PHYS-338: Physique Statistique **2025-2026**

Lecture 1: Un peu de probabilités

Prof.: Florent Krzakala

1.1 Rappels et Définitions

1.1.1 Probabilités et distributions

On définit une variable aléatoire X comme une variable dont la valeur dépend du hasard : elle peut prendre une valeur différente à chaque essai.

- Pour X une variable aléatoire discrète, chaque issue a une probabilité associée $p_i \in (0,1)$ correspondant à la probabilité que la variable aléatoire prenne la valeur de l'issue.

Exemple 1 *Dans le cas d'un lancer de dé on peut écrire:*

$$\mathbb{P}(X=i) = \tfrac{1}{6}, \quad i=1,2,3,4,5,6 \ ; \qquad \textit{avec} \qquad \sum_{i=1}^{6} p_i = \sum_{i=1}^{6} \mathbb{P}(X=i) = 1. \qquad \blacksquare$$

- Pour X une variable aléatoire continue il est impossible de déterminer une probabilité pour une valeur précise[1]. Néanmoins il est possible d'exprimer la probabilité de trouver la variable aléatoire dans un certain intervalle : elle est définie par la fonction de densité de probabilité.

Définition 1.1 *(Fonction de densité de probabilité p_X (p.d.f.))*
Fonction mathématique non-négative décrivant la probabilité qu'une variable aléatoire continue X prenne une certaine valeur x ou plus précisément appartienne à l'intervalle $[x, dx]$ pour dx un nombre réel positif infiniment petit. On définit $p_X(x)$ t.q.

$$p_X(x)dx = \mathbb{P}(X \in [x, x+dx]) \qquad \textit{avec } p_X \textit{ normalisée} \qquad \int_{-\infty}^{\infty} p_X(x)dx = 1. \tag{1.1}$$

Il est également possible de décrire le cas des variables aléatoires discrètes avec une p.d.f.
En reprenant l'exemple du jet de dé :

$$p_X(x) = \frac{1}{6}\sum_{i=1}^{6} \delta(i-x). \tag{1.2}$$

R Le facteur de $\frac{1}{6}$ devant la somme de l'Eq.(1.2) correspond à la normalisation de la fonction.

[1] À proprement parler, chaque point a probabilité nulle, voir Prop. 1.2 ci-après.

Proposition 1.2 *On peut élargir la définition précédente à un intervalle choisi [a,b], avec a, b* $\in \mathbb{R}$. *La probabilité que X prenne une valeur comprise dans cet intervalle est alors :*

$$\mathbb{P}(X \in [a,b]) = \int_a^b p_X(x)dx. \tag{1.3}$$

Exemple 2 *Voici un exemple de plusieurs distributions de base avec la p.d.f correspondante :*

1. *Distribution de Dirac : voir Fig.1(a)*
   Pour une moyenne c et une variance de 0, la p.d.f. s'écrit

   $$p_X = \delta(x-c) \tag{1.4}$$

   Ainsi dans le cas de variables aléatoires discrètes, le graphe de la p.d.f. correspond à un ensemble de pics placés selon les valeurs possibles de l'expérience.

2. *Loi normale (ou Gaussienne) : voir Fig.1(b)*
   Pour une moyenne μ *et une variance* Δ, *la p.d.f. s'écrit*

   $$p_X(x) = \mathcal{N}(x;\mu,\Delta) = \frac{1}{\sqrt{2\pi\Delta}} e^{-\frac{(x-\mu)^2}{2\Delta}} \tag{1.5}$$

   R *Noter que pour* $\Delta \to 0$ *la Gaussienne tends vers une distribution de Dirac centrée en* μ.

3. *Loi uniforme : voir Fig.1(c)*

   $$p_X(x) = \mathcal{U}(x;a,b) = \begin{cases} \frac{1}{b-a} & si\ a \le x \le b \\ 0 & sinon. \end{cases}$$

4. *Loi exponentielle : voir Fig.1(d)*

   $$p_X(x) = \frac{1}{\lambda} e^{-\frac{x}{\lambda}} \tag{1.6}$$

   avec $x \ge 0$ *et* $\lambda \ge 0$ *la moyenne.*

   R *En radioactivité,* λ *est nommée constante de temps (ou durée de vie moyenne) .*

■

Définition 1.3 *(Espérance) L'espérance d'une variable aléatoire X est définie*

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x p_X(x)dx. \tag{1.7}$$

Le cas discret est défini de manière analogue avec une somme. Pour une fonction g mesurable, on a

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) p_X(x)dx. \tag{1.8}$$

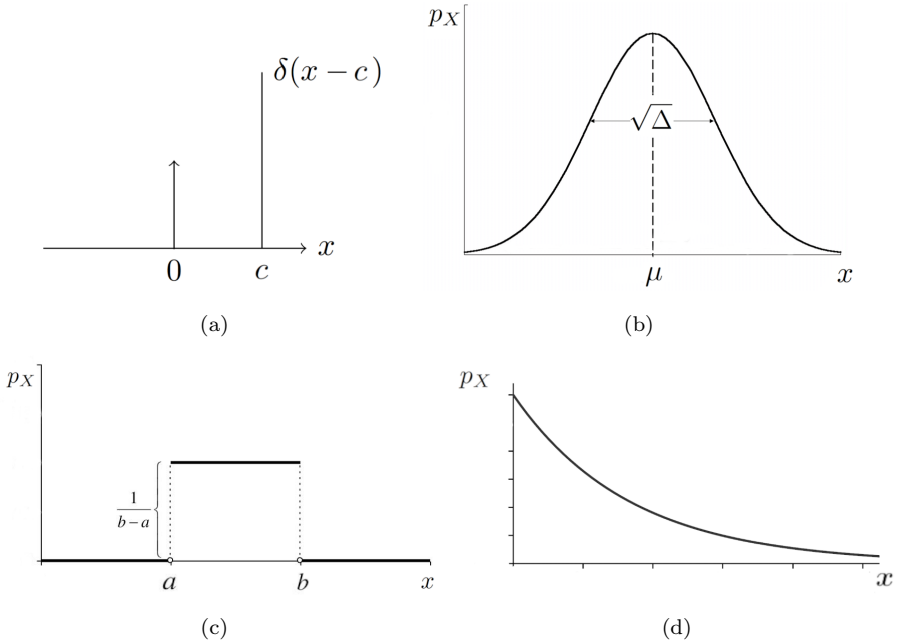

Figure 1.1: Distributions: (a) de Dirac, (b) Normale, (c) Uniforme et (d) Exponentielle.

