Excursions le long de la Gaussienne

Yvik Swan*

Université de Liège

Abstract

Nous introduisons la définition d'entropie différentielle $H(X)$ d'un signal aléatoire X et démontrons qu'un signal Gaussien est à entropie maximale. Nous étudions l'entropie d'un signal $X_\tau = X + \sqrt{\tau}Z$ composé d'une source X brouillée par un bruit Gaussien Z. Nous utilisons le fait que la loi Gaussienne satisfait une équation de la chaleur pour donner une formule exprimant l'accroissement instantané $\partial_\tau H(X_\tau)$ en terme d'une quantité $I(X_\tau)$, appelée information de Fisher. Cette formule permet, après quelques détours, de résoudre une conjecture datant de 1948 selon laquelle l'entropie d'une somme normalisée de variables aléatoires $X_1, X_2, \ldots, X_n$ indépendantes et de même loi croît de façon monotone (en n) vers son maximum. Ce résultat donne une interprétation naturelle du célèbre *théorème central limite*.

Mots clés: Variables et vecteurs aléatoires, entropie de Shannon, entropie différentielle, information de Fisher, semi-groupe d'Ornstein-Uhlenbeck, monotonicité de l'entropie et conjecture de Shannon, théorème central limite entropique

Contents

*Département de Mathématique, Grande Traverse 12, Sart Tilman, B-4000 Liège

1 Préambule

L'article qui suit est issu d'un cours donné lors de la 7ème édition de la BSSM (Août 2014). La lecture de cet article ne nécessite aucune connaissance particulière, hormis des notions élémentaires de calcul différentiel et intégral. Les résultats présentés sont classiques; d'excellentes références sur ces sujets sont les livres [9, 12], dans lesquels le lecteur intéressé pourra trouver d'autres références. Les résultats plus avancés présentés dans les dernières sections proviennent, pour la plupart, des articles [1, 2, 4]. D'autres références seront données dans le texte.

2 Introduction : notions de probabilité

Considérons une expérience aléatoire dont toutes les issues possibles sont regroupées en un ensemble Ω. Sur Ω nous collons une notion de probabilité en définissant un couple $(\mathcal{A}, P)$ avec

- $\mathcal{A}$ une collection d'événements (appelée σ-algèbre)
- P une fonction qui assigne à chaque événement A de $\mathcal{A}$ un nombre $P(A) \in [0, 1]$ (une probabilité).

On appelle la quantité $P(A)$ la probabilité de l'événement A. Intuitivement, $P(A)$ représente notre confiance en l'occurrence de l'événement A avec $P(A) = 1$ indiquant que nous sommes certains que l'événement va se produire (exemple : le soleil va se lever) et $P(A) = 0$ indiquant que nous sommes certains que l'événement ne va pas se produire (exemple : la terre va s'arrêter de tourner).

Remarque 2.1. *Tant la collection $\mathcal{A}$ que la fonction P doivent satisfaire certaines règles dont nous n'aurons pas besoin dans la suite; nous renvoyons le lecteur intéressé à un cours d'introduction à la probabilité (par exemple [14]) pour plus de détails.*

Soit maintenant $X : \Omega \to \mathbb{R}^n$ une fonction "régulière"[1] qui permet de résumer une expérience (d'issues Ω) en termes d'un certain nombre de quantités ($X = (X_1, \ldots, X_n) \in \mathbb{R}^n$). On appelle X un n-vecteur aléatoire (réel); quand $n = 1$ on parle généralement d'une variable aléatoire.

Exemple 2.1. *L'expérience consiste à regarder une personne; Ω contient toute les informations la concernant. La fonction X nous renvoie divers indicateurs comme par exemple sa taille, son poids, sa tension artérielle, le temps avant le prochain appel téléphonique que la personne recevra, ...*

Etant donnés le triplet $(\Omega, \mathcal{A}, P)$ (parfois appelé trinité probabiliste) et la fonction $X : \Omega \to \mathbb{R}^n$ nous introduisons la fonction

$$B \mapsto P_X(B) \overset{\text{not}}{=} P(X \in B) = P\left(\{\omega \in \Omega : X(\omega) \in B\}\right) = P\left(X^{-1}(B)\right)$$

qui agit sur les ensembles $B \subset \mathbb{R}^n$ "réguliers"[2]. Intuitivement la fonction P_X calcule la probabilité d'un sous-ensemble de $\mathbb{R}^n$ en fonction du comportement de X; on l'appelle la mesure de probabilité induite par X sur $\mathbb{R}^n$. Nous allons dans la suite nous intéresser principalement aux X tels que

$$P_X(B) = \int_B f^X(x)dx \tag{2.1}$$

[1] Borel-mesurable, pour les intimes
[2] Boréliens, pour les intimes

avec $f^X : \mathbb{R}^n \to \mathbb{R}$ une fonction positive intégrable telle que $\int_{\mathbb{R}^n} f^X(x)dx = 1$. Cette fonction est appelée la densité de X (ou encore sa loi) et on note

$$X \sim f^X \tag{2.2}$$

pour indiquer que (2.1) a lieu.

