

CORRECTION DU TD 5 : THÉORIE DE L'APPRENTISSAGE ET BORNES PAC.

COURS D'APPRENTISSAGE, ECOLE NORMALE SUPÉRIEURE, AUTOMNE 2015

Jean-Baptiste Alayrac
jean-baptiste.alayrac@inria.fr

1. EXERCICE 1 : INÉGALITÉ DE Hoeffding

1.1. **Preuve du lemme d'Hoeffding.** On va d'abord prouver le lemme d'Hoeffding.

1) Montrons d'abord que, presque sûrement, on a $\exp(sX) \leq \frac{b-X}{b-a} \exp(sa) + \frac{X-a}{b-a} \exp(sb)$. Cette inégalité nous permet alors d'écrire une inégalité concernant la transformée de Laplace. L'inégalité voulue se déduit facilement en utilisant la définition de la convexité de la fonction $t \rightarrow \exp(st)$.

Par positivité de l'espérance, on peut passer à l'espérance dans l'inégalité et on en déduit l'inégalité suivante sur la transformée de Laplace :

$$\begin{aligned} \mathbb{E}[\exp(sX)] &\leq \frac{b}{b-a} \exp(sa) + \frac{-a}{b-a} \exp(sb) \\ &\leq (1-\theta) \exp(sa) + \theta \exp(sb) \\ &\leq \exp(sa)(1-\theta + \theta \exp(b-a)) \\ &\leq \exp(-(b-a)\theta)(1-\theta + \theta \exp(s(b-a))). \end{aligned}$$

On pose $u = s(b-a)$. On voit alors naturellement intervenir la fonction que l'on étudie à la question suivante en passant au log.

2) En introduisant $\theta = \frac{-a}{b-a}$ et en considérant la fonction $\phi(u) = -\theta u + \log(1-\theta + \theta \exp(u))$ trouvons un majorant quadratique à ϕ .

Il suffit d'appliquer la majoration de Taylor-Lagrange à la fonction ϕ . On a $\forall u, \phi''(u) = \frac{\exp(u)\theta(1-\theta)}{(1-\theta + \theta \exp(u))^2} \leq \frac{1}{4}$ (on reconnaît en effet que la dérivée seconde est une quantité du type $\frac{ab}{a+b}$ et que ce rapport est majoré par $\frac{1}{4}$). On en déduit donc une majoration de ϕ par $u^2/8$.

3) Par ce qui précède, on en déduit alors le lemme d'Hoeffding car la log-Laplace est majorée par $\exp(s^2(b-a)^2/8)$.

4) On écrit l'inégalité de Markov pour la variable aléatoire *positive* $\exp(S_n - \mathbb{E}[S_n])$ et on obtient :

$$\forall s, t, \mathbb{P}(S_n - \mathbb{E}[S_n] \geq t) = \mathbb{P}(\exp(s(S_n - \mathbb{E}[S_n])) \geq \exp(st)) \leq \frac{\mathbb{E}[\exp(s(S_n - \mathbb{E}[S_n]))]}{\exp(st)}.$$

5) Comme la variable aléatoire $S_n - \mathbb{E}[S_n]$ est bornée et centrée, on peut combiner les réponses des trois questions précédentes, on obtient alors la borne suivante :

$$\forall s, t, \mathbb{P}(S_n - \mathbb{E}[S_n] \geq \epsilon) \leq \exp\left(\frac{\sum_{i=1}^n (b_i - a_i)^2 s}{8} - \epsilon s\right).$$

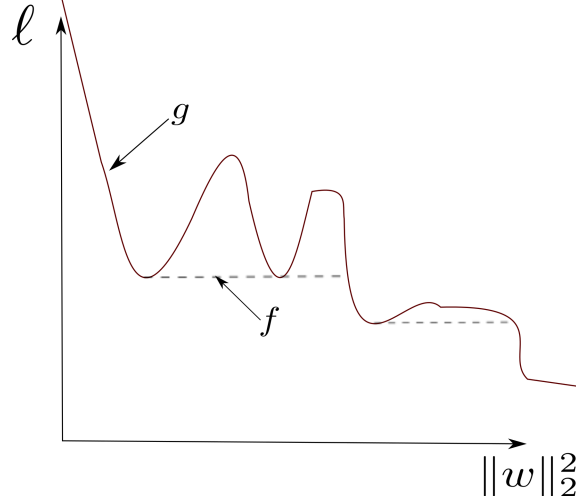


FIGURE 1. Dédution de la frontière de Pareto à partir du graphe de la fonction g .