Définition 1.4 *(Moment d'ordre n)*

$$\mu_n = \mathbb{E}[X^n] = \int_{-\infty}^{\infty} x^n p_X(x) dn \tag{1.9}$$

le moment d'ordre $n = 1$: $\mu_1 = \mu = \int_{-\infty}^{\infty} x p_X(x) dx$ est plus communément appelé la moyenne.

Définition 1.5 *Soient deux variables X et Y:*

1. *Loi jointe: $\mathbb{P}(X, Y) \equiv$ probabilité que X et Y soient vrais.*

2. *Loi conditionnelle: $\mathbb{P}(X|Y) \equiv$ probabilité que X soit vrai quand Y est vrai*
   $\Rightarrow \mathbb{P}(X, Y) = \mathbb{P}(X|Y)\mathbb{P}(Y)$

Définition 1.6 *(Covariance)*

$$\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = Cov(X, Y) \tag{1.10}$$

R (X et Y indépendants) $\Leftrightarrow \mathbb{P}(X, Y) = \mathbb{P}(X)\mathbb{P}(Y)$.

R Attention, (X et Y indépendants) $\Rightarrow$ Cov$(X, Y) = 0$, mais la réciproque est fausse en général. L'équivalence tient pour X, Y gaussiennes.

1.1.2 Propriétés de base

Proposition 1.7 *(Changement de variable)*
Soit X distribuée selon p_X et une autre variable Y t.q. $Y = g(X)$, on trouve:

$$p_Y(y) = \int_{-\infty}^{\infty} p_X(x)\delta(y - g(x))dx \tag{1.11}$$

En utilisant le résultat de la fonction de Dirac $\forall x_i$ t.q. $f(x_i) = 0$

$$\delta(f(x)) = \frac{\delta(x - x_i)}{|f'(x_i)|}, \tag{1.12}$$

on obtient alors $\forall x_i$ solutions de $g(x_i) = y$:

$$p_Y(y) = \sum_i \int dx p_X(x) \frac{\delta(x - x_i)}{|g'(x_i)|} = \sum_i \frac{1}{|g'(x_i)|} p_X(x_i). \tag{1.13}$$

Pour g une fonction monotone, la p.d.f. de Y peut donc être écrite :

$$p_Y(y) = p_X(x) \left| \frac{dx}{dy} \right| \tag{1.14}$$

tirée également de l'égalité des probabilités $p_Y(y)dy = p_X(x)dx$.

R Attention à ne pas oublier le jacobien $|\frac{dx}{dy}|$ dans les calculs!

Exemple 3 *Soit une variable aléatoire $X \sim \mathcal{U}[0, 1]$. On définit $Y = -\ln(X)$ $(\Leftrightarrow \quad e^{-y} = x)$*

La distribution de Y est donc donnée par $p_Y(y) = p_X(x)|\frac{dx}{dy}| = |\frac{dx}{dy}| = e^{-y}$. ∎

1.2 Bornes de bases

Proposition 1.8 *(Inégalité de Markov) - Soit X une variable aléatoire non-négative, et $a > 0$, alors*

$$\mathbb{P}[X \geq a] \leq \frac{\mathbb{E}[X]}{a} \tag{1.15}$$

Preuve 1

$$\begin{aligned} \mathbb{P}(X \geq a) &= \int_a^{+\infty} p_X(x)dx \\ &\leq \int_a^{+\infty} \frac{x}{a} p_X(x)dx \\ &\leq \frac{1}{a} \int_0^{+\infty} x p_X(x)dx \\ &= \frac{\mathbb{E}[X]}{a}, \end{aligned} \tag{1.16}$$

en utilisant que $x \geq a$, et ensuite que $xp_X(x) \geq 0$.

Proposition 1.9 *(Inégalité de Chebyshev) - Soit X une variable aléatoire qui admet une variance non-nulle σ^2 ainsi qu'une moyenne. Alors pour tout $k > 0$,*

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq k\sigma) \leq \frac{1}{k^2}. \tag{1.17}$$

Preuve 2

$$\begin{aligned} &\textit{Considérons la variable aléatoire } (X - \mathbb{E}[X])^2 \textit{ de variance } \sigma^2 \\ &\mathbb{P}(|X - \mathbb{E}[X]| \geq k\sigma) = \mathbb{P}((X - \mathbb{E}[X])^2 \geq k^2\sigma^2) \\ &\leq \frac{\mathbb{E}[(X - \mathbb{E}[X])^2]}{k^2\sigma^2} = \frac{\mathbb{E}[X^2] - \mathbb{E}[X]^2}{k^2\sigma^2} = \frac{\sigma^2}{k^2\sigma^2} \textit{ en utilisant l'inégalité de Markov} \\ &= \frac{1}{k^2} \end{aligned} \tag{1.18}$$

R Cette inégalité permet de borner la probabilité qu'une variable aléatoire dévie de la moyenne, et ce en utilisant la variance de cette variable.

Exemple 4 *Soit $Y_n = \frac{1}{n}\sum_{i=1}^n X_i$ la moyenne empirique, avec X_i des variables aléatoires i.i.d. d'espérance μ et de variance Δ.*

Alors $\mathbb{E}[Y_n] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[X_i] = \frac{1}{n} \cdot n\mu = \mu$ tandis que $Var[Y_n] = \frac{1}{n^2} \cdot n\Delta = \frac{\Delta}{n}$. Appliquant l'inégalité de Chebyshev, nous obtenons $\mathbb{P}(|Y_n - \mu| \geq \epsilon) \leq \frac{\Delta}{n\epsilon^2}$. Cette probabilité tend vers 0 lorsque $n \to \infty$. Ce résultat mène à la loi des grands nombres, qui est énoncée maintenant. ∎

Définition 1.10 *(Convergence en probabilité) - Soient $X, X_1, X_2, ...$ des variables aléatoires. On dit que X_n tend vers X en probabilité, $X_n \xrightarrow{P} X$, si pour tout $\epsilon > 0$,*

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| > \epsilon) = 0 \tag{1.19}$$

Définition 1.11 *(Convergence en distribution) - Soient $X, X_1, X_2, ...$ des variables aléatoires, ayant une fonction de répartition $F_X, F_{X_1}, F_{X_2}, ...$ On dit que X_n tend vers X en distribution, $X_n \xrightarrow{D} X$, si*

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x) \tag{1.20}$$

en tout x où $F_X(x)$ est continue.

Théorème 1.12 *(Loi des grands nombres) - Soit (X_n) une suite de variables aléatoires i.i.d d'espérance μ. Alors la moyenne empirique $Y_n = \frac{1}{n}\sum_{k=1}^n X_k$ converge en probabilité vers l'espérance μ*

$$Y_n \xrightarrow{P} \mu \tag{1.21}$$

Ainsi, $\forall \epsilon > 0$, $\mathbb{P}[|X - \mu| \geq \epsilon] \to 0$.