Exemple 2.2. *L'expérience consiste à choisir un point au hasard dans un rectangle $R \subset \mathbb{R}^2$. Prenons $X = (X_1, X_2)$ les coordonnées de ce point. Supposons que le rectangle a longueur L et largeur l. Alors*

$$P_X(B) = \frac{\text{Aire}(R \cap B)}{\text{Aire}(R)} = \int_B \frac{\mathbb{I}((x,y) \in R)}{L \times l} dxdy \tag{2.3}$$

pour $\mathbb{I}(\cdot \in R)$ la fonction indicatrice du rectangle R, qui vaut 1 si $\cdot \in R$ et 0 sinon. En définissant $f^X(x,y) = \mathbb{I}((x,y) \in R)/(L \times l)$ on a bien $X \sim f^X$.

Définition 2.1. *Soit $e \in S^{n-1}$ un vecteur unité de $\mathbb{R}^n$. On appelle* marginale de X *dans la direction e la fonction*

$$t \mapsto h(t) = \int_{te+e^\perp} f^X(x)dx \tag{2.4}$$

pour $t \in \mathbb{R}$, avec $te + e^\perp$ l'hyperplan de $\mathbb{R}^n$ de vecteur directeur e passant par le point te.

Etant donné un système de coordonnées de base canonique $e_1 = (1, 0, \ldots, 0), \cdots, e_n = (0, 0, \ldots, 1)$ on définit les *marginales canoniques*

$$\begin{aligned} f^{X_j}(t) &= \int_{te_j + e_j^\perp} f^X(x)dx \\ &= \int_{\mathbb{R}^{n-1}} f^X(x_1, \ldots, x_{j-1}, t, x_{j+1}, \ldots, x_n)dx_1 \cdots dx_{j-1}dx_{j+1} \cdots dx_n \end{aligned} \tag{2.5}$$

pour tout $1 \leq j \leq n$.

Exemple 2.3. *Reprenons la densité de l'exemple 2.2. Choisissons le système de coordonnées de façon à ce que les côtés du rectangle soient parallèles aux axes, et donc que*

$$R = \left\{(x,y) \in \mathbb{R}^2 \text{ tels que } a \leq x \leq a + L \text{ et } b \leq y \leq b + l\right\}.$$

On calcule alors aisément

$$f^{X_1}(x) = \int_{\mathbb{R}} f^X(x,y)dy = \frac{1}{L}\mathbb{I}(x \in [a, a+L])$$

et

$$f^{X_2}(y) = \int_{\mathbb{R}} f^X(x,y)dx = \frac{1}{l}\mathbb{I}(y \in [b, b+l]).$$

Remarquez que

$$f^X(x,y) = f^{X_1}(x)f^{X_2}(y). \tag{2.6}$$

Evidemment les choses sont plus compliquées si le rectangle n'est pas placé parallèlement aux axes; en particulier on perdrait la propriété (2.6).

Définition 2.2. *Les variables aléatoires* $X_1, X_2, \ldots, X_n$ *sont* indépendantes *si*

$$f^X(x_1, \ldots, x_n) = \prod_{j=1}^{n} f^{X_j}(x_j) \tag{2.7}$$

pour $X = (X_1, \ldots, X_n)$*, i.e. si leur loi jointe se factorise en le produit des marginales.*

3 La Gaussienne et la loi des séries

Soit $X = (X_1, X_2, \ldots, X_n)$. Plutôt qu'au comportement du vecteur complet on s'intéresse souvent à une statistique issue de X; parmi les statistiques les plus naturelles on trouve la somme partielle

$$Y_n = \sum_{j=1}^{n} X_j.$$

Partant des lois individuelles des X_j il est aisé d'obtenir une expression pour la loi de Y_n, à tout le moins sous l'hypothèse d'indépendance (2.7). En effet, la loi de la variable aléatoire $X_1 + X_2$ n'est rien d'autre que la marginale du couple (X_1, X_2) dans la direction $(1, 1)$; par conséquent

$$f^{X_1+X_2}(t) = \int_{\{(x_1,x_2)\,|\,x_1+x_2=t\}} f^{(X_1,X_2)}(x_1, x_2)dx_1dx_2 \tag{3.1}$$

$$= \int_{\mathbb{R}} f^{X_1}(t-v)f^{X_2}(v)dv \tag{3.2}$$

où, pour passer de (3.1) à (3.2), nous avons utilisé le fait que $f^{(X_1,X_2)}(x_1, x_2) = f^{X_1}(x_1)f^{X_2}(x_2)$ et appliqué le changement de variables $u = x_1 + x_2$ et $v = x_2$. En itérant le raisonnement, il est aisé de déduire le résultat suivant.

