

# Régression linéaire et gradient stochastique

Ce TD-TP a pour objectif de plonger de manière un peu plus individualisée dans le cours d'optimisation stochastique. Il n'est pas noté, mais j'encourage très vivement les étudiants à rédiger consciencieusement leurs réponses et leurs idées. La rédaction force à mieux présenter les choses et surtout à mieux les cerner. Ça me permettra aussi plus facilement de corriger d'éventuelles incompréhensions.

## 1 Introduction

L'objectif de ce TP est d'illustrer la première partie du cours d'optimisation stochastique sur les erreurs en apprentissage et d'implémenter les premières versions du gradient stochastique. Il a aussi pour objectif de vous entraîner à effectuer des calculs avec des variables aléatoires pour les rendre plus accessibles et mieux comprendre les cours à venir.

Nous allons nous placer dans le cadre de travail le plus simple : la régression linéaire. Ce cadre présente plusieurs avantages :

- C'est probablement le plus simple d'un point de vue théorique et il permet d'appréhender de nombreux phénomènes avec des mathématiques relativement élémentaires.
- C'est probablement encore le plus utilisé dans les applications, et il me semble nécessaire de le comprendre profondément.

## 2 Le cadre

Soit  $X$  un vecteur aléatoire de  $\mathbb{R}^d$ , pour  $d \in \mathbb{N}$  suivant une certaine distribution de probabilité inconnue  $P_X$ . Pour un certain vecteur  $\theta \in \mathbb{R}^d$ , on construit une variable aléatoire  $Y \in \mathbb{R}$  définie par :

$$Y = \langle \theta, X \rangle + B \quad \text{où} \quad B \sim \mathcal{N}(0, \sigma^2) \text{ est une variable aléatoire gaussienne indépendante de } X.$$

L'objectif de ce TP est d'apprendre le vecteur  $\theta$  inconnu à partir de  $n \in \mathbb{N}$  observations  $(x_i, y_i)_{1 \leq i \leq n}$  tirées indépendamment suivant la loi  $P$ .

Pour ce faire, on peut simplement résoudre le problème de minimisation du risque empirique suivant :

$$\inf_{\omega \in \mathbb{R}^d} E_n(\omega) = \frac{1}{2n} \sum_{i=1}^n (\langle \omega, x_i \rangle - y_i)^2 \quad (1)$$

On notera  $\omega_n^*$  n'importe quel minimiseur (sous réserve d'existence) du problème ci-dessus.

### 3 Questions préliminaires

1. Déterminer les espaces  $\mathcal{X}$  et  $\mathcal{Y}$  du cours. Quelle est la fonction perte  $l$  utilisée ici ? Quelle est la famille  $\mathcal{H}$  de prédicteurs utilisés ?

On a :  $h : \mathcal{X} \rightarrow \mathcal{Y}$ .

Fonction perte :  $l : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$

$$(x, y) \mapsto \frac{1}{2}(\langle \omega, x \rangle - y)^2$$

La famille des prédicteurs est  $\mathcal{H} = \{x \mapsto \langle \omega, x \rangle, \omega \in \mathbb{R}^d\}$ .

2. Déterminer la loi conditionnelle  $P(Y|X)$ .

Loi conditionnelle  $P(Y|X)$  :

$$\begin{aligned} P(Y = y|X = x) &= P(\langle \theta, X \rangle + B = y|X = x) \\ &= P(\langle \theta, x \rangle + B = y) \\ &= P(B = y - \langle \theta, x \rangle) \end{aligned}$$

$$\Rightarrow Y|X = x \sim \mathcal{N}(\langle \theta, x \rangle, \sigma^2)$$

3. Quelle est la définition du risque moyen ici ?

Risque moyen :

$$\begin{aligned} E(\omega) &= \mathbb{E}_{(X,Y)} [l(X, Y)] \\ &= \mathbb{E}_{(X,Y)} \left[ \frac{1}{2}(\langle \omega, X \rangle - Y)^2 \right] \\ &= \frac{1}{2} \mathbb{E} [(\langle \omega, X \rangle - \langle \theta, X \rangle - B)^2] \\ &= \frac{1}{2} \mathbb{E} [(\langle \omega - \theta, X \rangle - B)^2] \\ &= \frac{1}{2} \mathbb{E} [\langle \omega - \theta, X \rangle^2] - \mathbb{E} [\langle \omega - \theta, X \rangle B] + \frac{1}{2} \mathbb{E} [B^2] \\ &= \frac{1}{2} \mathbb{E} [\langle \omega - \theta, X \rangle^2] - \mathbb{E} [\langle \omega - \theta, X \rangle] \mathbb{E} [B] + \frac{1}{2} \text{Var}(B) \\ &= \frac{1}{2} \mathbb{E} [\langle \omega - \theta, X \rangle^2] + \frac{\sigma^2}{2} \end{aligned}$$

