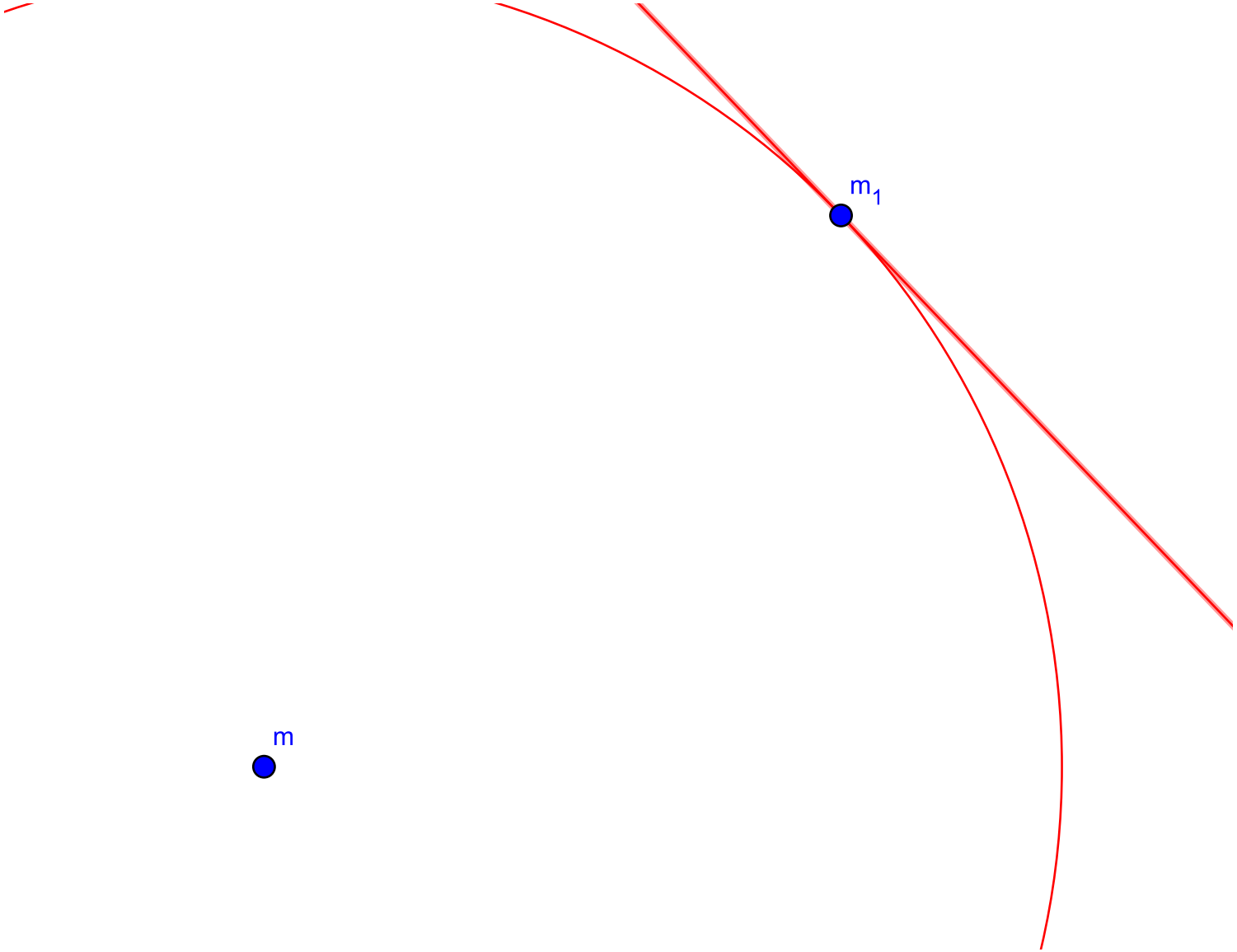
$$G(h) = \mathbb{E}\left[g\left(X, h\right)\right].$$