Proposition 3.1. *Soient* $X_1, \ldots, X_n$ *des variables aléatoires indépendantes. Alors*

$$f^{X_1+\ldots+X_n} = f^{X_1} * \cdots * f^{X_n} \tag{3.3}$$

pour $f * g(x) = \int_{\mathbb{R}} f(x-v)g(v)dv$ *le* produit de convolution *des fonctions* f *et* g*.*

Nous disposons donc d'une formule qui permet de calculer, du moins en théorie, la loi d'une somme partielle quelle que soit la loi des incréments. Malheureusement on se rend compte avec un peu de pratique que, bien qu'élégante, l'expression (3.3) est très peu amène aux calculs dès que n devient un peu grand.

Il se trouve néanmoins que, pour n grand, les variables aléatoires de la forme Y_n ont à peu près toute un comportement similaire : leur distribution de probabilité finit toujours par ressembler à une "courbe en cloche" (chapeau de Napoléon). Ce résultat est intuitivement facile à expliquer : la plupart des variables auront une valeur proche d'une certaine valeur moyenne, il y aura donc une grande concentration autour d'un centre (ce qui donne la bosse de la cloche) et peu de variables auront une valeur excessivement grande ou petite (donc la courbe ne se poursuit pas à l'infini). Il existe bien entendu un grand nombre de courbes en cloche; toutefois nous savons depuis longtemps que, parmi toutes, une seule est élue. En effet, selon le célèbre *théorème central limite*, on a toujours

$$f^{Y_n} \approx e^{-x^2/2} \tag{3.4}$$

pour n grand et ce *quelle que soit la loi des* X_j! La convolution (3.3) a un effet lissant sur ses entrées et mène *toujours* à la même courbe en cloche.

Remarque 3.1. *Nous renvoyons le lecteur intéressé d'avoir un énoncé complet ou plus de détails sur cette équation et sa signification à la page wikipedia appropriée, de même qu'au livre [10].*

On déduit de (3.4) que la courbe $e^{-x^2/2}$ joue un rôle remarquable et unique parmi toutes les fonctions ayant un graphe "en cloche". La loi de probabilité issue de cette courbe est appelée la *loi normale* ou encore la *loi Gaussienne* en l'honneur du mathématicien allemand Carl-Friedrich Gauss (1777-1855).

Définition 3.1. *Une variable aléatoire Z est de loi Gaussienne standard si*

$$f^Z(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} =: \phi(x); \tag{3.5}$$

on écrit $Z \sim \mathcal{N}(0,1)$.

La loi Gaussienne est dite centrée et réduite; en vocabulaire statistique elle a moyenne nulle et variance 1. Dans la suite nous nous contentons de faire remarquer (exercice à faire) que

$$\int_{\mathbb{R}} \phi(x)dx = 1, \quad \int_{\mathbb{R}} x\phi(x)dx = 0, \text{ et } \int_{\mathbb{R}} x^2\phi(x)dx = 1. \tag{3.6}$$

Le TCL nous enseigne donc que, pour n grand, on a toujours

$$\sum_{j=1}^{n} X_j \overset{\mathcal{L}}{\approx} \alpha_n Z + \beta_n \tag{3.7}$$

pour des constantes α_n et β_n à déterminer et $\overset{\mathcal{L}}{\approx}$ désignant l'approximation en loi, i.e. le membre de gauche a presque la même loi que le membre de droite.

Une question naturelle à ce stade est de déterminer *pourquoi* la fonction $e^{-x^2/2}$ et donc les variables $Z \sim \mathcal{N}(0,1)$ jouent ce rôle particulier. Nous allons maintenant donner une explication élégante de ce phénomène (appelé universalité de la loi Gaussienne) en étudiant la notion d'entropie.

4 Entropie discrète

Soit $X : \Omega \to \mathbb{R}$ une variable aléatoire; par définition cette quantité sert à décrire un phénomène fortuit. Claude Shannon (1916 - 2001, MIT + Bell + services secrets américains) fut le premier à étudier la notion de "surprise" associée à un phénomène fortuit X.

Définition 4.1. *Si un événement A a probabilité $P(A)$ alors*

$$I(A) = -\log_2 P(A) \tag{4.1}$$

est l'information de A.

Intuitivement, (4.1) reflète le fait que plus A est rare, plus son information $I(A)$ est grande : on apprend beaucoup plus en observant un événement rare qu'en observant un événement commun.

