TD 2 : Statistique exhaustive et Loi de Zipf

Exercice (Une statistique exhaustive)

Soit une séquence de variables aléatoires indépendantes et identiquement distribuées (i.i.d.) $X_1, X_2, \ldots, X_n$, où chaque $X_i \in \{0, 1\}$ représente un lancer de pièce. La probabilité d'obtenir un 1 est donnée par un paramètre inconnu $\theta = \Pr(X_i = 1)$.

1. Remarquez que la probabilité d'obtenir une séquence particulière avec exactement k résultats de 1 est la même pour toutes ces séquences et calculer cette probabilité coditionnée à $T(X_1, \ldots, X_n) = k$.
2. Montrer que T est une statistique exhaustive.

Exercice (Loi de Zipf)

Dans cet exercice nous nous proposons d'explorer un modèle de Mandelbrot pour expliquer la loi de Zipf par l'optimisation d'un ratio coût/information.

Pour observer la loi de Zipf, commençons par classer les fréquences d'apparition des mots notée $p_1 \geq p_2 \geq \ldots$. Considérons un mot w_k de fréquence p_k dans le vocabulaire. Mandelbrot fait l'hypothèse qu'un langage cherche à maximiser l'information moyenne transmise par chaque mot, c'est à dire son entropie $H = -\sum_k p_k \log_2(p_k)$.
Par ailleurs, mobiliser un mot a un coût — notamment en mémoire et en prononciation — que l'on note C_k (et dont l'expression sera précisées plus tard). Le coût moyen du langage est alors donnée par $C = \sum_k p_k C_k$. Mandelbrot propose donc d'étudier un modèle de langage où l'on tente de minimiser le ratio coût/information, C/H.

Méthode des multiplicateurs de Lagrange La méthode des multiplicateurs de Lagrange est une technique utilisée en optimisation pour trouver les extrémums (maximums ou minimums) d'une fonction sous contrainte. Supposons que nous voulions maximiser ou minimiser une fonction $f(x_1, x_2, \ldots, x_d)$ pour un vecteur sous une contrainte donnée par $g(x_1, x_2, \ldots, x_d) = c$. On introduit un nouveau paramètre λ, appelé multiplicateur de Lagrange, qui permet de transformer notre problème en un problème d'optimisation sans contrainte pour la fonction suivante.

$$L(x_1, \ldots, x_d, \lambda) = f(x_1, \ldots, x_d) + \lambda \cdot g(x_1, \ldots, x_d)$$

Une condition nécessaire pour être un point extremal sous contrainte est alors donnée par le fait que les dérivées partielles de L par rapport à chaque variable x_i doivent être nulles :

$$\frac{\partial L}{\partial x_1} = 0, \ \ldots, \ \frac{\partial L}{\partial x_d} = 0.$$

1. En utilisant la méthode des multiplicateurs de Lagrange, montrer que la distribution de fréquence qui minimise le ratio coût/information $C^* = C/H$ sous la contrainte $\sum_k p_k = 1$ est de la forme $p_k = \lambda' 2^{-HC_k/C}$, où λ' est une constante de normalisation.
2. Quel doit être la forme de C_k pour que p_k suive une loi de puissance $p_k \propto k^{-B}$. Peut-on interpréter cette forme en coût d'accès mémoire des mots dans le cerveau ?

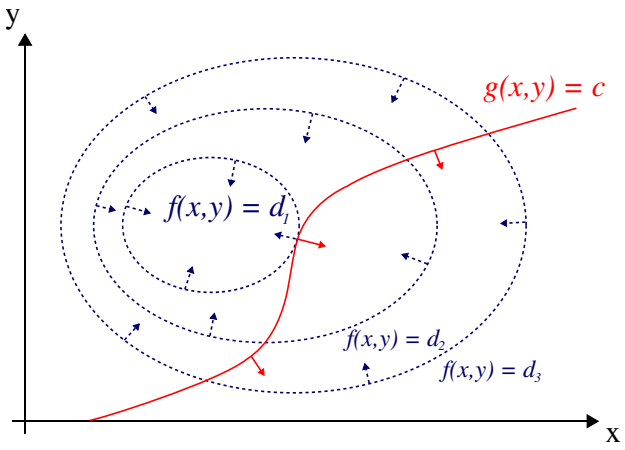

FIGURE 1 – Illustration de la méthode dans un espace bidimensionnel.

3. Calculez la constante B en fonction de H et C. En déduire que C/C_0 ne dépend pas de C_0.
4. Soit D la constante telle que $p_k = \dfrac{1}{D} \cdot k^{-B}$. Déduire de la formule sur B précédente la valeur de D.
5. Pour quelle valeurs de B peut-on approcher cette valeur de D ?
   Mandelbrot propose un modèle un peu plus général avec $C_k = C_0 \log(k + k_0)$. On pose

$$\zeta(s, q) = \sum_0^\infty (n + q)^{-s}.$$

6. Montrer que dans ce modèle,

$$p_k = \frac{1}{\zeta(B, 1 + k_0)} (k + k_0)^{-B}.$$

7. Calculer $\frac{\partial}{\partial s}\zeta(s, q)$ que l'on notera $\zeta'(s, q)$.
8. Exprimer $\frac{C}{C_0}$ et H en fonction de ζ et ζ'.
9. En utilisant la relation entre B et C/C_0, trouver une condition sur B.
10. Montrer que quand $k_0 \to 0$, $B \to \infty$. Puis que pour $k_0 \to \infty$, $B \to 1$. Estimer cette dernière vitesse de convergence.