1.3 Fonctions génératrices

Définition 1.13 *(Fonction génératrice des moments, MGF)*

Soit X une variable aléatoire de densité $p_X(x)$ et de support S. La MGF $M_X(t)$ est définie comme

$$M_X(t) = \mathbb{E}[e^{tX}] = \int_S dx p_X(x) e^{tx}. \tag{1.22}$$

Le cas discret est analogue en remplaçant l'intégrale par une somme.

R La MGF est la transformée de Laplace de la densité, et n'existe pas $\forall t$ en général (il faut que la densité décroisse "assez vite" quand $x \to \infty$).

Par construction la proposition suivante est vraie :

Proposition 1.14

$$\left.\frac{\partial^n M_X(t)}{\partial t^n}\right|_{t=0} = \mu_n = \mathbb{E}[X^n], \tag{1.23}$$

où μ_n est le n-ieme moment.

On peut définir de manière analogue la fonction caractéristique en utilisant la transformée de Fourier et non celle de Laplace :

Définition 1.15 *(Fonction caractéristique)*

Soit X une variable aléatoire de densité $p_X(x)$ et de support S. La fonction caractéristique $\phi_X(t)$ est définie comme

$$\phi_X(t) = \mathbb{E}[e^{itX}] = \int_S dx p_X(x) e^{itx}. \tag{1.24}$$

R Dans ce cas, la transformée est toujours bien définie pour l'intégrale de Lebesgue puisque e^{itx} est dominé en module par 1, et que la densité est normalisée (théorème d e convergence dominée).

On a de la même manière que pour la MGF:

Proposition 1.16

$$\left.\frac{\partial^n \phi_X(t)}{\partial t^n}\right|_{t=0} = i^n \mu_n. \tag{1.25}$$

Définition 1.17 *(Fonction génératrice des cumulants, CGF)*

Soit X une variable aléatoire de densité $p_X(x)$ et de support S. La CGF $K_X(t)$ est définie comme

$$K_X(t) = \ln(M_X(t)) = \ln(\mathbb{E}[e^{tX}]). \tag{1.26}$$

Les cumulants sonts alors définis $\kappa_n = \left.\frac{\partial^n K_X(t)}{\partial t^n}\right|_{t=0}$.

Exemple 5 *Les deux premiers cumulants sont donnés par*

$$\kappa_1 = \left.\frac{\partial \ln(M_X(t))}{\partial t}\right|_0 = \overbrace{\frac{1}{M_X(0)}}^{1}\overbrace{\left.\frac{\partial}{\partial t}M_X(t)\right|_0}^{\mu} = \mu, \tag{1.27}$$

$$\kappa_2 = \left.\frac{\partial^2 \ln(M_X(t))}{\partial t^2}\right|_0 = -\overbrace{\frac{1}{M_X^2(0)}}^{1}\overbrace{\left.\left(\frac{\partial M_X(t)}{\partial t}\right)^2\right|_0}^{\mu_1^2} + \overbrace{\frac{1}{M_X(0)}}^{1}\overbrace{\left.\frac{\partial^2}{\partial t^2}M_X(t)\right|_0}^{\mu_2} = \mu_2 - {\mu_1}^2 = \Delta. \tag{1.28}$$

■

Proposition 1.18 *(Additivité des fonctions génératrices)*

Soient X,Y deux variables aléatoires indépendantes. En général

$$M_{X+Y}(t) = M_X(t)M_Y(t), \tag{1.29}$$

$$K_{X+Y}(t) = K_X(t) + K_Y(t). \tag{1.30}$$

Preuve 3

$$M_{X+Y}(t) = \mathbb{E}[e^{t(X+Y)}] = \mathbb{E}[e^{tX}e^{tY}] = \mathbb{E}[e^{tX}]\mathbb{E}[e^{tY}] = M_X(t)M_Y(t) \tag{1.31}$$

en utilisant l'indépendance pour factoriser les espérances. En prenant le logarithme de 1.31 prouvée ci-dessus, il vient la relation 1.30.

R Cela signifie en particulier que pour $Y = \sum_{i=1}^n X_i$, où les $\{X_i\}_{i=1,\dots,n}$ sont i.i.d, $M_Y(t) = (M_{X_1}(t))^n$ et $K_Y(t) = nK_{X_1}(t)$.

R Il suit en dérivant la relation 1.30 que pour deux variables aléatoires indépendantes, les cumulants des sommes sont les sommes des cumulants.

Notez qu'il est souvent plus commode d'utiliser la fonction caracteristique plutôt que le CGF, qui a le bon goût de toujours exister (ce qui n'est la cas de la CGF):

Définition 1.19 *(Fonction caractéristique) Soit X une variable aléatoire de densité $p_X(x)$ et de support S. La CGF $K_X(t)$ est définie comme*

$$\varphi_X(t) = \ln(M_X(it)) = \ln(\mathbb{E}[e^{itX}]). \tag{1.32}$$

Les cumulants sonts alors définis $i^n\kappa_n = \left.\frac{\partial^n \varphi_X(t)}{\partial t^n}\right|_{t=0}$.

1.4 Petites déviations (ou fluctuations typiques)

Un résultat fondamental, et général, est que les écarts à la moyenne d'une variable aléatoire du type $Y = \sum_{i=1}^n X_i$ avec $\{X_i\}_{i=1,\dots,n}$ i.i.d. sont répartis comme une gaussienne d'écart type $\sqrt{\Delta/n}$, du moins pour les petites déviations: en regle generale, on s'attend a ce que **les petites fluctuations sont presque toujours Gaussiennes**! C'est l'objet de l'un des theoremes les plus fondamentaux des probabilités et statistiques:

Théorème 1.20 *(Théorème Central Limite) - Soit (X_n) une suite de variables aléatoires i.i.d. de moyenne μ et de variance Δ. On considère également la moyenne empirique $Y_n = \frac{1}{n}\sum_{i=1}^n X_i$, alors la variable aléatoire*

$$S_n = \frac{\sqrt{n}(Y_n - \mu)}{\sqrt{\Delta}} \tag{1.33}$$

converge en distribution vers une loi normale de moyenne 0 et de variance 1 : $S_n \xrightarrow{D} X \sim \mathcal{N}(0,1)$.