Exemple 4.1. *Clairement, I(ne pas gagner le loto) ≈ 0; notre vie n'est pas bouleversée quand on ne gagne pas. En revanche I(gagner le loto) $\approx \infty$; notre vie est tourneboulée si cet événement se réalise.*

Remarque 4.1. *Bien sûr on est en droit de questionner la légitimité du choix de la fonction* $-\log_2$ *dans* (4.1)*. En première approximation, on pourrait penser que n'importe quelle fonction décroissante ferait l'affaire. Il y a toutefois des raisons pratiques qui justifient la définition* (4.1)*. Premièrement on souhaite naturellement que* $I(A+B) = I(A) + I(B)$ *lorsque* A *et* B *sont indépendants. Or dans ce cas* $P(A \cap B) = P(A)P(B)$ *et donc il est préférable de choisir une fonction logarithmique. Reste encore à expliquer le choix de la base 2. Pour cela considérons l'expérience la plus simple qui consiste à jeter une pièce équilibrée. Il est alors naturel d'associer à 'événement "la pièce rend pile" (dont la probabilité vaut 1/2) l'information 1. On dira* $I(\text{la pièce rend pile}) = 1$ *bit, le mot* bit *ayant été popularisé comme mesure d'information par Shannon dans son article fondateur de 1948, cf [15]. D'autres définitions sont toutefois possibles, et nous renvoyons le lecteur intéressé aux références [9, 12] pour plus de détails.*

Définition 4.2. *L'*entropie de Shannon *d'une variable aléatoire qui prend les valeurs* $x_1, \dots, x_n$ *avec probabilités* $p_1, \dots, p_n$ *est*

$$H(X) = -\sum_{j=1}^{n} p_j \log_2 p_j = \sum_{j=1}^{n} p_j I(\{X = x_j\}) \tag{4.2}$$

L'entropie mesure donc l'information moyenne apportée par la variable aléatoire X.

Remarque 4.2. *On lance une pièce équilibrée; soit* X *une variable aléatoire qui vaut* x_1 *si la pièce rend pile et* x_2 *sinon. On a* $H(X) = 1$ *si la pièce est équilibrée. En revanche si la pièce montre Pile avec probabilité* p *au lieu de 1/2 on obtient directement*

$$H(X) = p \log_2(1/p) + (1-p) \log_2(1/(1-p)) < 1. \tag{4.3}$$

Remarquez que les résultats ci-dessous ne dépendent pas des valeurs prises par X*.*

5 Entropie continue et le saut d'entropie

Soit X un vecteur aléatoire de loi f sur $\mathbb{R}^n$. Son entropie (différentielle) est la quantité

$$H(X) = -\int_{\mathbb{R}^n} f \log f$$

avec $\log = \log_e$ et la convention que $0 \times \log 0 = 0$. En particulier si $n = 1$ et $Z \sim \phi$ la loi Gaussienne standard on calcule

$$H(Z) = -\int_{-\infty}^{\infty} \left(-\frac{x^2}{2} - \log\sqrt{2\pi} \right) \phi(x) dx = \log\left(\sqrt{2\pi e}\right).$$

Supposons maintenant que $\int x f(x) dx = 0$ et $\int x^2 f(x) dx = 1$. Considérons la quantité

$$H(Z) - H(X)$$

qu'on appelle l'*entropy jump* entre la loi de X et la Gaussienne.

Proposition 5.1.

$$H(Z) - H(X) = \int f \log \frac{f}{\phi} \tag{5.1}$$

Proof. Bien entendu on a $H(Z) - H(X) = \log\left(\sqrt{2\pi e}\right) - H(X)$. On peut également écrire

$$\begin{aligned}\int f \log \frac{f}{\phi} &= \int f \log f - \int f \log \phi \\ &= -H(X) + \int \frac{x^2}{2} f + \log\sqrt{2\pi} \int f \\ &= -H(X) + 1/2 + \log\sqrt{2\pi},\end{aligned}$$

dont le résultat découle. □

Ce résultat d'apparence anodine a de grandes conséquences. Tout d'abord il permet d'exprimer l'entropy jump entre X et Z en termes de la quantité

$$D(X|Z) = \int f \log \frac{f}{\phi} \tag{5.2}$$

connue, depuis les années 50, sous le nom de divergence de Kullback-Leibler. Cette quantité a été introduite indépendamment de Shannon afin de mesurer la dissimilarité entre f et ϕ; elle apparaît de manière naturelle lors de la comparaison de tests statistiques. Deuxièmement, l'observation (5.1) nous enseigne que

$$\begin{aligned}H(Z) - H(X) &= \int f \log \frac{f}{\phi} \\ &\geq \int f \left(1 - \frac{\phi}{f}\right) \qquad \left(\text{car } \log u \geq 1 - \frac{1}{u}\right) \\ &= \int f - \int \phi = 0.\end{aligned}$$

En d'autres termes on a le résultat suivant.