Remarquons que la somme dans l'exponentielle vient de l'indépendance des variables aléatoires qui permet d'écrire que la transformée de Laplace de la somme est le produit des transformées de Laplace.

Optimiser la borne sur s conduit à prendre $s^* = \frac{4\epsilon}{\sum_i (b_i - a_i)^2}$. La borne correspondante est précisément celle de l'inégalité de Hoeffding.

6) L'inégalité de Chernoff se déduit assez simplement de cette inégalité en considérant des variables aléatoires de Bernoulli (à valeur dans $\{0, 1\}$ donc $\forall i, (b_i - a_i) = 1$) Z_i d'espérance p .

2. EXERCICE 2 : CORRESPONDANCE ENTRE OPTIMISATION SOUS CONTRAINTE ET OPTIMISATION RÉGULARISÉE

7) Soit $\lambda \geq 0$, et w_λ une solution de $(\mathcal{R}_\lambda)$. On pose $R^2 = \|w_\lambda\|_2^2$. On a alors w_λ qui est solution du problème $(\mathcal{C}_R)$. En effet, si $(\mathcal{C}_R)$ avait pour optimum $w' \neq w_\lambda$, w' aurait soit une norme égale à celle de w_λ (auquel cas elle serait solution de $(\mathcal{R}_\lambda)$ avec même valeur objectif que w_λ et donc w_λ est aussi solution de $(\mathcal{C}_R)$), soit une norme plus petite que w_λ (et donc, comme $\|w_\lambda\| > \|w'\|$, elle aurait une valeur dans l'objectif de $(\mathcal{R}_\lambda)$ plus petite que celle de w_λ).

8) $f(R)$ est la valeur de la solution du problème de minimisation $(\mathcal{C}_R)$. Soit $R' > R \geq 0$, on a alors $f(R') \geq f(R)$ car les domaines d'optimisation de $(\mathcal{C}_R)$ et $(\mathcal{C}_{R'})$ sont emboîtés. Le graphe de f se déduit simplement de g car $f(R) = \inf_{R' \leq R} g(R)$, cf Fig. 1.

Remarque 1 utile pour la suite : A tout point $A = (u, v)$ de la frontière de Pareto, on peut faire correspondre des solutions w_A correspondant aux solutions des problèmes $\mathcal{C}_R$.

9) Supposons que ℓ soit convexe, alors les points au delà de la frontière de Pareto forment bien un convexe. On peut s'en convaincre en revenant à la définition de la convexité d'un ensemble. Soit $u = (u_1, u_2), v = (v_1, v_2) \in \mathcal{P}, \lambda \in [0, 1]$. Par définition de $\mathcal{P}$, il existe w_1 et w_2 tels que $u_1 \geq \|w_1\|_2^2, v_1 \geq \|w_2\|_2^2, u_2 \geq \sum_i \ell(w_1^T x_i, y_i)$ et $v_2 \geq \sum_i \ell(w_2^T x_i, y_i)$.

Ainsi, $\lambda u_2 + (1 - \lambda)v_2 \geq \lambda \sum_i \ell(w_1^T x_i, y_i) + (1 - \lambda) \sum_i \ell(w_2^T x_i, y_i) \geq \sum_i \ell((\lambda w_1^T (1 - \lambda)w_2^T)x_i, y_i)$, par définition de la convexité de ℓ . Il en va de même pour u_1 et v_1 . Donc la combinaison convexe de u et v est bien dans $\mathcal{P}$.

Remarque 2 utile pour la suite : L'ensemble des paires de points du plan "faisables" (c'est à

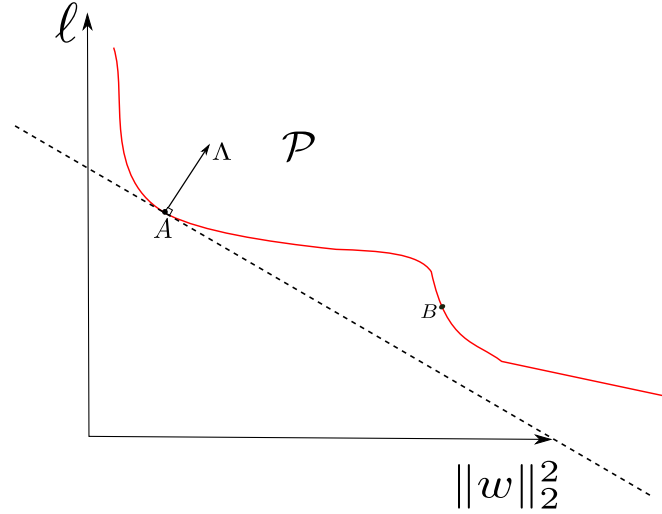


FIGURE 2. La frontière de Pareto, l'ensemble $\mathcal{P}$ et la méthode de scalarisation illustrée graphiquement. Le point A peut être obtenu par scalarisation, mais pas le point B .

dire correspondant à une paire de valeur $(\|w\|^2, \sum_i \ell(w^T x_i, y_i))$ pour un certain w) est inclus dans $\mathcal{P}$.

