Sujet 1 — Probabilités et statistiques

Consignes, notations et rappels

Le but du problème est d'étudier « l'erreur » dans le théorème central limite. La méthode utilisée sera d'abord appliquée dans le cadre habituel d'une suite de variables aléatoires indépendantes et identiquement distribuées (ce que l'on abrégera en i.i.d.), puis on verra dans un cas dépendant que cette technique peut montrer une convergence vers une loi normale.

La première partie est l'étude d'une distance sur un ensemble de lois de probabilité. La deuxième partie est analytique, et établit les propriétés des solutions d'une certaine équation différentielle. La troisième partie utilise ces résultats pour contrôler la distance entre une loi quelconque et la loi normale. Dans la quatrième partie on applique ce résultat au cas classique d'une suite de variables aléatoires i.i.d. La dernière partie étudie une suite où l'hypothèse d'indépendance n'est plus vérifiée.

Liens entre les parties. Les parties sont relativement indépendantes ; en particulier, les deux premières parties, et les 3 premières questions de la cinquième partie, sont indépendantes.

Notations et rappels. Dans tout le problème on fixe un espace probabilisé $(\Omega, \mathcal{F}, \mathbb{P})$ sur lequel toutes les variables aléatoires considérées sont définies. Pour une variable aléatoire X à valeurs réelles, on note μ_X la loi de X. Si X est dans $L^1(\Omega)$, c'est-à-dire si

$$\int_\Omega |X(\omega)| \, d\mathbb{P}(\omega) < \infty,$$

on dit que X est *intégrable* et l'on note $\mathbf{E}[X]$ l'espérance de X :

$$\mathbf{E}[X] = \int_\Omega X(\omega) d\mathbb{P}(\omega) = \int_\mathbb{R} x d\mu_X(x).$$

Si X^2 est intégrable, on note $\mathbf{Var}(X)$ la variance de X :

$$\mathbf{Var}(X) = \mathbf{E}\left[X^2\right] - \mathbf{E}[X]^2 .$$

On note $\mathcal{M}$ l'ensemble des mesures de probabilité sur $\mathbb{R}$ qui ont un premier moment, c'est-à-dire l'ensemble des mesures de probabilité μ telles que $\int_\mathbb{R} |x| \, d\mu(x) < \infty$, ou encore l'ensemble des lois de variables aléatoires intégrables.

Si $(\mu_n)_{n \in \mathbb{N}}$ est une suite de mesures de probabilité sur $\mathbb{R}$, et que μ est une mesure de probabilité, on dit que μ_n *converge étroitement* vers μ si, pour toute fonction $f : \mathbb{R} \to \mathbb{R}$ continue bornée,

$$\int_\mathbb{R} f d\mu_n \xrightarrow{n \to \infty} \int_\mathbb{R} f d\mu.$$

On admettra que, si la convergence ci-dessus a lieu pour toute fonction f Lipschitzienne bornée, alors μ_n converge étroitement vers μ.

On rappelle qu'une suite de variables aléatoires (X_n) *converge en loi* vers une variable aléatoire X si et seulement si la suite de lois $(\mu_{X_n})_{n\in\mathbb{N}}$ converge étroitement vers la loi de X.

On dit que la suite $(\mu_n)_{n\in\mathbb{N}}$ est *tendue* si pour tout $\epsilon > 0$ il existe un compact K_ϵ de $\mathbb{R}$ tel que :

$$\forall n \geq 0, \quad \mu_n(K_\epsilon) \geq 1 - \epsilon.$$

On rappelle le résultat de compacité suivant : si une suite de mesures de probabilité $(\mu_n)_{n\in\mathbb{N}}$ est tendue, alors il existe une sous-suite qui converge étroitement vers une mesure de probabilité.

Enfin, pour toute fonction bornée f de $\mathbb{R}$ dans $\mathbb{R}$, on note $\|f\|_\infty = \sup_{x\in\mathbb{R}}\{|f(x)|\}$.