Théorème 5.1. *La loi normale ϕ est à entropie maximale parmi toutes les lois de moyenne nulle et de variance 1.*

On peut également aller plus loin. Souvenons-nous que

$$f^{X+Y}(t) = \int f^X(t-u) f^Y(u) du;$$

or la fonction $u \mapsto u \log u =: \Psi(u)$ est une fonction convexe donc, par l'inégalité de Jensen[3], on a

$$\begin{aligned}\Psi(f^{X+Y}(t)) &= \Psi\left(\int \left(f^X(t-u)\right) f^Y(u) du\right) \\ &\leq \int \Psi\left(f^X(t-u)\right) f^Y(u) du.\end{aligned}$$

[3]sorte d'inégalité triangulaire généralisée d'après laquelle $\Psi(\int h(x)g(x)dx) \leq \int \Psi(h(x))g(x)dx$ pour toute fonction de densité g et toute fonction convexe Ψ; cf par exemple Wikipedia pour plus de détails

En intégrant sur toutes les valeurs de t et en échangeant l'ordre d'intégration on obtient

$$\begin{aligned}\int \Psi(f^{X+Y}(t))dt &\leq \int \left\{ \int \Psi\left(f^X(t-u)\right) f^Y(u)du \right\} dt \\ &\leq \int \left\{ \int \Psi\left(f^X(t-u)\right) dt \right\} f^Y(u)du.\end{aligned}$$

Or, pour tout $u \in \mathbb{R}$, on a

$$\int \Psi\left(f^X(t-u)\right) dt = -H(X)$$

de même que

$$\int \Psi(f^{X+Y}(t)) = -H(X+Y).$$

En utilisant le fait que f^Y intègre à 1 (c'est une densité), on peut enfin conclure.

Proposition 5.2. *Si X et Y sont indépendants alors*

$$H(X+Y) \geq H(X). \tag{5.3}$$

L'inégalité (5.3) est plutôt grossière. Shannon [15] et Stam [16] ont, les premiers, prouvé un résultat un peu plus précis, à savoir

$$H\left(\frac{X+Y}{\sqrt{2}}\right) \geq H(X). \tag{5.4}$$

En itérant (5.4) on déduit que, pour $Y_n = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} X_j$ avec X_j indépendants de même loi on a toujours

$$H(Y_{2^{j+1}}) \geq H(Y_{2^j}) \tag{5.5}$$

et donc l'entropie d'une somme croît le long des puissances de 2. Ceci nous mène sur la voie d'un théorème central limite entropique dans lequel on verrait la convergence en loi de Y_n vers la Gaussienne comme la résultante d'une attraction inextricable en termes de l'entropie.

Conjecture (Shannon, 1948)

$$H(Y_{n+1}) \geq H(Y_n). \tag{5.6}$$

La preuve de (5.6) est restée cachée jusqu'en 2004, année où deux groupes de chercheurs ont, indépendamment et de façon différente, construit des outils suffisamment fins pour fournir une preuve de (5.6). Les deux approches (présentées pour la première fois dans [4] et [13], respectivement) empruntent le même chemin le long d'une courbe menant de la loi de X à celle de Z.

6 La Gaussienne sous la chaleur

Rappelons que la densité d'une variable Gaussienne Z est $\phi(x) = (2\pi)^{-1}e^{-x^2/2}$. Pour $\tau > 0$ on a que la densité de la variable $\sqrt{\tau}Z$ est

$$\phi_\tau(x) = \frac{1}{\sqrt{2\pi\tau}} e^{-\frac{x^2}{2\tau}}. \tag{6.1}$$

Il est raisonnablement aisé (mais un peu fastidieux) de montrer que la fonction définie en (6.1) satisfait l'équation aux dérivées partielles

$$\partial_\tau \phi_\tau(x) = \frac{1}{2}\partial_x^2 \phi_\tau(x), \tag{6.2}$$

que d'aucuns reconnaîtront comme la célèbre équation de la chaleur. Prenons maintenant X une variable aléatoire de loi f, indépendante de Z et considérons la quantité aléatoire

$$X_\tau = X + \sqrt{\tau}Z.$$

La variable X_τ s'interprète comme un signal X brouillé par un bruit aléatoire $\sqrt{\tau}Z$. Sa loi, notée q_τ, est donnée (une fois de plus il faut se rappeler de (3.2)) par

$$q_\tau(u) = \int \phi_\tau(u-x) f(x) dx.$$

Alors, moyennant des conditions de régularité (qui se trouvent être *toujours* satisfaites) on déduit

$$\begin{aligned}\partial_\tau q_\tau(u) &= \int \left(\partial_\tau \phi_\tau(u-x)\right) f(x) dx = \int \frac{1}{2}\partial_u^2 \phi_\tau(u-x) f(x) dx \\ &= \frac{1}{2}\partial_u^2 \left(\int \phi_\tau(u-x) f(x) dx \right)\end{aligned}$$

et donc

$$\partial_\tau q_\tau(u) = \frac{1}{2}\partial_u^2 q_\tau(u) \tag{6.3}$$

c'est-à-dire que la loi de la variable $X + \sqrt{\tau}Z$ satisfait également une équation de la chaleur. Cette découverte aura dans la suite une importance capitale.