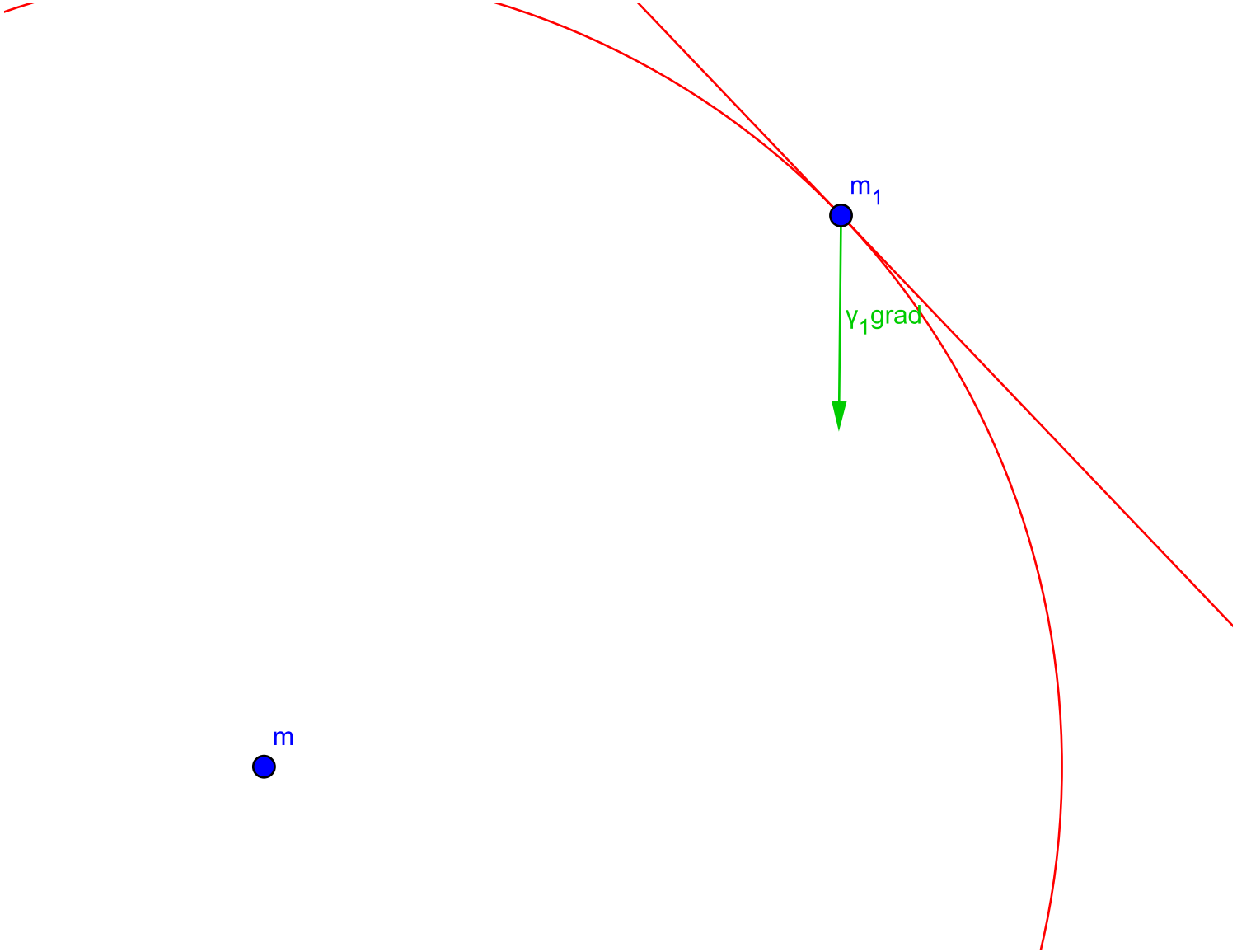
L'algorithme de gradient est défini de manière itérative par