On peu en proposer une dérivation simple comme suit: Considérons la fonction generatice des cumulants pour S_N (avec pour simplifier $\mu = 0$): On a

$$K_{\sum_i X_i}(t) = NK_X(t) \tag{1.34}$$

mais aussi

$$K_{X/a}(t) = K_X(t/a) \tag{1.35}$$

et donc

$$K_{S_n}(t) = nK_X(t/\sqrt{n}) \tag{1.36}$$

On en deduit donc que les cumulants de S_n sont

$$\kappa_m(S_n) = n\partial_t^m K_X(t/\sqrt{n})|_{t=0} = \frac{n}{n^{m/2}}\kappa_m(X) \tag{1.37}$$

par concequent, dans la limite $n \to \infty$ on trouve que tout les cumulants sont nuls, sauf $\kappa_2 = \Delta$, ce qui est bien le cas pour la Gaussienne!

1.5 Grandes déviations

Si l'on a vu à l'aide du CLT que les écarts à la moyenne d'une variable aléatoire du type $Y = \sum_{i=1}^n X_i$ avec $\{X_i\}_{i=1,\dots,n}$ i.i.d. sont répartis comme une gaussienne d'écart type $\sqrt{\Delta/n}$, il est aussi intéressant de considérer les événements beaucoup plus rares lors desquels une valeur extrême est atteinte. Dans ce cas la, **les grandes fluctutations ne sont pas universalles** et **ne sont pas en general Gaussiennes**. Pour ce faire, les quelques résultats suivants sont utiles.

Proposition 1.21 *(Borne de Chernof)*

Soient X, X_1, ..., X_n i.i.d, $a \in \mathbb{R}$. Alors

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^n X_i \geq a\right] \leq e^{-n(\lambda a - K_X(\lambda))} \qquad \forall \lambda > 0. \tag{1.38}$$

Preuve 4

$$\mathbb{P}[\frac{1}{n}\sum_{i=1}^n X_i \geq a] = \mathbb{P}[\sum_{i=1}^n X_i \geq na] \overset{\lambda \geq 0}{=} \mathbb{P}[\lambda \sum_{i=1}^n X_i \geq \lambda na] = \mathbb{P}[\overbrace{e^{\lambda \sum_{i=1}^n X_i}}^{>0} \geq e^{\lambda na}] \overset{Markov}{\leq} \frac{\mathbb{E}[e^{\lambda \sum_{i=1}^n X_i}]}{e^{\lambda na}} \overset{i.i.d}{=} \frac{\mathbb{E}[e^{\lambda X}]^n}{e^{\lambda na}} = e^{n[\ln(\mathbb{E}[e^{\lambda x}]) - \lambda a]} = e^{-n(\lambda a - K_X(\lambda))} \tag{1.39}$$

Puisque l'inégalité 1.38 est valide pour toute valeur de λ, il est naturel de chercher à trouver une valeur minimisant la borne, i.e. prendre le supremum sur λ de la valeur absolue de l'exponant. Le résultat suivant indique que cette borne optimale est saturée pour $n \to \infty$:

Théorème 1.22 *(Théorème de Cramér)*

$$\frac{1}{n} \ln \mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} X_i \geq a\right] \xrightarrow{n\to\infty} - \sup_{\lambda>0} \left(\lambda a - K_X(\lambda)\right) \tag{1.40}$$

On définit alors la fonction de grande déviation

$$I(a) := \sup_{\lambda>0} \left(\lambda a - K_X(\lambda)\right). \tag{1.41}$$

dont la derivee est donnee par λ.

R Le théorème peut se reformuler $\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} X_i \geq a\right] \asymp e^{-nI(a)}$, où $\asymp$ se lit "se comporte, quand n est grand, comme". C'est la notation standard pour le formalisme des grandes déviations (nous l'avons déjà vue dans la Série 1 pour la méthode de Laplace).

R Il est intéressant de noter que la fonction $I(a)$ est la transformée de Legendre de la fonction génératrice des cumulants. Reportez vous a l'exercice de la Série 1 pour bien comprendre les propriétés de la transformée de Legendre.

Il est facile de voir que le theoreme se generalise facilement pour la probabilite que la somme soit plus petite ou egale a "a"! C'est le meme theoreme, mais cette fois λ est negatif. De fait une fois que nous ecrivons le theoreme des deux cote, le théorème de Cramér nous dit que :

$$\mathbb{P}(Y_N \in [a, a + da]) \asymp e^{-NI(a)}$$

où $I(a)$ est la fonction de grande déviation, donnée par la transformée de Legendre de la fonction génératrice des cumulants. Notez que nous avons plutôt écrit $\mathbb{P}(Y \geq a)$, mais je vous laisse réfléchir au fait que c'est la même chose à l'échelle exponentielle! Pour résumer, Cramer nous enseigne que l'on peut écrire le taux de

grande déviation avec la Tranformée de Legendre de la CGF [2], nous avons donc:

$$I(a) := \sup_{\lambda} (\lambda a - K_X(\lambda)) . \tag{1.45}$$

$$K_X(\lambda) := \sup_{a} (\lambda a - I(a)) . \tag{1.46}$$

$$I'(a) = \lambda \tag{1.47}$$

Preuve 5 *Une preuve complète et rigoureuse prendrait un peu de temps, mais il est simple de vérifier le théorème à posteriori! En effet, si l'on suppose qu'il est existe une fonction de grande déviation pour une variable aléatoire S_n, alors nous pouvons écrire que*

$$\frac{1}{n} \log \mathbb{E}[e^{n\lambda S_n)}] = \frac{1}{n} \log \int dS_n e^{n(-I(S_n)+\lambda S_n)} \tag{1.48}$$

et donc, en utilisant la méthode de Laplace (Serie 1) on trouve

$$K(\lambda) = \lim_{n\to\infty} \frac{1}{n} \log \mathbb{E}[e^{n\lambda S_n)}] = \sup[\lambda s - I(s)] . \tag{1.49}$$

En supposant $I(s)$ convexe, on trouve donc en inversant la transformée de Legendre:

$$I(s) = \sup[\lambda s - K(\lambda)] . \tag{1.50}$$

En appliquant cette propriété a $S_n = \sum_i X_i/n$ on trouve bien le résulat annoncé[a]

[a] Le lecteur astucieux aura par ailleurs remarquer que cette "preuve" fonctionne aussi pour le théorème plus général Thm.1.23.