10) Commençons par noter que le Λ de l'énoncé doit être pensé comme un vecteur colonne. Le problème de minimisation proposé dans cette question est exactement un problème du type $(\mathcal{R}_\lambda)$ et, la question 7 permet de conclure naturellement qu'une solution de ce problème est sur la frontière de Pareto.

Pour faire cette question, nous aurons besoin de la notion de droite support. On notera ici pour deux points du plan A et B , $B - A = \overrightarrow{AB}$.

Pour rappel, une droite $(\mathcal{D})$ de vecteur normal n du plan est appelée droite support à $\mathcal{P}$ en A si :

- $(\mathcal{D})$ passe par A .
- $(\forall B \in \mathcal{P}, (B - A)^T n \geq 0)$ ou $(\forall u \in \mathcal{P}, (B - A)^T n \leq 0)$.

Par exemple, sur la figure 2, l'ensemble $\mathcal{P}$ admet une droite support en A de vecteur normal Λ .

Remarquons que, si en un point A de la frontière de Pareto, Λ est vecteur normal à une droite support, alors n'importe quel w_A associé au point A (Remarque 1) est solution du problème régularisé. En effet, par définition de la droite support, tout point B de $\mathcal{P}$ vérifiera $(B - A)^T \Lambda \geq 0$, or, par la remarque 2, toutes les paires $(\|w\|^2, \ell(w^T x_i, y_i))$ de points "faisables" sont incluses dans $\mathcal{P}$. Ainsi, si on considère une paire "faisable" $(\|w\|^2, \ell(w^T x_i, y_i))$ et le point B associé, on a forcément $(B - A)^T \Lambda \geq 0$, i.e, $\ell(w^T x_i, y_i) + \lambda \|w\|_2^2 \geq \ell(w_A^T x_i, y_i) + \lambda \|w_A\|_2^2$. Ainsi, connaissant la frontière de Pareto, pour résoudre le problème régularisé, il suffit de trouver les points de la frontière de Pareto où Λ est normal à une droite support.

Il reste à s'assurer de l'existence d'un tel point où Λ soit normal à une droite support. Soit une solution w_λ d'un problème régularisé, on lui fait correspondre un certain R correspondant à un problème contraint $\mathcal{C}_R$. Ainsi, ce w_λ correspond à un point de la frontière de Pareto B . On peut regarder la droite dirigée par Λ passant par B . Elle est forcément droite support, car la valeur de $\mathcal{R}_\lambda$ est optimale pour w_λ . Cette situation se visualise à l'aide de la figure 2.

11) De manière réciproque, si un point A est sur la frontière de Pareto, il ne peut être atteint par scalarisation que si l'ensemble $\mathcal{P}$ admet une droite support en A (sinon il y aurait des points de $\mathcal{P}$ de l'autre côté de la droite de vecteur normal Λ passant par A). Ainsi la méthode de scalarisation,

qui revient à considérer les problèmes de type $\mathcal{R}_\lambda$, ne permet de retrouver que certains des points de la frontière de Pareto (i.e, seulement certaines des solutions des problèmes $\mathcal{C}_R$). En un certain sens, la formulation d'optimisation contrainte est, dans le cas général plus riche que la formulation d'optimisation régularisée puisque, contraindre l'optimisation sur une boule permet d'obtenir des solutions qui ne seront jamais atteintes en régularisant. Cela correspond à la situation du point B de la figure 2.

12) Un ensemble convexe de $\mathbb{R}^2$ admet en tout point de sa frontière une droite support (il s'agit d'une généralisation du résultat vu en TD d'analyse convexe). Ainsi, par les questions 9 et 11, tous les points de la frontière de Pareto peuvent, dans le cas convexe, être atteints par la méthode de scalarisation et sont donc solution d'un problème régularisé ($\mathcal{R}_\lambda$).