4. Quel est le prédicteur optimal  $\omega^*$  ? Est-il unique ? Que vaut  $E(\omega^*)$  ?

Prédicteur optimal :  $\omega^* = \theta$ .

Il n'est pas unique si  $X = 0$  presque partout.

5. Si  $X \sim \mathcal{N}(0, I_d)$  que vaut  $E(\omega)$  ?

Si  $X \sim \mathcal{N}(0, I_d)$ , alors :

$$\begin{aligned}
 E(\omega) &= \frac{1}{2} \mathbb{E} [\langle \omega - \theta, X \rangle^2] + \frac{\sigma^2}{2} \\
 &= \frac{1}{2} \mathbb{E} [\langle \omega - \theta, X \rangle \langle \omega - \theta, X \rangle] + \frac{\sigma^2}{2} \\
 &= \frac{1}{2} \mathbb{E} [(\omega - \theta)^T X X^T (\omega - \theta)] + \frac{\sigma^2}{2} \\
 &= \frac{1}{2} (\omega - \theta)^T \mathbb{E} [X X^T] (\omega - \theta) + \frac{\sigma^2}{2} \\
 &= \frac{1}{2} (\omega - \theta)^T \text{Var}(X) (\omega - \theta) + \frac{\sigma^2}{2} \\
 &= \frac{1}{2} (\omega - \theta)^T I_d (\omega - \theta) + \frac{\sigma^2}{2} \\
 &= \frac{1}{2} (\omega - \theta)^T (\omega - \theta) + \frac{\sigma^2}{2} \\
 &= \frac{1}{2} \|\omega - \theta\|^2 + \frac{\sigma^2}{2} \text{ (Norme de Mahalanobis)}
 \end{aligned}$$

6. Déterminer l'erreur d'approximation  $\mathcal{E}_{app}$  pour ce problème.

On a :  $\mathcal{E}_{app} = E(h^*) - E(h_{\mathcal{H}}^*)$ .

avec :  $h^* = \underset{h \in \mathbb{R}^d}{\operatorname{argmin}} E(h)$

$h_{\mathcal{H}}^* = \underset{h \in \mathcal{H}}{\operatorname{argmin}} E_n(h)$

$$\begin{aligned}
 E(h) &= \mathbb{E} [l(h(X), Y)] \\
 &= \frac{1}{2} \mathbb{E} [(h(X) - Y)^2] \\
 &= \frac{1}{2} \mathbb{E} [(h(X) - \langle \theta, X \rangle - B)^2] \\
 &= \frac{1}{2} \mathbb{E} [(h(X) - \langle \theta, X \rangle)^2] + \frac{\sigma^2}{2} \geq \frac{\sigma^2}{2}
 \end{aligned}$$

$h^*(x) = \langle \theta, x \rangle \in \mathcal{H}$ , donc  $E(h^*) - E(h_{\mathcal{H}}^*) = 0$  : Pas d'erreur d'approximation.

7. Est-ce que la fonction  $E_n$  est convexe ou non convexe ?

On a :  $E_n(\omega) = \frac{1}{2n} \sum_{i=1}^n (\langle \omega, x_i \rangle - y_i)^2$ .

avec :  $\omega \mapsto \langle \omega, x_i \rangle - y_i$  et  $t \mapsto t^2$  convexes.

Donc  $E_n$  est convexe par somme et composition de fonctions convexes.

