TD 5 : INÉGALITÉS DE CONCENTRATION

COURS D'APPRENTISSAGE, ECOLE NORMALE SUPÉRIEURE, PRINTEMPS 2013

Rémi Lajugie
remi.lajugie@ens.fr

Résumé. Ce TD/Tp comporte deux grandes parties : dans la première, théorique, on cherchera à trouver un équivalent au maximum de variables aléatoires gaussiennes, dans la seconde on illustrera quelques bornes et inégalités de concentration classiques.

Remarques préliminaires : On rappelle que la densité d'une variable aléatoire gaussienne centrée réduite est donnée simplement par $f(x) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2})$.

On rappelle également que l'on appelle fonction de répartition associée à une variable aléatoire X la fonction F, la fonction $\forall t \in \mathbb{R}, F(t) = \mathbb{P}(X \leq t)$
Il peut aussi être utile de connaître le résultat suivant : pour une variable aléatoire Gaussienne de moyenne μ et de variance σ, la fonction génératrice des moments (ou transformation de Laplace) est donnée par $\mathbb{E}[\exp tX] = \exp(\mu t + \frac{\sigma^2 t^2}{2})$. Aussi, une variable sous gaussienne est une variable dont la transformée de Laplace est majorée par celle d'une gaussienne centrée pour un certain σ, on appelle ce σ^2 le paramètre de sous-gaussiannité.

1. Exercice 1 : Comportement asymptotique de l'espérance du maximum de variables gaussiennes

On commence par considérer $N_1, \dots N_K$, K variables aléatoires sous Gaussiennes de même paramètre v.

1) Montrez qu'on a la majoration suivante :

$$\mathbb{E}[\max_{1 \leq j \leq K} N_j] \leq \sqrt{2v \log(K)}.$$

Indice : On pourra utiliser Jensen ainsi que la majoration, a priori triviale du maximum par la somme. Ce type d'argument s'appelle un argument à la Pisier.

Le reste de cet exercice a pour but de montrer que cette borne est en réalité optimale, au sens où l'espérance du maximum de variables gaussiennes va se comporter asymptotiquement comme le terme de droite de la borne de la question précédente.

On considère ainsi K variables aléatoires **i.i.d** $N_1, \dots, N_K$ suivant des loi normales centrées réduites.
2) En notant la partie positive $(x)_+ = \max(0, x)$, prouvez l'encadrement suivant :

$$\mathbb{E}[(\max N_j)_+] \geq \mathbb{E}[(\max N_j)] \geq \mathbb{E}[(\max N_j)_+] - 1$$

3) On appelle désormais Φ la fonction de répartition de la loi normale centrée réduite, et Ψ_K celle de $(\max N_j)_+$.

a) Que peut on dire de $\Psi_K(0)$?

b) En statistiques, il est très courant d'avoir besoin, non pas de la fonction de répartition mais d'une certaine manière de son "inverse", c'est à dire de la fonction qui, à un niveau de probabilité $\alpha \in [0, 1]$ associe le plus petit $x \in \mathbb{R}$ tel que la probabilité de l'événement $\{t \leq x\} = \alpha$.

Plus formellement, étant donné une fonction de répartition F, on construit son inverse généralisé comme $\forall t \in [0,1], F^{-1}(t) = \inf\{x \in \mathbb{R}, F(x) \geq t\}$. Commencez par vous convaincre que, dans le cas où la fonction de répartition est inversible, cette définition coincide avec celle, classique de l'inverse.

c) Donnez la fonction inverse généralisée de la fonction Ψ_K en fonction de Φ.

4) a) Considérez une variable aléatoire U de loi uniforme sur $[0,1]$, montrez que, pour une variable aléatoire X de fonction de répartition F_X, on a $F_X^{-1}(U)$ qui a même loi que X.

b) On fixe maintenant un niveau $\delta > 0$. A partir de la question précédente, montrez que, pour K assez grand on aura l'inégalité :

$$\mathbb{E}[(\max N_j)_+] \geq (1-\delta)\Phi^{-1}(\delta^{1/K})$$

5) a) En effectuant deux Intégrations par Parties, trouvez un équivalent de $1-\Phi^x$ quand x tend vers $+\infty$.

b) Utilisez l'équivalent précédent pour en trouver un pour $\Phi^{-1}(t)$ quand t tend vers 1.

c) Démontrez que

$$\liminf_K (1-\delta)\Phi^{-1}(\delta^{1/K}) \geq (1-\delta)\sqrt{2\log(K)}$$

d) Conclure.

2. Exercice : Visualisation de bornes et inégalités de concentration

Cet exercice s'effectuera sur machine. Il a pour but de vous aider à mieux visualiser les bornes et les inégalités vues en cours.

6) On commence par visualiser l'équivalent de l'exercice précédent. Avec la fonction `randn` de Matlab, générez des données suivant K gaussiennes indépendantes centrées réduites. On estimera la moyenne empirique du maximum de ces variables aléatoires. Représentez en fonction de K sur un graphique à la fois l'équivalent de l'exercice précédent ainsi que la moyenne empirique que vous calculerez.

7) On considère maintenant des variables aléatoires S_k, sommes de k variables aléatoires distribuées suivant des lois de Bernoulli de paramètre $p = \mathbb{P}(X_i = 1) = 0.25$. Représentez graphiquement la borne de Hoeffding sur la probabilité de déviation de S_n par rapport à sa moyenne. Par des simulations, illustrez cette borne.

8) On considère maintenant un mélange de deux distributions Gaussiennes de moyennes 2 et -2, de variance unité et de paramètre de mélange $\pi_1 = 0.7$ et $\pi_2 = 1 - \pi_1 = 0.3$. Quelle est la moyenne théorique de ce mélange de distributions ? Pour des tailles d'échantillon $n = 100$, représentez la fonction de répartition (ou la fonction quantile) de la déviation en valeur absolue de la moyenne empirique à la moyenne théorique.