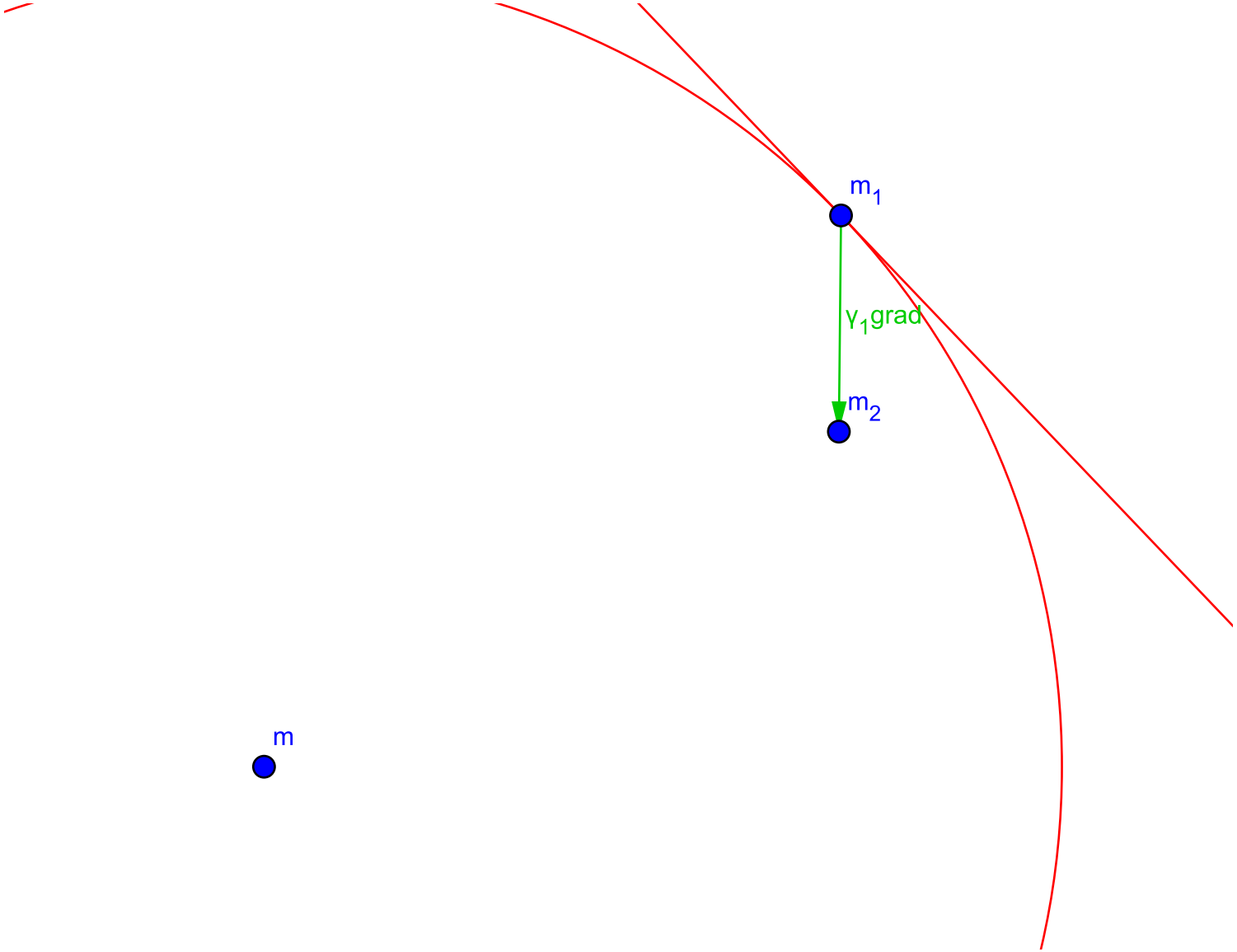
$$m_{t+1} = m_t - \gamma_t \nabla G\left(m_t\right)$$