I Une distance entre lois de probabilité

Dans cette partie les questions 4 à 6 étudient des exemples et sont indépendantes de la suite du problème.

I. 1°) Soit $\mathcal{H}$ l'ensemble des fonctions Lipschitziennes :

$$\mathcal{H} = \{h : \mathbb{R} \to \mathbb{R}, \exists C \in \mathbb{R}_+, \forall (x,y) \in \mathbb{R}^2, |h(x) - h(y)| \leq C\,|x-y|\}.$$

Pour toute $h \in \mathcal{H}$, on notera

$$\|h\|_{Lip} = \sup_{(x,y)\in\mathbb{R}^2; x\neq y} \frac{|h(x) - h(y)|}{|x-y|}.$$

la semi-norme de Lipschitz de h. Montrer que si μ est dans $\mathcal{M}$ et h dans $\mathcal{H}$, la fonction h est intégrable pour μ :

$$\int_{\mathbb{R}} |h(x)|\, d\mu(x) < \infty.$$

I. 2°) Pour deux mesures μ et ν dans $\mathcal{M}$, on définit

$$d_{\mathcal{W}}(\mu,\nu) = \sup_{h\in\mathcal{H}, \|h\|_{Lip}\leq 1} \left| \int_{\mathbb{R}} h(x) d\mu(x) - \int_{\mathbb{R}} h(x) d\nu(x) \right|.$$

Montrer que l'on définit ainsi une distance sur $\mathcal{M}$.

I. 3°) Soit (X,Y) un couple de variables aléatoires réelles. On suppose que la loi marginale de X est μ et celle de Y est ν, et que μ et ν sont dans $\mathcal{M}$. Montrer que :

$$d_{\mathcal{W}}(\mu,\nu) \leq \mathbf{E}\left[|X - Y|\right].$$

I. 4°) Pour tout $a \in \mathbb{R}$ on note δ_a la mesure de Dirac en a. Montrer que $d_{\mathcal{W}}(\delta_a, \delta_b) = |a - b|$.

I. 5°) Pour tout $m \in \mathbb{R}$ et tout $s > 0$, notons $\mathcal{N}(m,s)$ la loi normale de moyenne m et d'écart-type s, dont la densité est donnée par :

$$\frac{1}{\sqrt{2\pi}\, s} \exp\left(-\frac{(x-m)^2}{2s^2}\right).$$

Montrer que :

$$\forall s > 0, \forall (a,b) \in \mathbb{R}^2, \quad d_{\mathcal{W}}\left(\mathcal{N}(a,s), \mathcal{N}(b,s)\right) = |b-a|.$$

Dans un des sens on pourra utiliser le fait que, si $X \sim \mathcal{N}(m,s)$, alors pour tout couple $(\alpha, \beta) \in \mathbb{R}^2$, $\alpha X + \beta$ suit une loi normale (dont on rappellera la moyenne et l'écart-type).

I. 6°) Montrer que

$$\forall s \in \mathbb{R}_+^\star, \quad d_{\mathcal{W}}\left(\mathcal{N}(0,s), \mathcal{N}(0,1)\right) = \sqrt{\frac{2}{\pi}}\,|1-s|.$$

I. 7°) Soit $(\mu_n)_{n\in\mathbb{N}}$ une suite de mesures dans $\mathcal{M}$, et μ une mesure de $\mathcal{M}$, telles que :

$$d_{\mathcal{W}}(\mu_n, \mu) \xrightarrow[n\to\infty]{} 0.$$

Montrer que $(\mu_n)_{n\in\mathbb{N}}$ converge étroitement vers μ.