8. Calculer  $\nabla E_n(\omega)$ .

$$\text{On a : } E_n(\omega) = \frac{1}{2n} \sum_{i=1}^n (\langle \omega, x_i \rangle - y_i)^2.$$

$$\text{On pose : } X = (x_1, \dots, x_n) \in \mathbb{R}^{d \times n}$$

$$\text{et } Y = (y_1, \dots, y_n) \in \mathbb{R}^n$$

$$\text{On a donc : } E_n(\omega) = \frac{1}{2n} \|X^T \omega - Y\|^2.$$

$$\begin{aligned} \text{Donc : } \nabla E_n(\omega) &= \frac{1}{n} (X^T)^T (X^T \omega - Y) \\ &= \frac{1}{n} X (X^T \omega - Y) \end{aligned}$$

9. Est-ce que le problème (1) possède une solution ? Une solution unique ?

La fonction  $E_n$  est convexe, différentiable et coercive (car  $E_n(\omega) \xrightarrow{\|\omega\| \rightarrow \infty} \infty$ ).

On a la condition d'optimalité :  $\nabla E_n(\omega) = 0 \Leftrightarrow \frac{1}{n} X (X^T \omega - Y) = 0 \Leftrightarrow X X^T \omega = X Y$ .

On montre que ce problème admet une solution : on montre  $X Y \in \text{Im}(X X^T)$ .

- On a :  $X Y \in \text{Im}(X)$  (trivial).
- Avec une SVD, on a :  $X = U \Sigma V^T$  avec  $U$  et  $V$  unitaires  
 $U = (u_1, \dots, u_r)$  avec  $r = \text{rg}(X)$

$$\text{Donc : } \text{Im}(X) = \text{Vect}(u_1, \dots, u_r)$$

- On a aussi :  $X X^T = U \Sigma V^T V \Sigma^T U^T = U \Sigma \Sigma^T U^T$

$$\text{Donc : } \text{Im}(X X^T) = \text{Vect}(u_1, \dots, u_r) = \text{Im}(X)$$

On obtient donc :  $X Y \in \text{Im}(X X^T)$ , donc le problème admet une solution.

## 4 Erreur d'estimation

Dans cette partie, on se propose de borner l'erreur d'estimation  $\mathcal{E}$

$$\mathcal{E}_{est} = E(\omega_n^*) - E(\omega^*),$$

pour se faire une idée de la vitesse de convergence du risque empirique.

Soient  $(Z_i)_{1 \leq i \leq n}$  un ensemble de variables aléatoires i.i.d. de variance  $\sigma^2$ .

1. Que vaut  $Var(\frac{1}{n} \sum_{i=1}^n Z_i)$  ?

Les variables aléatoires  $(Z_i)$  sont indépendantes donc :

$$\begin{aligned} Var(\frac{1}{n} \sum_{i=1}^n Z_i) &= \frac{1}{n^2} Var(\sum_{i=1}^n Z_i) \\ &= \frac{1}{n^2} \sum_{i=1}^n Var(Z_i) \text{ (car les } Z_i \text{ sont indépendantes)} \\ &= \frac{1}{n^2} n Var(Z_1) \\ &= \frac{1}{n} Var(Z_1) \\ &= \frac{\sigma^2}{n} \end{aligned}$$

2. Que peut-on en déduire sur la différence  $E_n(\omega) - E(\omega)$  ?

On cherche la variance de l'estimateur :  $E_n(\omega) - E(\omega)$ .

$$\begin{aligned} \mathbb{E}[(E_n(\omega) - E(\omega))^2] &= \mathbb{E}\left[\left[\frac{1}{n} \sum_{i=1}^n \left(\frac{1}{2}(\langle \omega, x_i \rangle - y_i)^2 - E(\omega)\right)\right]^2\right] \\ &= 1 \end{aligned}$$

Malheureusement, ce résultat est valable pour un vecteur  $\omega \in \mathbb{R}^d$  quelconque, mais ne permet pas de contrôler  $E(\omega_n^*) - E(\omega^*)$  directement. On va donc affiner ce résultat. On suppose que les paramètres optimaux  $\omega^*$  et  $\omega_n^*$  vivent dans une boule de rayon  $R > 0$ . On utilise la décomposition suivante :

$$\mathcal{E}_{est} = [E(\omega_n^*) - E_n(\omega_n^*)] + [E_n(\omega_n^*) - E_n(\omega^*)] + [E_n(\omega^*) - E(\omega^*)].$$

1. Que pouvez-vous dire de  $E_n(\omega_n^*) - E_n(\omega^*)$  ?

Erreur d'estimation :  $E(\omega_n^*) - E(\omega^*)$ .

$$\begin{aligned} \text{avec : } \quad \omega_n^* &= \operatorname{argmin}_{\omega \in \mathbb{R}^d} E_n(\omega) \\ \text{et } \quad \omega^* &= \operatorname{argmin}_{\omega \in \mathbb{R}^d} E(\omega) \end{aligned}$$

On a :  $E_n(\omega_n^*) - E_n(\omega^*) \leq 0$  car  $\omega_n^* \stackrel{def}{=} \operatorname{argmin}_{\omega \in \mathbb{R}^d} E_n(\omega)$ .