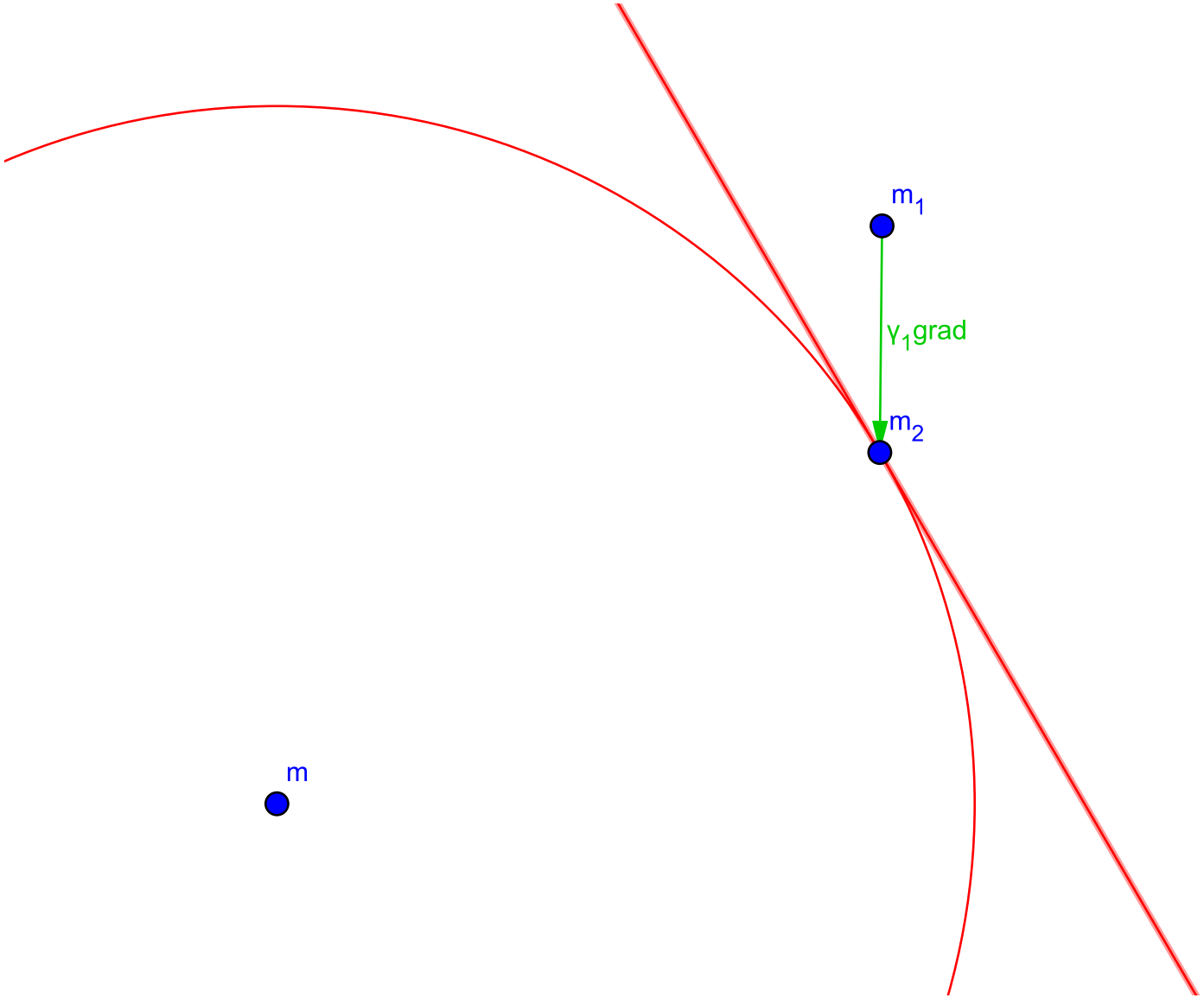
où η_t est une suite de pas positifs.