7 Dérivée de l'entropie

Reprenons X_τ de la section précédente et considérons la fonction $\tau \mapsto H(X_\tau)$ qui décrit l'évolution de l'entropie le long d'un chemin partant de X et évoluant le long d'une déformation Gaussienne de plus en plus importante. En utilisant (6.3) on calcule (toujours moyennant des hypothèses de régularité qui sont toujours satisfaites)

$$\begin{aligned}\partial_\tau H(X_\tau) &= -\partial_\tau \int q_\tau \log q_\tau \\ &= -\int (\partial_\tau q_\tau) \log q_\tau - \int q_\tau \left(\partial_\tau \log q_\tau\right) \\ &= -\frac{1}{2}\int \left(\partial_x^2 q_\tau\right) \log q_\tau - \int \partial_\tau q_\tau.\end{aligned}$$

On a, d'une part,

$$\int \partial_\tau q_\tau = \partial_\tau \int q_\tau = \partial_\tau 1 = 0$$

et, d'autre part,

$$\begin{aligned}\int \left(\partial_x^2 q_\tau\right) \log q_\tau &= \left[\partial_x q_\tau \log q_\tau\right]_{-\infty}^{\infty} - \int \left(\partial_x q_\tau\right) \partial_x \log q_\tau \\ &= \left[\partial_x q_\tau \log q_\tau\right]_{-\infty}^{\infty} - \int \frac{\left(\partial_x q_\tau\right)^2}{q_\tau}.\end{aligned}$$

En remarquant que $\left[\partial_x q_\tau \log q_\tau\right]_{-\infty}^{\infty} = 0$ on conclut

$$\partial_\tau H(X_\tau) = \frac{1}{2} \int \frac{\left(\partial_x q_\tau\right)^2}{q_\tau}. \tag{7.1}$$

Loin d'être anodine, l'identité (7.1) constitue une sorte de révolution copernicienne dans l'étude de l'entropie. En effet la quantité du membre de droite de cette équation est ce qu'on appelle l'*information de Fisher de* X_τ, une quantité étudiée par les statisticiens depuis le début du 20ème siècle et bien connue également des analystes.

Définition 7.1. *Soit X de loi f. L'information de Fisher de X est la quantité*

$$I(f) = I(X) = \int \left((\log f)'\right)^2 f = \int \frac{(f')^2}{f}$$

Remarque 7.1. *La quantité $I(f)$ apparaît naturellement dans l'étude de la fonction f, et est liée aux constantes apparaissant dans l'étude des* inégalités de Poincaré. *La quantité $I(X)$ apparaît également dans l'étude de certains estimateurs via la* borne de Cramer-Rao*; elle donne, en gros, une bonne inférieure sur la qualité d'estimation que l'on peut avoir d'un paramètre dans un modèle paramétrique de loi X.*

On déduit de tous les arguments ci-dessous le théorème suivant.

Théorème 7.1 (Identité de de Bruijn (1918-2012))**.** *Soit $X_\tau = X + \sqrt{\tau} Z$ avec X et Z indépendants. Alors*

$$\partial_\tau H(X_\tau) = \frac{1}{2} I(X_\tau) \tag{7.2}$$

si $Z \sim \mathcal{N}(0,1)$.

Un peu d'exploration autour de (7.2) montrera rapidement que le choix du chemin $X + \sqrt{\tau} Z$ n'est pas primordial pour obtenir une identité telle que (7.2) et, plus généralement, on obtiendra un résultat similaire pour n'importe quelle mixture $X_t = \gamma_1(t) X + \gamma_2(t) Z$ avec γ_1 et γ_2. On déduit

Lemme 7.1. *Soit $t \mapsto \gamma(t) = (\gamma_1(t), \gamma_2(t))$ un chemin C^1 de $(1,0)$ à $(0,1)$ avec coordonnées positives pour tout $t > 0$. Alors*

$$\begin{aligned}&\frac{d}{dt} H(\gamma_1(t) X + \gamma_2(t) Z) \\ &\quad = \frac{\gamma_1'(t)}{\gamma_1(t)} + \left(\gamma_2(t)\gamma_2'(t) - \gamma_2^2(t)\frac{\gamma_1'(t)}{\gamma_1(t)}\right) I(\gamma_1(t) X + \gamma_2(t) Z).\end{aligned} \tag{7.3}$$