Exemple 6 *Soit X_i des v.a. iid selon la loi de probabilité Rademacher, i.e. que $X_i = \pm 1$ avec probabilité $\frac{1}{2}$. Définissons alors la v.a. $Y_n := \frac{1}{n}\sum_i^n X_i$. L'esperance des X_i est de 0, leur variance est de 1. On voit donc que $Y_n \approx 0 \pm \frac{1}{\sqrt{n}}$. Intéressons nous maintenant aux grandes déviations de Y_n. Pour cela, calculons la fonction de grande déviation $I(a)$:*

$$M_X(t) = \mathbb{E}[e^{tX}] = \frac{1}{2}\left(e^t + e^{-t}\right) = \cosh(t). \tag{1.51}$$

[2] Même si nous avons écrit cela pour les moyennes empiriques $S_N = 1/N \sum_i X_i$, cela est encore bien plus général! Le théorème de Gärtner–Ellis nous enseigne que le même résultat est vérifié même avec des fonctions bien plus intéressante que la moyenne:

Théorème 1.23 (Gärtner–Ellis, informel) *Si*

$$K(\lambda) = \lim_{n\to\infty} \frac{1}{n} \log \mathbb{E}(\exp(n\lambda A_n))$$

existe et est différentiable pour tout $\lambda \in \mathbb{R}$, alors en définissant

$$I(a) = \sup_{\lambda}(\lambda a - K(\lambda)) , \tag{1.42}$$

on obtient le principe de grandes déviations

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}(A_n = a) \leq -I(a) \tag{1.43}$$

avec égalité pour les points exposés (c.a.d. les points ou la fonction est égale a son enveloppe convexe):

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}(A_n = a) = -I(a) \quad \forall a \in \{\text{points exposes}\} . \tag{1.44}$$

Nous verrons plus tard que c'est à cause de cette généralité que les transformées de Legendre apparaissent en physique statistique et en thermodynamique pour les changements d'ensembles.

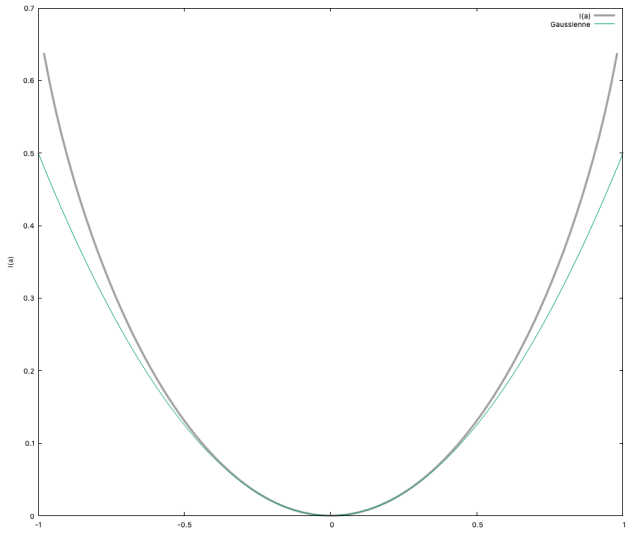

Figure 1.2: Le taux $I(a)$ dans l'exemple (1.53), compare a celui d'une Gaussienne.

Pour obtenir $I(a)$*, il reste à trouver* $\sup_{\lambda\in\mathbb{R}}(\lambda a - \ln(\cosh(\lambda))) = \sup_{\lambda\in\mathbb{R}} \mathrm{g}_a(\lambda)$*:*

$$\frac{\mathrm{dg}_a}{\mathrm{d}\lambda}(\lambda^*) = a - \tanh(\lambda^*) = 0 \Rightarrow \lambda^* = \operatorname{atanh}(a),\ a \in [0\,;1[. \tag{1.52}$$

La seconde dérivée étant strictement négative, λ^* *correspond bien au maximum de* g_a *pour* $a \in [0\,;1[$.
Si $a = 1$, $\sup_{\lambda\in\mathbb{R}}(\lambda - \ln(\cosh(\lambda))) = \ln(2)$, *si* $a > 1$, $\sup_{\lambda\in\mathbb{R}}(\lambda a - \ln(\cosh(\lambda))) = +\infty$, *et si* $a < 0$, *le supremum de* g_a *est 0.*
Ainsi,

$$I(a) = \begin{cases} \left(\ln(\sqrt{1-a^2}) + a\operatorname{atanh}(a)\right), \text{si } a \in]-1\,;1[\\ \ln(2),\ \text{si } a = -1, 1 \\ +\infty,\ \text{si } |a| > 1, \end{cases} \tag{1.53}$$

où on a utilisé $\cosh(\operatorname{atanh}(a)) = \frac{1}{\sqrt{1-a^2}}$*. Enfin, en utilisant le théorème de Cramér sur* X_i *et* $-X_i$*, on obtient que*

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i\right| \geq a\right] \asymp e^{-nI(a)} = \begin{cases} e^{-n\left(\ln(\sqrt{1-a^2})+a\operatorname{atanh}(a)\right)}, \text{si } a \in [0\,;1[\\ \frac{1}{2^n},\ \text{si } a = 1 \\ 0,\ \text{si } a > 1. \end{cases} \tag{1.54}$$

■

Conclusion: Nous avons donc vu pour une variable aléatoire X de moyenne μ et de variance Δ, si l'on observe une moyenne empirique $Y_N = \frac{1}{N}\sum X_i$, on s'attend à ce que $Y_N$soit proche de μ avec des fluctuations typiques d'ordre $\sqrt{\frac{\Delta}{N}}$, et que ces fluctuations soient distribuées comme une gaussienne en vertu du théorème de la limite centrale. Par contre, si l'on s'intéresse aux grandes déviations, qui sont très rares (et même exponentiellement rares), nous avons vu que le théorème de Cramér nous dit que : $\mathbb{P}(Y_N \in [a, a+da]) \asymp e^{-NI(a)}$ où $I(a)$ est la fonction de grande déviation, donnée par la transformée de Legendre de la fonction génératrice des cumulants, c'est-à-dire $I(x) = \sup_{\lambda\in\mathbb{R}}\left(\lambda x - \log\mathbb{E}\left[e^{\lambda X_1}\right]\right)$.

1.6 Grandes déviations et entropie

Ce que nous avons vu pour les calcul de moyenne, c.a.d les variables du type $Y_n = \frac{1}{n}\sum_i X_i$ est en fait une loi très générale, même pour des variables aléatoires bien plus complexes et corellees que les valeurs moyennes (voir le thm 1.23) Les fluctuations typiques sont presque toujours gaussiennes (sauf au point critique en physique statistique), tandis que les grandes déviations ne le sont pas et demandent une analyse plus fine.

Nous allons maintenant présenter un résultat très général, qui permet de mettre en évidence le lien fondamental entre histogramme empirique, entropie, et fonctions de grandes déviations.