2. En déduire que :  $\mathcal{E}_{est} \leq 2 \sup_{\|\omega\|_2 \leq R} |E(\omega) - E_n(\omega)|$ .

On a :

$$\begin{aligned}
\mathcal{E}_{est} &= [E(\omega_n^*) - E_n(\omega_n^*)] + [E_n(\omega_n^*) - E_n(\omega^*)] + [E_n(\omega^*) - E(\omega^*)] \\
&\leq [E(\omega_n^*) - E_n(\omega_n^*)] + [E_n(\omega^*) - E(\omega^*)] \\
&\leq |E(\omega_n^*) - E_n(\omega_n^*) + E_n(\omega^*) - E(\omega^*)| \\
&\leq |E(\omega_n^*) - E_n(\omega_n^*)| + |E_n(\omega^*) - E(\omega^*)| \\
&\leq \sup_{\|\omega\|_2 \leq R} |E(\omega) - E_n(\omega)| + \sup_{\|\omega\|_2 \leq R} |E(\omega) - E_n(\omega)| \\
&\leq 2 \sup_{\|\omega\|_2 \leq R} |E(\omega) - E_n(\omega)|
\end{aligned}$$

3. Etablir l'identité suivante :

$$\begin{aligned}
E_n(\omega) - E(\omega) &= \frac{1}{2} \langle \omega, (\frac{1}{n} \sum_{i=1}^n x_i x_i^T - \mathbb{E}(X X^T)) \omega \rangle \\
&\quad - \langle \omega^T, (\frac{1}{n} \sum_{i=1}^n y_i x_i^T - \mathbb{E}(Y X)) \rangle \\
&\quad + \frac{1}{2} (\frac{1}{n} \sum_{i=1}^n y_i^2 - \mathbb{E}(Y^2)).
\end{aligned}$$

On a :

$$\begin{aligned}
E(\omega) &= \frac{1}{2} (\omega - \theta)^T \mathbb{E}[X X^T] (\omega - \theta) + \frac{\sigma^2}{2} \\
&= \frac{1}{2} (\omega^T \mathbb{E}[X X^T] - \theta^T \mathbb{E}[X X^T]) (\omega - \theta) + \frac{\sigma^2}{2} \\
&= \frac{1}{2} (\langle \omega, \mathbb{E}[X X^T] \omega \rangle - \langle \theta, \mathbb{E}[X X^T] \omega \rangle \\
&\quad - \langle \omega, \mathbb{E}[X X^T] \theta \rangle + \langle \theta, \mathbb{E}[X X^T] \theta \rangle) + \frac{\sigma^2}{2}
\end{aligned}$$

et

$$\begin{aligned}
E_n(\omega) &= \frac{1}{2n} \sum_{i=1}^n (\langle \omega, x_i \rangle - y_i)^2 \\
&= \frac{1}{2n} \sum_{i=1}^n (\langle \omega, x_i \rangle^2 + y_i^2 - 2 \langle \omega, x_i \rangle y_i) \\
&= \frac{1}{2n} \sum_{i=1}^n (\langle \omega, x_i x_i^T \omega \rangle + y_i^2 - 2 \langle \omega, y_i x_i \rangle)
\end{aligned}$$

Donc :

$$|E_n(\omega) - E(\omega)| = \frac{1}{2n} \quad \text{Texte Manquant}$$

4. Borner la valeur absolue des erreurs ci-dessus en espérance. Vous pourrez par exemple utiliser des inégalités de Bernstein (scalaires, vectorielles et matricielles). Que conclure sur le taux de convergence de  $\mathcal{E}_{est}$  vers 0 ?

On pose :

$$\begin{aligned} |\mathbb{E}_c| &= \sup |\mathbb{E}(\langle \omega, \left( \frac{1}{n} \sum_{i=1}^n x_i x_i^T - \mathbb{E}[XX^T] \right) \omega \rangle)| \\ &= \sup |\mathbb{E}(\langle \omega, \left( \frac{1}{n} \sum_{i=1}^n Z_i \omega \right) \rangle)| \text{ avec } Z_i = x_i x_i^T - \mathbb{E}(XX^T) \end{aligned}$$

*Texte Manquant*

## 5 Questions d'optimisation stochastique

Dans cette partie, on suppose que  $i_k$  est un indice aléatoire tiré uniformément dans l'ensemble  $\{1, \dots, n\}$ .