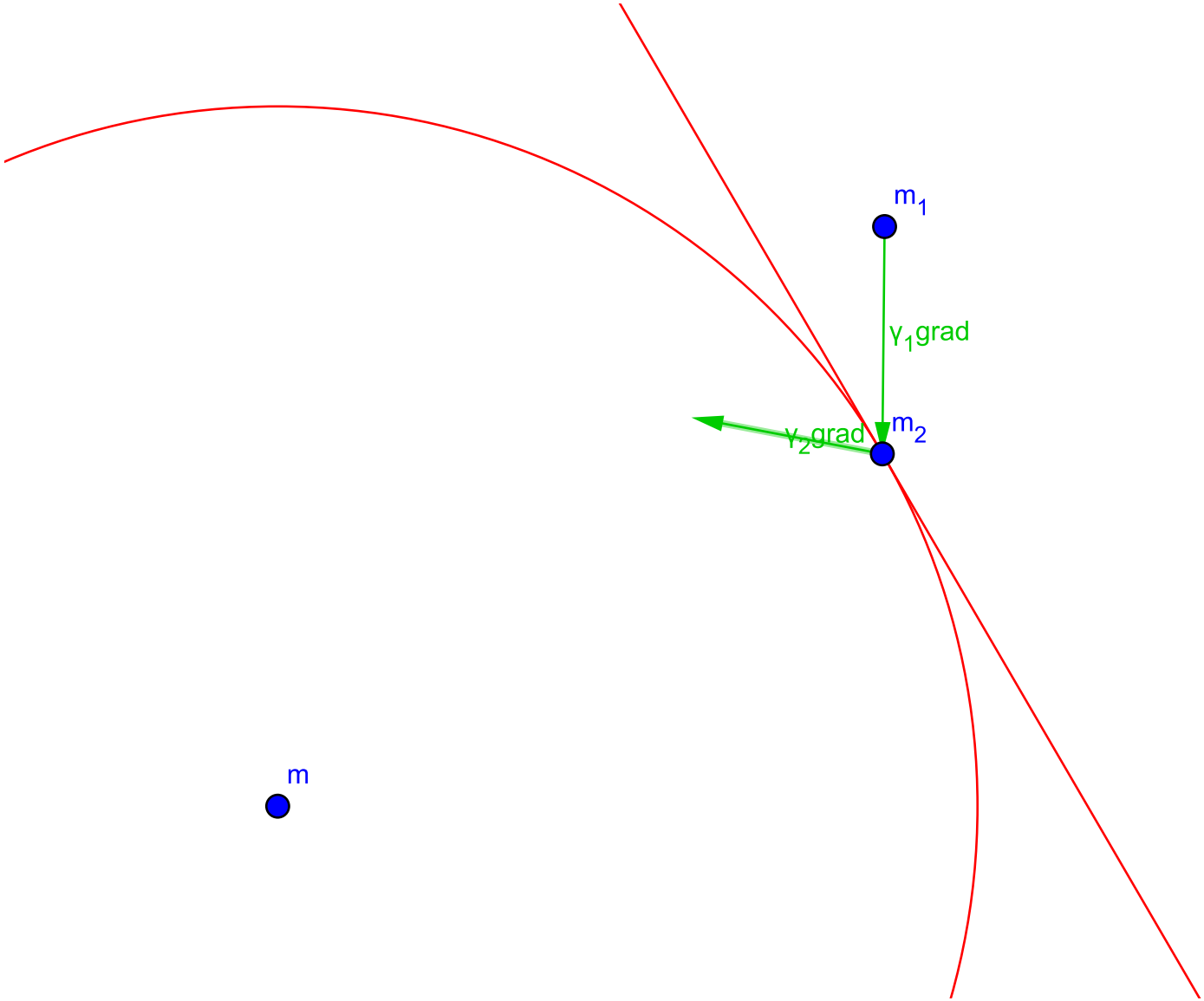

m_1
m

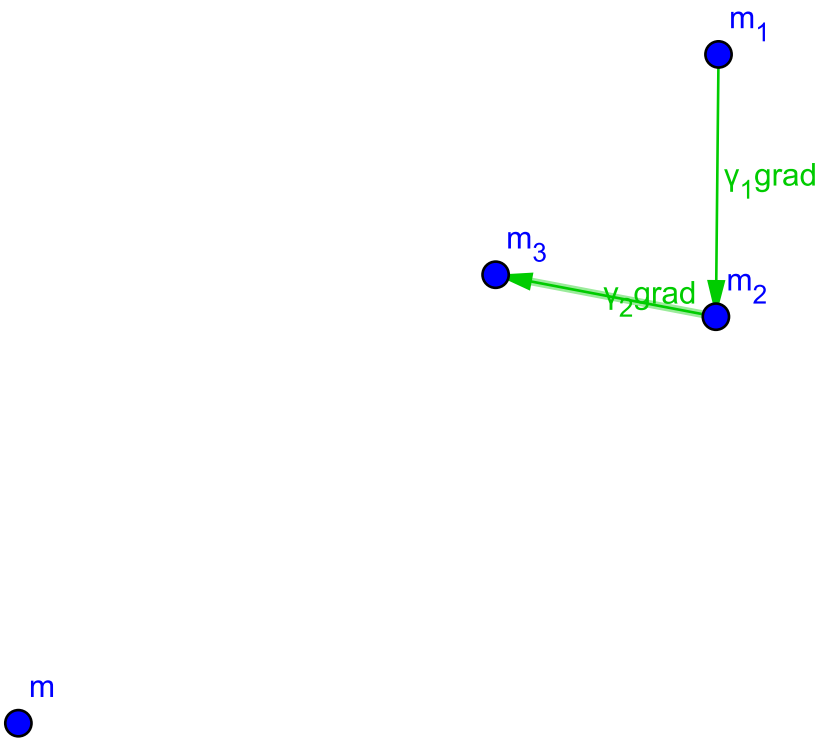

m_1
m

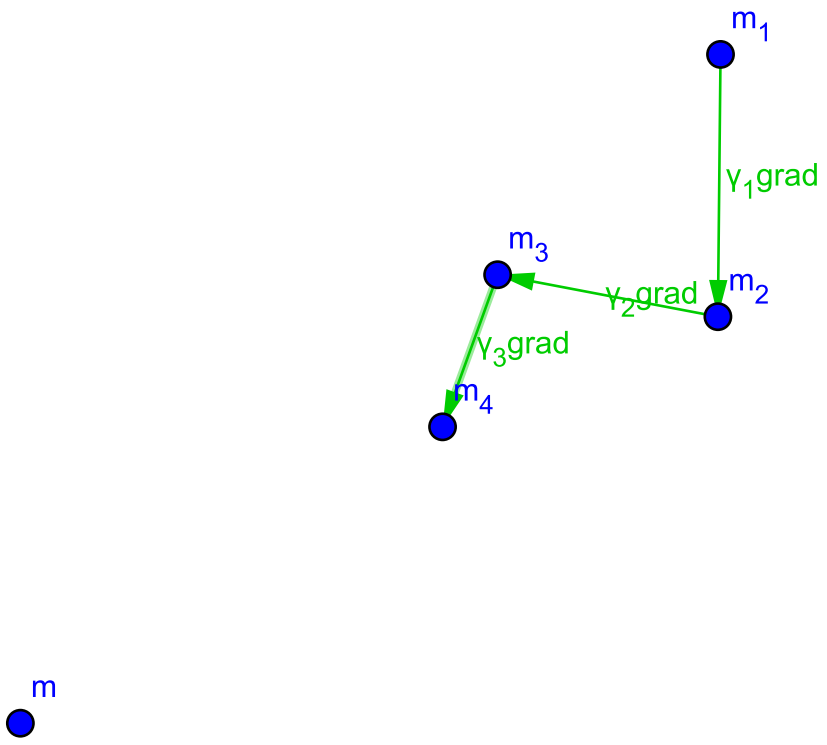

m_1
$\gamma_1$grad
m

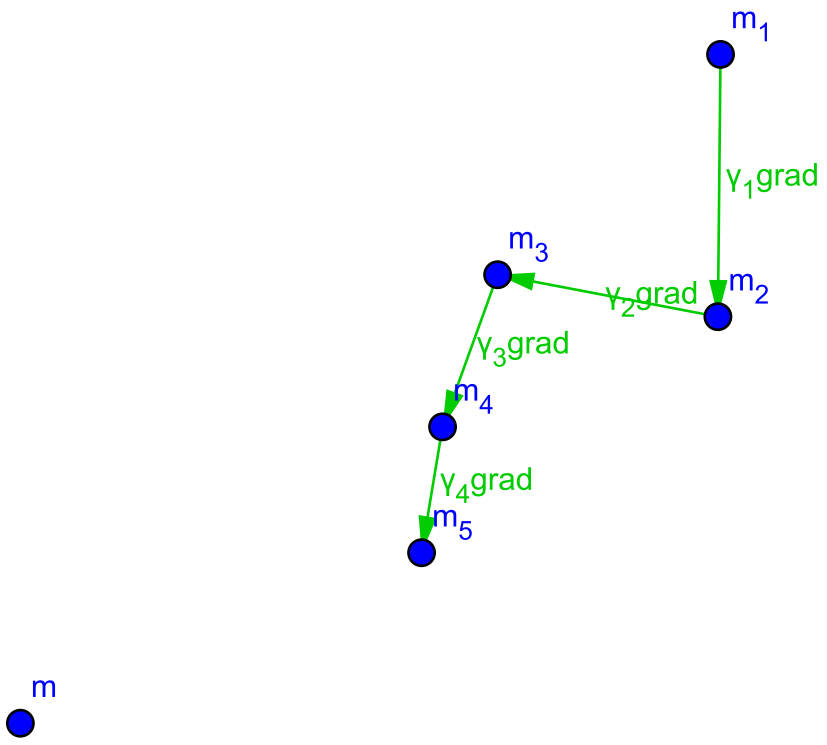

m_1
$\gamma_1$grad
m_2
m

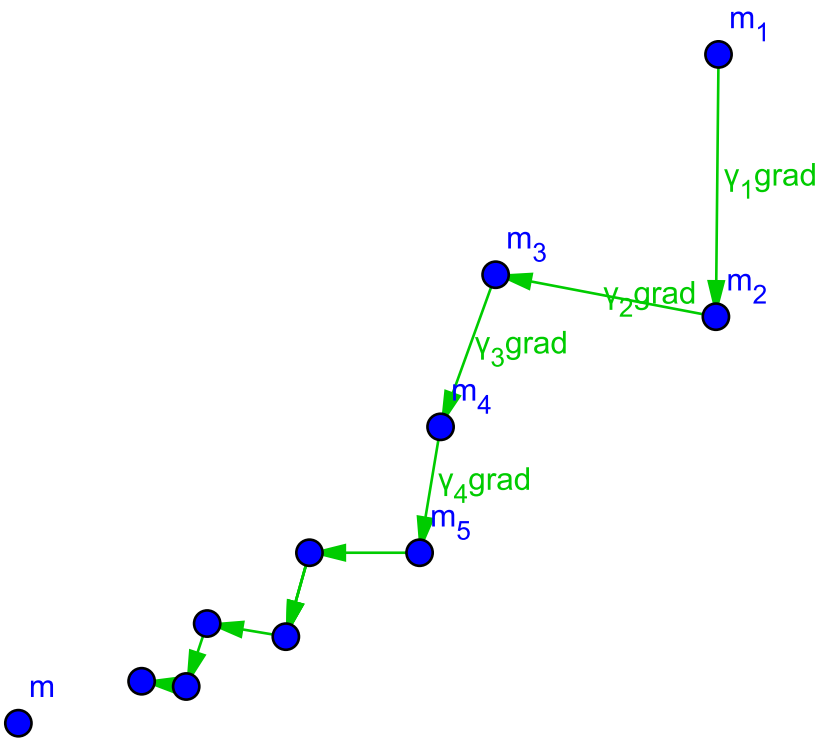

m_1
$\gamma_1$grad
m_2
m


m1
γ1grad
m2
γ2grad
m


m1
γ1grad
m3
m2
γ2grad
m


m1
γ1grad
m3
m2
γ2grad
γ3grad
m4
m


m1
γ1grad
m3
γ2grad
m2
γ3grad
m4
γ4grad
m5
m


m_1
$\gamma_1$grad
m_3
$\gamma_2$grad
m_2
$\gamma_3$grad
m_4
$\gamma_4$grad
m_5
m

Algorithmes de gradient stochastiques

L'algorithme de gradient stochastique est défini de manière récursive pour tout $n \geq 0$ par

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

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

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

Exemples

Régression linéaire : On a

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

Régression logistique : On a

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

Médiane géométrique : On a