Théorème 1.24 *(Théorème de Sanov)*
Soient $X_1, X_2, \ldots, X_k$ des variables aléatoires i.i.d. suivant une distribution de probabilité P, où $P = \{p_i\}_{i=1}^k$ désigne les probabilités théoriques associées à chaque X_i.
La probabilité d'observer une distribution empirique $\hat{P} = \{\hat{p}_i\}_{i=1}^k$ après n observations est donnée par l'expression suivante :

$$\mathbb{P}[\hat{P}] \asymp e^{-n \cdot D_{KL}(\hat{P}||P)}, \tag{1.55}$$

où $D_{KL}(\hat{P}||P)$ désigne la divergence de Kullback-Leibler (aussi appelée entropie relative) définie par :

$$D_{KL}(\hat{P}||P) := \sum_{i=1}^{k} \hat{p}_i \ln\left(\frac{\hat{p}_i}{p_i}\right). \tag{1.56}$$

et où le symbole $\asymp$ signfie que c'est $\frac{1}{n}$ fois le logarithme des deux expressions qui est égal dans la mesure où n tend vers l'infini $\left(ie\ a_n \asymp b_n \leftrightarrow lim_{n\to\infty} \frac{1}{n}\ln(a_n) = lim_{n\to\infty}\frac{1}{n}\ln(b_n)\right)$

R En probabilité et en théorie de l'information, le théorème de Sanov est un peu diffèrent, il traite de la probabilité de trouver la loi empirique *a l'intérieur* d'un ensemble A, plutôt que de trouver exactement une loi empirique $\hat{p}$. Le théorème prend alors un forme un peu différente avec des préfacteurs qui apparaissent devant l'exponentielle. Nous nous contenterons dans ce cours de la forme donnée ci-dessus.

Exemple 7 *Pour illustrer le théorème, on peut s'imaginer la situation suivante : Munis de n boules, on considère un ensemble de k boîtes où lors d'un lancée d'une boule, la probabilité que cette dernière tombe dans la boîte k est de p_k.*
Maintenant, on fixe nos distributions de probabilités P et $\hat{P}$. On va supposer que P correspond à la distribution de probabilité réelle qui dépend de la taille des boîtes les unes par rapport aux autres tandis que $\hat{P}$ correspond à la distribution de probabilité étudiée par le lanceur.

Dans notre exemple, faisons l'hypothèse que le lanceur cherche à déterminer la probabilité d'observer une distribution uniforme au bout de n lancers, ie $\hat{P} \sim U([1,k])$.
En ce qui concerne la distribution de probabilité réelle P, on va considérer 2 cas de figures :

- *Distribution conforme : les boîtes sont de tailles identiques donc la distribution de probabilité réelle P est aussi uniforme ie $P \sim U([1,k])$.*
- *Distribution biasée : les boîtes ne sont pas de la même taille. On suppose que les boîtes impaires sont deux fois plus grandes que les boîtes paires ce qui peut se modéliser par la loi de probabilité suivante :*

$$p_i = P(X_i = m) = \begin{cases} \frac{4}{3k} & si\ m\ est\ impaire \\ \frac{2}{3k} & si\ m\ est\ paire \end{cases} \tag{1.57}$$

Sanity check : $\sum_i p_i = \frac{k}{2}\cdot(p_{paire} + p_{impaire}) = \frac{k}{2}\cdot(\frac{2}{3k} + \frac{4}{3k}) = \frac{k}{2}\cdot\frac{2}{k} = 1$

À présent, on calcule la divergence de Kullbach-Leibler $D_{KL}(\hat{P}||P)$ dans les deux cas de figures :

- *Dans le premier cas où P est uniforme,* $\Rightarrow \quad D_{KL}(\hat{P}||P) = \sum_{i=1}^{k} p_i \ln\left(\frac{p_i}{p_i}\right) = 0$
- *Dans le deuxième cas où P est biaisée,* $\Rightarrow \quad D_{KL}(\hat{P}||P) = \sum_{i=1}^{k} \frac{1}{k} \ln\left(\frac{1/k}{p_i}\right) = \sum_{p_{paire}} \frac{1}{k} \ln\left(\frac{1/k}{p_{paire}}\right) + \sum_{p_{impaire}} \frac{1}{k} \ln\left(\frac{1/k}{p_{impaire}}\right) = \frac{1}{2}\left[\ln\left(\frac{1}{k}\cdot\frac{3k}{2}\right) + \ln\left(\frac{1}{k}\cdot\frac{3k}{4}\right)\right] = \frac{1}{2}\ln\left(\frac{9}{8}\right) = \ln\left(\frac{3}{2\sqrt{2}}\right) \approx 0.0257$

À partir de ces coefficients, on va pouvoir illustrer notre exemple, à savoir la probabilité d'observer une certaine distribution $\hat{P}$, ici uniforme, en fonction d'une distribution donnée P (en supposant n assez grand pour que le théorème s'applique), ici conforme et biasée.

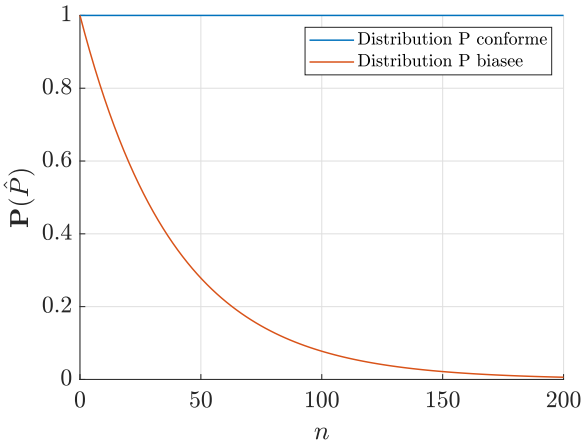

Figure 1.3: Probabilité d'observation d'une distribution $\hat{P}$ uniforme selon de la distribution réelle P en fonction du nombre de lancers n

De cette exemple, on peut conclure que lorsque la distribution réelle P est uniforme, la probabilité d'observer une distribution $\hat{P}$ uniforme vaut logiquement 1 tandis que pour la distribution $\hat{P}$ biasée, la probabilité de l'observer tend vers zéro à mesure que le nombre de lancers n augmente. ∎

Avant de prouver le théorème de Sanov, il est nécessaire de s'intéresser à la notion d'entropie, en présentant l'entropie de Gibbs-Shannon et l'entropie relative.

1.6.1 Entropie de Gibbs-Shannon

Définition 1.25 *(Entropie de Gibbs-Shannon)*

Cas discret *Soit $\{p_1, ..., p_n\}$ avec $0 \leq p_i \leq 1$, $\Sigma_{i=1}^{n} p_i = 1$*

$$H(p) = -\Sigma_{i=1}^{n} p_i \ln p_i. \tag{1.58}$$

Cas continu *Soit* $0 \leq p_X(x)$ *avec* $\int_{-\infty}^{+\infty} p_X(x)dx = 1$

$$H(p) = -\int_{-\infty}^{+\infty} dx p_X(x) \ln p_X(x). \tag{1.59}$$

Théorème 1.26 *Soit le coefficient binomial* $C_n^k = \binom{n}{k} = \frac{n!}{k!(n-k)!}$*, alors ce coefficient est borné inférieurement et supérieurement tel que*

$$\frac{\exp(nH(p=\frac{k}{n}))}{n+1} \leq \binom{n}{k} \leq \exp(nH(p=\frac{k}{n})) \tag{1.60}$$

On établit ainsi que $\binom{n}{k} \asymp \exp(nH(p=\frac{k}{n}))$.