1. Est-ce que  $E_n$  est fortement convexe ?

On a :  $E_n(\omega) = \frac{1}{2n} \sum_{i=1}^n (\langle \omega, x_i \rangle - y_i)^2$ .

Son gradient est :  $\nabla E_n(\omega) = \frac{1}{n} X(X^T \omega - Y)$ .

Sa hessienne est :  $H_{E_n}(\omega) = \frac{1}{n} XX^T$ .

Si  $\lambda_{\min}(XX^T) > 0$ , alors  $E_n$  est strictement convexe.

2. Soit  $f_i(\omega) = \frac{1}{2} \|\langle \omega, x_i \rangle - y_i\|_2^2$ . Calculer  $\nabla f_i(\omega)$ .

On a :  $f_i(\omega) = \frac{1}{2} \|\langle \omega, x_i \rangle - y_i\|_2^2$ .

Donc :  $\nabla f_i(\omega) = x_i(x_i^T \omega - y_i)$ .

3. Calculer  $\mathbb{E}_{i_k}(\nabla f_{i_k}(\omega))$ .

On a :  $\mathbb{E}_{i_k}[\nabla f_{i_k}(\omega)] = \mathbb{E}[x_{i_k}(x_{i_k}^T \omega - y_{i_k})] = \mathbb{E}[x_{i_k} x_{i_k}^T] \omega - \mathbb{E}[x_{i_k} y_{i_k}]$ .

4. Calculer la constante de Lipschitz  $L$  de  $\nabla E_n(\omega)$ .

Soit  $f$  une fonction  $L$ -lipschitzienne.

$\forall x, y \in \mathbb{R}^d, \quad \|f(x) - f(y)\|_2 \leq L \|x - y\|_2$ .

$$\begin{aligned} \text{On a : } \|\nabla E_n(\omega) - \nabla E_n(\omega')\|_2 &= \left\| \frac{1}{n} XX^T(\omega - \omega') \right\|_2 \\ &\leq \left\| \frac{1}{n} XX^T \right\|_2 \|\omega - \omega'\|_2 \end{aligned}$$

Donc :  $L = \left\| \frac{1}{n} XX^T \right\|_2$  (c'est la norme spectrale de la matrice  $\frac{1}{n} XX^T$ , qui est égale à sa plus grande valeur propre).

5. Calculer  $\mathbb{E}_{i_k}(\|\nabla f_{i_k}(\omega)\|_2^2)$  en fonction de  $\|\nabla E_n(\omega)\|_2^2$ .

$$\begin{aligned}
\mathbb{E}_{i_k} [\|\nabla f_{i_k}(\omega)\|_2^2] &= \mathbb{E}_{i_k} [\|x_{i_k}(x_{i_k}^T \omega - y_{i_k})\|_2^2] \\
&= \mathbb{E}_{i_k} [\langle x_{i_k}(x_{i_k}^T \omega - y_{i_k}), x_{i_k}(x_{i_k}^T \omega - y_{i_k}) \rangle] \\
&= \mathbb{E}_{i_k} [\langle x_{i_k} x_{i_k}^T \omega, x_{i_k} x_{i_k}^T \omega \rangle - 2 \langle x_{i_k} x_{i_k}^T \omega, x_{i_k} y_{i_k} \rangle + \langle x_{i_k} y_{i_k}, x_{i_k} y_{i_k} \rangle] \\
&= \langle \mathbb{E}_{i_k} [x_{i_k} x_{i_k}^T] \omega, \mathbb{E}_{i_k} [x_{i_k} x_{i_k}^T] \omega \rangle \\
&\quad - 2 \langle \mathbb{E}_{i_k} [x_{i_k} x_{i_k}^T] \omega, \mathbb{E}_{i_k} [x_{i_k} y_{i_k}] \rangle \\
&\quad + \langle \mathbb{E}_{i_k} [x_{i_k} y_{i_k}], \mathbb{E}_{i_k} [x_{i_k} y_{i_k}] \rangle
\end{aligned}$$