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

Approche non asymptotique

On prend $\gamma_n = c_\gamma n^{-\alpha}$ avec $c_\gamma > 0$ et $\alpha \in (1/2, 1)$ et on suppose que les hypothèses suivantes sont vérifiées :

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

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

(PS0) Il existe $C \geq 0$ tel que pour tout $h \in \mathbb{R}^d$,

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

Alors, en notant $C' = \max\left\{C, 2\mu^2\right\}$, pour tout $n \geq 1$,

$$\mathbb{E}\left[\left\| m_n - m \right\|^2 \right] \leq 2 \exp\left(C' c_\gamma^2 \frac{2\alpha}{2\alpha - 1} \right) \exp\left(-\frac{\mu c_\gamma}{4} n^{1-\alpha} \right) \left(\mathbb{E}\left[\left\| m_0 - m \right\|^2 \right] + 1 \right) + \frac{2 c_\gamma C}{\mu n^\alpha}$$

Preuve

La preuve repose sur le lemme suivant (admis) :

Lemme

Soit (δ_n) *une suite positive vérifiant*

$$\delta_{n+1} \leq \left(1 - 2\mu\gamma_{n+1} + 2L^2\gamma_{n+1}^2\right)\delta_n + 2\sigma^2\gamma_{n+1}^2$$