I. 8°) Soit $(\mu_n)_{n\in\mathbb{N}}$ une suite d'éléments de $\mathcal{M}$. On suppose que $(\mu_n)_{n\in\mathbb{N}}$ est une suite de Cauchy pour la distance $d_{\mathcal{W}}(\cdot,\cdot)$. Montrer que la suite d'intégrales $\int |x|\, d\mu_n(x)$ converge. Montrer qu'il existe une mesure μ et une sous-suite (notée $(\mu_{\phi(n)})_{n\in\mathbb{N}}$) telles que $(\mu_{\phi(n)})_{n\in\mathbb{N}}$ converge étroitement vers μ. Montrer que μ est nécessairement dans $\mathcal{M}$.

Indication. Pour le dernier point, on pourra introduire, pour tout $M > 0$, la fonction $f_M : x \mapsto \min(|x|, M)$.

I. 9°) Montrer que $\mathcal{M}$ est complet pour la distance $d_{\mathcal{W}}(\cdot,\cdot)$.

Indication. On pourra introduire pour tout $M > 0$ et tout $h \in \mathcal{H}$ la fonction h_M définie par

$$h_M(x) = \begin{cases} h(x) & \text{si } h(x) \in [-M, M], \\ M & \text{si } h(x) \geq M, \\ -M & \text{si } h(x) \leq -M, \end{cases}$$

et montrer qu'elle est Lipshitzienne bornée.

II Une équation différentielle

II. 1°) Soit $\phi(x)$ la densité, $\Phi(x)$ la fonction de répartition et $\Psi(x)$ la « fonction de survie » de la loi normale $\mathcal{N}(0,1)$:

$$\begin{aligned}
\phi(x) &= \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \\
\Phi(x) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-y^2/2} dy, \\
\Psi(x) &= 1 - \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{x}^{\infty} e^{-y^2/2} dy.
\end{aligned}$$

Montrer les deux inégalités suivantes pour $x > 0$:

$$\Psi(x) \leq \frac{1}{x}\phi(x),$$
$$e^{x^2/2}\int_x^\infty e^{-y^2/2}dy = \frac{\Psi(x)}{\phi(x)} \leq \sqrt{\frac{\pi}{2}}.$$

II. 2°) Déterminer toutes les fonctions $f : \mathbb{R} \to \mathbb{R}$ dérivables telles que :

$$\forall x \in \mathbb{R}, \quad f'(x) - xf(x) = 0.$$

II. 3°) Soit h une fonction dans $\mathcal{H}$ et Z une variable aléatoire de loi $\mathcal{N}(0,1)$. Justifier que $h(Z)$ est d'espérance finie. On cherche à résoudre l'équation différentielle (d'inconnue f) :

$$\forall x \in \mathbb{R}, \quad f'(x) - xf(x) = h(x) - \mathbf{E}\left[h(Z)\right]. \tag{1}$$

On définit une fonction f_h par :

$$f_h(x) = -e^{x^2/2}\int_x^\infty e^{-y^2/2}(h(y) - \mathbf{E}\left[h(Z)\right])dy.$$

Justifier cette définition. Montrer que f_h est la seule solution bornée de l'équation (1), et qu'on a :

$$\forall h \in \mathcal{H}, \quad \|f_h\|_\infty \leq 2\,\|h\|_{Lip}\,.$$

Indication. On pourra établir

$$|h(y) - \mathbf{E}\left[h(Z)\right]| \leq \|h\|_{Lip}\left(|y| + \sqrt{2/\pi}\right).$$

Pour traiter le cas $x < 0$ on pourra utiliser l'expression

$$f_h(x) = e^{x^2/2}\int_{-\infty}^x e^{-y^2/2}(h(y) - \mathbf{E}\left[h(Z)\right])dy.$$

Dans la suite, **on admettra** que l'on peut de façon similaire montrer que f_h est de classe $\mathcal{C}^2$ et établir les bornes :

$$\forall h \in \mathcal{H}, \quad \begin{cases} \|f_h'\|_\infty & \leq \sqrt{2/\pi}\,\|h\|_{Lip}\,, \\ \|f_h''\|_\infty & \leq 2\,\|h\|_{Lip}\,. \end{cases}$$