Proof. Commençons par remarquer que $H(cX) = H(X) + \log|c|$. On déduit

$$\begin{aligned} H(\gamma_1(t)X + \gamma_2(t)Z) &= \log|\gamma_1(t)| + H\left(X + \frac{\gamma_2(t)}{\gamma_1(t)}Z\right) \\ &= \log(\gamma_1(t)) + H\left(X + \sqrt{\left(\frac{\gamma_2(t)}{\gamma_1(t)}\right)^2}Z\right) \end{aligned}$$

et donc

$$\frac{d}{dt}H(\gamma_1(t)X + \gamma_2(t)Z) = (\log|\gamma_1(t)|)' + \frac{d}{d\tau}H\left(X + \sqrt{\tau}Z\right)\frac{d}{dt}\left(\frac{\gamma_2(t)}{\gamma_1(t)}\right)^2.$$

Remarquons maintenant que $I(aX) = \frac{1}{a^2}I(X)$. En appliquant (7.2), et en nettoyant un peu l'identité ainsi obtenue, on déduit (7.3). □

8 Excursion le long de la Gaussienne

Prenons une fois de plus X dont la loi est telle que

$$\int xf(x)dx = 0 \text{ et } \int x^2 f(x)dx = 1.$$

Dans la suite nous allons nous intéresser à un choix γ_1, γ_2 très particulier, appelé *smart path* en anglais (le chemin malin). Définissons, pour $t \geq 0$, la variable aléatoire

$$X_t = \sqrt{e^{-t}}X + \sqrt{1 - e^{-t}}Z. \tag{8.1}$$

La loi de X_t est telle que $X_0 = X$ et $X_\infty = Z$. De même, si f_t est la densité de X_t on remarque facilement que

$$\int xf_t(x)dx = 0 \text{ et } \int x^2 f_t(x)dx = 1.$$

La variable X_t s'interprète donc effectivement comme une promenade dans la famille des lois de probabilité sur $\mathbb{R}$, de moyenne nulle et de variance unité, promenade qui part de la loi f de X et se termine en la loi ϕ de Z[4]. En appliquant le lemme 7.1 on déduit l'identité

$$\partial_t H(X_t) = \frac{1}{2}\left(I(X_t) - 1\right). \tag{8.2}$$

Explorons le membre de droite de cette dernière équation. On calcule aisément

$$I(Z) = \int \left(\partial_x \log\phi\right)^2 \phi(x) = \int x^2\phi = 1.$$

On a également

$$\begin{aligned} 0 \leq \int \left(\frac{f'(x)}{f(x)} + x\right)^2 f(x)dx &= \int \left(\frac{f'(x)}{f(x)}\right) f(x)dx + 2\int x\frac{f'(x)}{f(x)}f(x)dx + \int x^2 f(x)dx \\ &= I(X) + 2\int xf'(x)dx + 1. \end{aligned}$$

[4] D'aucuns reconnaîtront en (8.1) le semi-groupe d'Ornstein-Uhlenbeck.

Or, par intégration par parties, on a

$$\int xf'(x)dx = [xf(x)]_{-\infty}^{\infty} - \int x^2 f(x)dx = -1$$

donc

$$0 \leq I(X) - 1.$$

L'information de Fisher d'une variable aléatoire de moyenne nulle et de variance 1 est donc toujours supérieure à 1, qui est l'information de la Gaussienne centrée réduite (ce résultat est à comparer avec le théorème 5.1).

Théorème 8.1. *La loi normale ϕ est à information de Fisher minimale parmi toutes les lois de moyenne nulle et de variance 1.*

Remarquons finalement que

$$\int_0^{\infty} \partial_t H(X_t)dt = H(X_\infty) - H(X_0) = H(Z) - H(X)$$

(il faut, en principe, vérifier des conditions techniques pour cette dernière affirmation; ce travail a été réalisé dans [6]). On est enfin en mesure de conclure.

Théorème 8.2 (Formule de de Bruijn, version 2)**.** *Le saut d'entropie est la moitié de l'intégrale du saut d'information le long du semi-groupe d'Ornstein-Uhlenbeck, c'est-à-dire*

$$H(Z) - H(X) = \frac{1}{2}\int_0^{\infty} (I(X_t) - 1)\, dt. \tag{8.3}$$

Les premières preuves rigoureuses de ce résultat sont dues à [3, 7]; cf. également [8]. Il se trouve qu'il est la porte d'entrée pour l'étude du saut d'entropie car, contrairement à l'entropie, l'information de Fisher est un outil très amène aux calculs et est la voie suivie par tous ceux qui ont travaillé sur le sujet. Parmi toutes les illustrations nous ciblons [11, 13, 17] dont nous ne parlerons pas dans ces notes. La fin de cet article est consacrée à une découverte relativement récente dont nous n'avons toujours pas fini d'explorer les conséquences.

9 Caractérisation variationnelle de l'information

Une voie d'accès à l'étude de l'information de Fisher en énonçant un résultat remarquable qui a mené, depuis sa découverte en 2004, a plusieurs découvertes majeures concernant le saut d'entropie $H(Z) - H(X)$.