Preuve 6 *On se concentre d'abord sur la borne supérieure.*

On utilise la propriété $\sum_{i=0}^{n} \binom{n}{i} (\frac{k}{n})^i (1-\frac{k}{n})^{n-i} = 1$ *(binôme de Newton), et en particulier*

$$\begin{aligned} &\binom{n}{k} (\frac{k}{n})^k (1-\frac{k}{n})^{n-k} \leq 1 \\ &\binom{n}{k} \exp(n[\frac{k}{n}\ln(\frac{k}{n}) + \frac{n-k}{n}\ln(1-\frac{k}{n})]) \leq 1 \\ &\binom{n}{k} \leq \exp(nH(p=\frac{k}{n})) \end{aligned} \tag{1.61}$$

Pour la borne inférieure, en reprenant le terme du binôme $i = k$*, on a cette fois ci*

$$\begin{aligned} &(n+1)\binom{n}{k} (\frac{k}{n})^k (1-\frac{k}{n})^{n-k} \geq 1 \\ &\binom{n}{k} \geq \frac{\exp(nH(p=\frac{k}{n}))}{n+1} \end{aligned} \tag{1.62}$$

Le comportement du coefficient binomial lorsque n devient grand s'étend aux coefficients multinomiaux:

$$\binom{n}{k_1 k_2 ... k_n} = \frac{n!}{k_1!\, k_2! \, ... k_n!} \asymp \exp(nH(\{p_i = \frac{k_i}{n}\})). \tag{1.63}$$

Même si la preuve si dessus s'adapate bien au cas multinomial, une esquisse de preuve alternative est présentée ci-dessous en utilisant la formule de Stirling pour n et $\{k_i\}_{i=1,...,n} \gg 1$:

$$\frac{n!}{k_1!\, k_2! \, ... k_n!} \asymp \frac{n^n}{\Pi_i^n k_i^{k_i}} \frac{\overbrace{e^{-n}}^{1}}{e^{-\sum_i^n k_i}} = e^{-n\left(\sum_i^n \frac{k_i}{n} \ln(k_i) - \frac{\overbrace{\sum_i^n k_i}^{1}}{n} \ln(n)\right)} = e^{nH(\{p_i = \frac{k_i}{n}\})}, \tag{1.64}$$

en utilisant que $\frac{\sum_i^n k_i}{n} = 1$ et que les termes en $\sqrt{n}$, $\sqrt{k_i}$ de la formule de Stirling sont exponentiellement négligeables asymptotiquement.

1.6.2 Entropie relative

Définition 1.27 *(Entropie relative, ou divergence de Kullback-Leibler)*
L'entropie relative est définie entre deux distributions de probabilités, p et q, discrètes ou continues.
Cas discret *Soient* $p = \{p_i\}$, $q = \{q_i\}$ *avec* $0 \leq p_i, q_i \leq 1$, $\Sigma_{i=1}^n p_i = 1$ *et* $\Sigma_{i=1}^n q_i = 1$,

$$D_{KL}(p||q) = \sum_{i=1}^{n} p_i \ln\left(\frac{p_i}{q_i}\right). \tag{1.65}$$

Cas continu *Soient* $p_X(x)$ *et* $q_X(x)$ *les fonctions de densité de probabilité associées à respectivement p et q,*

$$D_{KL}(p||q) = \int_{\mathbb{R}} dx p_X(x) \ln\left(\frac{p_X(x)}{q_X(x)}\right) = \mathbb{E}\left[\ln\left(\frac{p_X(x)}{q_X(x)}\right)\right]. \tag{1.66}$$

R Noter que pour $q = \{\frac{1}{n}\}$ une distribution uniforme sur n valeurs discrètes, $D_{KL}(p||q) = -H(p) - \ln(\frac{1}{n})$, i.e. la divergence de Kullback-Leibler se comporte alors à une constante près comme l'entropie de Gibbs-Shannon.

Proposition 1.28 *(Propriété de Gibbs)*
Soient deux distributions de probabilités, p et q. Alors la divergence de Kullback-Leibler respecte

$$D_{KL}(p||q) \geq 0, \ \textit{avec} \ \ D_{KL}(p||q) = 0 \text{ ssi } p = q. \tag{1.67}$$

Preuve 7 *Puisque* $\ln(x) \leq x - 1 \ \forall x > 0$ *, en particulier*

$$\ln\left(\frac{q_i}{p_i}\right) \leq \frac{q_i}{p_i} - 1 \Rightarrow \sum_i p_i \ln\left(\frac{q_i}{p_i}\right) \leq \sum_i p_i \left(\frac{q_i}{p_i} - 1\right) = \sum_i p_i - \sum_i q_i = 1 - 1 = 0 \tag{1.68}$$

en utilisation la normalisation des deux lois. Il apparaît ensuite que le cas d'égalité est bien donné par $q_i = p_i \ \forall i$ *puisque l'inégalité sur le logarithme est saturée seulement en 1. Le cas continu se traite de manière similaire, à la différence qu'il est possible d'exiger uniquement* $p = q$ *presque partout pour le cas d'égalité.*

R La divergence de Kullback-Leibler exprime en quelque sorte à quel point deux lois de probabilité sont différentes l'une de l'autre puisqu'elle vaut 0 seulement quand les deux lois sont identiques. Cependant puisque elle n'est pas symétrique (et ne respecte pas l'inégalité triangulaire) on ne peut pas parler de distance au sens mathématique du terme, mais plutôt de divergence.

À partir de cela, montrons notre version du théorème de Sanov.