III Distance à la loi normale

Soit $\mathcal{F}$ l'ensemble des fonctions de classe $\mathcal{C}^2$, telles que f, f' et f'' sont bornées. Pour une fonction $f \in \mathcal{F}$ et une variable aléatoire X intégrable, on définit

$$\mathcal{S}(f, X) = \mathbf{E}\left[Xf(X) - f'(X)\right]. \tag{2}$$

III. 1°) Justifier que cette définition a un sens.

III. 2°) Si Z suit la loi normale standard $\mathcal{N}(0,1)$, montrer que

$$\forall f \in \mathcal{F}, \quad \mathcal{S}(f, Z) = 0.$$

III. 3°) En utilisant la question I. 2°) et les résultats de la partie II, montrer que, pour X une variable aléatoire intégrable quelconque et Z une variable aléatoire de loi $\mathcal{N}(0,1)$, on a :

$$d_{\mathcal{W}}\left(\mu_X, \mu_Z\right) \leq \sup\left\{|\mathcal{S}(f,X)|\,;\, f \in \mathcal{F}, \|f'\|_\infty \leq \sqrt{2/\pi}, \|f''\|_\infty \leq 2\right\}.$$

III. 4°) Si X est telle que $\mathcal{S}(f,X) = 0$ pour tout $f \in \mathcal{F}$, que peut-on dire de X ?

IV Au-delà du TCL

Soit $(X_k)_{k\in\mathbb{N}}$ une suite de variables indépendantes, de même loi, centrées et de variance 1 :

$$\mathbf{E}\left[X_k\right] = 0, \quad \mathbf{Var}\left(X_k\right) = 1.$$

On suppose de plus que les X_k admettent un moment d'ordre 3 :

$$\mathbf{E}\left[|X_1|^3\right] < \infty.$$

IV. 1°) Que disent la loi des grands nombres et le théorème central limite appliqués à la suite (X_k) ?

On fixe un entier n. Pour $1 \leq k \leq n$, on note :

$$W = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} X_j \qquad \text{et} \qquad W_k = W - \frac{X_k}{\sqrt{n}}.$$

IV. 2°) Soit f une fonction de $\mathcal{F}$. Montrer qu'il existe une fonction g, bornée par $\|f''\|_\infty/2$, telle que :

$$\forall (x,y) \in \mathbb{R}^2, \quad f(x) - f(y) = f'(y)(x-y) + g(x,y)(x-y)^2. \tag{3}$$

IV. 3°) On rappelle que $\mathcal{S}(f,W)$ est défini par l'équation (2). Montrer que

$$\begin{aligned}\mathcal{S}(f,W) = \frac{1}{\sqrt{n}}\mathbf{E}\left[\sum_{k=1}^{n} X_k f(W_k)\right] + \mathbf{E}\left[\frac{1}{\sqrt{n}}\sum_{k=1}^{n} X_k (f'(W_k)(W-W_k)) - f'(W)\right] \\ + \frac{1}{\sqrt{n}}\mathbf{E}\left[\sum_{k=1}^{n} X_k (W-W_k)^2 g(W,W_k)\right].\end{aligned}$$

IV. 4°) On note (dans l'ordre) A ,B et C les trois termes du membre de droite de l'équation ci-dessus. Montrer que A est nul, puis que :

$$B = \frac{1}{n}\mathbf{E}\left[\sum_{k=1}^{n}(f'(W_k) - f'(W))\right]$$
$$C = \frac{1}{n^{3/2}}\mathbf{E}\left[\sum_{k=1}^{n} X_k^3 g(W, W_k)\right].$$

IV. 5°) En déduire que

$$|\mathcal{S}(f, W)| \leq \frac{1}{\sqrt{n}} \|f''\|_\infty \left(\mathbf{E}\left[|X_1|\right] + \frac{1}{2}\mathbf{E}\left[\left|X_1^3\right|\right]\right).$$

En utilisant les résultats de la partie III, en déduire que W converge en loi quand n tend vers l'infini et donner sa limite. Que pensez-vous de ce résultat ?