avec $\gamma_n = c_\gamma n^{-\alpha}$, $c_\gamma, L \geq \mu > 0$, $\sigma^2 \geq 0$ *et* $\alpha \in (1/2, 1)$. *Alors pour tout* $n \geq 1$,

$$\delta_n \leq 2\exp\left(-\frac{\mu c_\gamma}{4}n^{1-\alpha}\right)\exp\left(2L^2c_\gamma^2\frac{2\alpha}{2\alpha-1}\right)\left(\delta_0 + \frac{\sigma^2}{L^2}\right) + \frac{4c_\gamma\sigma^2}{\mu n^\alpha}$$

Application à la régression linéaire

On suppose

- $\mathbb{E}\left[\epsilon^2\right] < +\infty$.
- $\mathbb{E}\left[X^4\right] < +\infty$
- La matrice $\mathbb{E}\left[XX^T\right]$ est définie positive et on note μ sa plus petite valeurs propre.

On a alors

$$\mathbb{E}\left[\|\theta_n - \theta\|^2\right] \leq 2\exp\left(Cc_\gamma^2 \frac{2\alpha}{2\alpha - 1}\right)\exp\left(-\frac{\mu c_\gamma}{4} n^{1-\alpha}\right)\left(\mathbb{E}\left[\|m_0 - m\|^2\right] + 1\right) + \frac{2c_\gamma C}{\mu n^\alpha}$$

avec $C = \max\left\{2\mathbb{E}\left[\epsilon^2\right]\mathbb{E}\left[\|X\|^2\right], 2\mathbb{E}\left[\|X\|^4\right]\right\}$.