Preuve 8 *(Théorème de Sanov) La probabilité d'observer la distribution empirique* $\hat{p} = \{p_i\}_{i=1}^k$ *à partir de N observations suivant la distribution "réelle" p est:*

$$\begin{aligned} \mathbb{P}\left[\{\hat{p}_1, ..., \hat{p}_k\}\right] &= p_1^{N\hat{p_1}} ... p_k^{N\hat{p_k}} \binom{N}{N\hat{p_1}, ..., N\hat{p_k}} \\ &= \binom{N}{N\hat{p_1}, ..., N\hat{p_k}} e^{N\sum_i \hat{p}_i \ln(p_i)}. \end{aligned} \tag{1.69}$$

Et en utilisant l'approximation des coefficients multinomiaux :

$$\mathbb{P}\left[\{\hat{p}_1, ..., \hat{p}_k\}\right] \asymp e^{N\sum_i \hat{p}_i \ln(p_i) - N\sum_i \hat{p}_i \ln(\hat{p}_i)} = e^{-ND_{KL}(\hat{p}||p)}. \tag{1.70}$$

Un exerice utile est d'utiliser le Theoreme de Sanov pour retrouver le resultat de l'example 6. En effet si l'on tire n fois une boule qui peut etre a gauche ou a droite avec probabilite 1/2, et que l'on se demande quelle est la probailite d'obervser une fraction ρ_+ et ρ_- on trouve que le taux est donnée par

$$D_{\mathrm{KL}}(\rho|\mathrm{uniform}) = \sum_i \rho_i \log rho_i/(1/2) = \log(2) - H(\rho) \tag{1.71}$$

L'entropie binaire est triviale a calculer! On trouve donc que le taux de grande deviation est

$$D_{\mathrm{KL}}(\rho|\mathrm{uniform}) = \log(2) + (\rho_+)\log(\rho_+) + (\rho_-)\log(\rho_-) \tag{1.72}$$

Si l'on choisi d'ecrire ce resultat avec la valeur moyenne $a = \rho_+ + \rho_-$ on trouve

$$D_{\mathrm{KL}}(\rho|\mathrm{uniform}) = \log(2) + \frac{1+a}{2}\log(\frac{1+a}{2}) + \frac{1-a}{2}\log(\frac{1-a}{2}) \tag{1.73}$$

Neme si cela n'est pas evident, c'est exactmenet la meme fonction que la fonction eq.(1.53).

1.7 Maximisation de l'entropie sous contraintes

La morale de cette histoire est que l'entropie mesure la probabilité d'observer une distribution $\{p\}$ lorsque la distribution sous-jacente est elle-même uniforme. Plus précisément, l'entropie est le logarithme (rapporté au nombre d'observations) de cette probabilité. Ainsi, plus l'entropie est grande, plus il est probable d'observer la distribution $\{p\}$. Il est donc évident que la distribution la plus " probable " est celle qui maximise l'entropie. C'est exactement ce qui justifie le principe de **MAXENT** (Maximum Entropy), selon lequel, en l'absence d'information supplémentaire, la meilleure estimation de la distribution est celle d'entropie maximale. Toutes les autres distributions sont (exponentiellement) plus rares !

Ce principe est fondamental. Il permet de répondre à des questions importantes, et justifie notamment l'utilisation de la famille des distributions exponentielles. En effet, considérons la question suivante : j'ai Q états possibles, et ma distribution originale sous-jacente est uniforme. Je réalise un grand nombre N de tirages, mais je ne m'intéresse qu'aux configurations telles que

$$\sum_{i=1}^{Q} \rho_i E_i = E\,.$$

Dans ce cas, il nous faut considérer la distribution $\{\rho_i\}$ qui maximise l'entropie sous la contrainte

$$\sum_{i=1}^{Q} \rho_i E_i - E = 0\,.$$

Cela se fait aisément avec la technique des multiplicateurs de Lagrange :

$$\max_{\{\rho\}} H(\{\rho\}), \quad \text{avec} \quad \sum_{i=1}^{Q} \rho_i E_i - E = 0 \tag{1.74}$$

est équivalent à

$$\text{extr}_{\{\rho\},\beta}\, H(\{\rho\}) + \beta(\sum_{i=1}^{Q} \rho_i E_i - E)\,, \tag{1.75}$$

ce qui nous donne

$$-\log \rho_i - 1 + \beta E_i = 0 \quad \Longrightarrow \quad \rho_i \propto e^{-\beta E_i}\,. \tag{1.76}$$

Ainsi,

$$\rho_i = \frac{1}{Z} e^{-\beta E_i}\,, \tag{1.77}$$

où Z est la constante de normalisation (fonction de partition).

Si l'on veut fixer une valeur moyenne, on trouve donc que la distribution d'entropie maximale est la loi exponentielle. On peut vérifier facilement que si l'on fixe à la fois une moyenne et une variance, la distribution d'entropie maximale est la loi gaussienne. Nous reviendrons sur ce point plus tard dans le cours.

1.8 L'Entropie selon Shannon

Pour finir, discutons brièvement d'une vision équivalente, mais alternative, de l'entropie, due au grand ingénieur Claude Shannon. Dans cette perspective, l'entropie peut être comprise comme une mesure de l'*information manquante* nécessaire pour décrire un système aléatoire, ou encore comme la *surprise moyenne* que l'on éprouve en observant la réalisation d'une variable aléatoire.

Mathématiquement, si X est une variable aléatoire distribuée selon p, on peut définir la *surprise* associée à un événement comme $-\log p(X)$. L'entropie s'interprète alors comme la valeur moyenne de cette surprise :

$$H(p) = \mathbb{E}_{X\sim p}[-\log p(X)] = -\sum_x p(x) \log p(x)\,. \tag{1.78}$$

Ainsi, une distribution très concentrée correspond à une faible entropie (peu de surprise), tandis qu'une distribution uniforme correspond à une entropie maximale (incertitude maximale).

On voit donc que les deux visions — la justification fréquentiste par la probabilité d'observer une distribution, et l'interprétation informationnelle comme surprise moyenne — conduisent toutes deux naturellement au principe de **MAXENT**. Ce principe apparaît ainsi comme un pont entre la théorie de l'information de Shannon et la vision de statistique de grandes deviations: dans les deux cas, l'entropie sélectionne la distribution la plus naturelle ou la plus probable en l'absence d'information supplémentaire.

L'entropie de Shannon reprend formellement l'expression introduite bien plus tôt par Boltzmann et Gibbs en mécanique statistique, $S = -k \sum p \log p$, où k est la constante de Boltzmann. Lorsque Shannon introduisit sa mesure de l'information, il demanda conseil à John von Neumann sur le choix du nom à lui donner. Von Neumann lui répondit :

> " Vous devriez l'appeler entropie, pour deux raisons : (i) la même formule $-\sum p \log p$ existe déjà en mécanique statistique sous le nom d'entropie, et (ii) personne ne sait vraiment ce qu'est l'entropie, ainsi, dans une discussion, vous aurez toujours l'avantage. "

C'est ainsi que la formule de Boltzmann et Gibbs, $S = -k \sum p \log p$, est entrée dans la théorie de l'information. La vision de grande deviation est arrivee un peu plus tards, notement avec les travaux du Mathematicien Srinivasa Varadhan. Cette convergence entre mathematiques, physique et théorie de l'information illustre la profondeur du concept : que l'on parle d'atomes ou de messages, c'est toujours la même mesure d'incertitude fondamentale qui intervient.