V Un jeu de pile ou face

Dans cette partie on considère un jeu répété de pile ou face, modélisé par une suite $(X_k)_{k\in\mathbb{N}}$ de variables aléatoires i.i.d. de loi de Bernoulli de paramètre 1/2 :

$$\forall k \in \mathbb{N}, \quad \mathbf{P}\left[X_k = 0\right] = \mathbf{P}\left[X_k = 1\right] = \frac{1}{2}.$$

On s'intéresse aux suites de deux symboles consécutifs. Pour tout $k \geq 1$, on pose

$$D_k = \begin{cases} 1 \text{ si } X_k = 0, X_{k+1} = 0, \\ 2 \text{ si } X_k = 0, X_{k+1} = 1, \\ 3 \text{ si } X_k = 1, X_{k+1} = 0, \\ 4 \text{ si } X_k = 1, X_{k+1} = 1. \end{cases}$$

V. 1°) Montrer que la suite $(D_k)_{k\in\mathbb{N}}$ est une chaîne de Markov. Donner sa matrice de transition.

V. 2°) Montrer qu'il existe une unique probabilité invariante pour $(D_k)_{k\in\mathbb{N}}$; la calculer.

V. 3°) On joue au jeu suivant : pour chaque temps k, on touche 3 euros si $D_k = 1$, sinon on perd 1 euro. On note Y_k le gain (algébrique) correspondant. En écrivant le gain jusqu'à l'étape n comme une somme :

$$G^{(n)} = \sum_{k=1}^{n} Y_k,$$

montrer que $\frac{G^{(n)}}{n}$ tend presque sûrement vers 0.

Dans la suite on adapte l'approche de la partie précédente pour établir un résultat du type « théorème central limite » pour $G^{(n)}$.

V. 4°) Pour $k,l \in (\mathbb{N}^*)^2$, les variables Y_k et Y_l sont-elles indépendantes ? Calculer leur covariance.

V. 5°) On fixe n et on définit les quantités suivantes :

$$\sigma_n = \sqrt{\mathbf{Var}\left(\sum_{k=1}^{n} Y_k\right)} \qquad\qquad V = \frac{G^{(n)}}{\sigma_n} = \frac{1}{\sigma_n}\sum_{k=1}^{n} Y_k$$

$$V_k = V - \frac{1}{\sigma_n} \sum_{\substack{j:\, 1\le j\le n \\ |k-j|\le 1}} Y_j \qquad\qquad A = \sum_{k=1}^{n} Y_k(V - V_k).$$

Montrer que $\mathbf{E}[A] = \sigma_n$ et que $\sigma_n^2 = 5n-2$.

V. 6°) Montrer que pour toute $f \in \mathcal{F}$, on a

$$\mathbf{E}[Vf(V) - f'(V)] = \mathbf{E}\left[f'(V)\left(\frac{A}{\sigma_n} - 1\right)\right] - \frac{1}{\sigma_n}\mathbf{E}\left[\sum_{k=1}^{n} Y_k(V - V_k)^2 g(V, V_k)\right]$$

où g est définie en fonction de f comme dans la question IV. 2°).

V. 7°) En utilisant le fait que les Y_k sont bornées, montrer qu'il existe une constante C telle que pour tout n,

$$\mathbf{Var}(A) \le \frac{Cn}{\sigma_n^2}.$$

En déduire qu'il existe C' (ne dépendant pas non plus de n) telle que

$$|\mathcal{S}(f,V)| \le C' \|f'\|_\infty \frac{\sqrt{n}}{\sigma_n^2} + C' \|f''\|_\infty \frac{n}{\sigma_n^3}.$$

V. 8°) En déduire que $\frac{G^{(n)}}{\sqrt{n}}$ converge en loi et donner sa limite.

⋆ ⋆ ⋆ **Fin du sujet** ⋆ ⋆ ⋆