Application à la régression linéaire

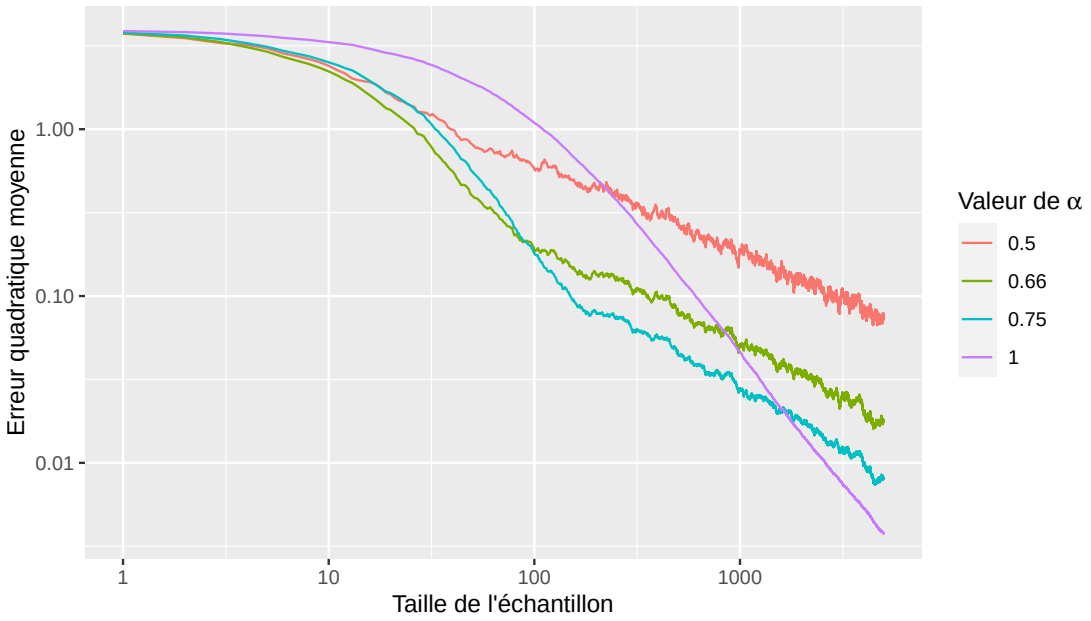


Figure – Evolution de l'erreur quadratique moyenne de θ_n en fonction de la taille d'échantillon n dans le cadre de la régression linéaire.

Une proposition utile

Proposition

Soit δ_n, γ_n deux suites positives telles que

- $\gamma_n = c_\gamma n^{-\alpha}$ *avec* $c_\gamma > 0$ *et* $\alpha \in (1/2, 1)$.
- *Il existe un rang* n_0*, une constante* $c_0 \in \left(0, \gamma_{n_0}^{-1}\right)$ *et une suite positive* v_n *telle que pour tout* $n \geq n_0$*,*

$$\delta_{n+1} \leq (1 - c_0 \gamma_{n+1})\, \delta_n + \gamma_{n+1} v_{n+1}.$$

Alors pour tout $n \geq 2n_0$*,*

$$\delta_n \leq \exp\left(-\frac{c_0 c_\gamma}{4} n^{1-\alpha}\right) \left(\delta_{n_0} + \sum_{k=n_0}^{n/2-1} \gamma_{k+1} v_{k+1}\right) + \max_{n/2 \leq k+1 \leq n-1} v_{k+1}$$

Pour simplifier

Corollaire

Soit δ_n, γ_n deux suites positives telles que

- $\gamma_n = c_\gamma n^{-\alpha}$ *avec* $c_\gamma > 0$ *et* $\alpha \in (1/2, 1)$.
- *Il existe un rang* n_0*, une constante* $c_0 \in \left(0, \gamma_{n_0}^{-1}\right)$ *et une suite positive* v_n *telle que pour tout* $n \geq n_0$*,*

$$\delta_{n+1} \leq (1 - c_0 \gamma_{n+1})\, \delta_n + \gamma_{n+1} v_{n+1}.$$

Si $v_n = C_v (\ln n)^\beta n^v$ *avec* $\beta, v \in \mathbb{R}$*, alors*

$$\delta_n = O\left(v_n\right).$$