Théorème 9.1 (Artstein, Ball, Barthe et Naor [1, 4])**.** *Soit $f : \mathbb{R}^n \to [0, \infty]$ une densité deux fois différentiable sur $\mathbb{R}^n$ telle que*

$$\int \frac{\|\nabla f\|^2}{f}, \int \|\mathrm{Hess} f\| < \infty. \tag{9.1}$$

Soit e un vecteur unitaire de $\mathbb{R}^n$ et $h(t) = \int_{te+e^\perp} f$ la marginale de f dans la direction e. Alors

$$I(h) \leq \int_{\mathbb{R}^n} \frac{(\mathrm{div}(pf))^2}{f} \tag{9.2}$$

pour tout champ de vecteur $p : \mathbb{R}^n \to \mathbb{R}^n$ continûment différentiable tel que $\langle p(x), e\rangle = 1$ pour tout x et $\int \|p\| f < \infty$.

Proof. Les conditions sur f nous assurent que

$$h'(t) = \int_{te+e^\perp} \partial_e f.$$

On a également

$$\operatorname{div}(pf) = \partial_e(pf) + \partial_{e^\perp}(pf) = \partial_e(f) + \partial_{e^\perp}(pf),$$

car p est constant dans la direction e. Si $\int_{\mathbb{R}^n} (\operatorname{div}(pf))^2/f < \infty$ alors $\operatorname{div}(pf)$ est intégrable sur $\mathbb{R}^n$ et donc sur presque tout hyperplan perpendiculaire à e. Si, de plus, $\int \|p\| f < \infty$ alors l'intégrale de la divergence $(n-1)$-dimensionnelle de pf dans chacun de ces hyperplans est nulle par le théorème de Gauss-Green. On conclut que

$$h'(t) = \int_{te+e^\perp} \operatorname{div}(pf)$$

Avant de continuer la démonstration illustrons cet argument dans un cas particulier.

Exemple 9.1. *Prenons $n = 2$ et $h(x) = \int f(x,y)dy$. Remarquons que, pour toute fonction $q : \mathbb{R}^2 \to \mathbb{R}$ intégrable on a $\int \partial_y(q(x,y)f(x,y))dy = 0$. Donc*

$$\begin{aligned} h'(x) &= \int_{\mathbb{R}} \partial_x f(x,y)dy \\ &= \int_{\mathbb{R}} \left[\partial_x f(x,y) + \partial_y q(x,y) f(x,y)\right] dy \\ &= \int_{\mathbb{R}} \operatorname{div}(p(x,y)f(x,y))dy \end{aligned}$$

pour $p : \mathbb{R}^2 \to \mathbb{R}^2 : (x,y) \mapsto (1, q(x,y))$.

Poursuivons. On a

$$I(h) = \int \left(\frac{h'(t)}{h(t)}\right)^2 h(t)dt = \int \frac{\left(\int_{te+e^\perp} \operatorname{div}(pf)\right)^2}{\int f}.$$

Or, grâce à l'inégalité de Cauchy-Schwarz, on a également

$$\begin{aligned} \left(\int_{te+e^\perp} \operatorname{div}(pf)\right)^2 &= \left(\int_{te+e^\perp} \frac{\operatorname{div}(pf)}{\sqrt{f}} (\sqrt{f})\right)^2 \\ &\leq \int_{te+e^\perp} \left(\frac{\operatorname{div}(pf)}{\sqrt{f}}\right)^2 \int_{te+e^\perp} (\sqrt{f})^2 \end{aligned}$$

et donc

$$I(h) \leq \int \frac{\int_{te+e^\perp} \left(\frac{\operatorname{div}(pf)}{\sqrt{f}}\right)^2 \left(\int f\right)}{\int f} = \int \frac{\operatorname{div}(pf)^2}{f}.$$

□

La force du théorème 9.1 réside en le fait qu'on peut en gros choisir n'importe quel champ de vecteur dans p car les contraintes imposées sont relativement petites. La preuve d'un résultat sur les l'information de Fisher d'une marginale repose alors construction d'un champ de vecteur approprié. Ceci est loin d'être un exercice facile et dépasse de loin le cadre "élémentaire" du présent article; nous renvoyons le lecteur intéressé aux remarquables articles [1, 2, 4, 5].

10 L'entopie croît le long des convolutions

Nous concluons l'article en illustrant la façon dont on peut utiliser le théorème 9.1 pour apporter une borne utile sur le saut d'information $I(X) - 1$ et donc, ce faisant, sur le saut d'entropie $H(Z) - H(X)$. Soit $f(x_1, \dots, x_n) = f_1(x_1) \dots f_n(x_n)$ la densité jointe des variables aléatoires indépendantes $X_1, \dots, X_n$ de lois respectives $f_1, \dots, f_n$. Soit $(a_1, \dots, a_{n+1}) \in S^n$ un vecteur unitaire. La densité de la variable aléatoire $\sum_{j=1}^{n+1} a_j X_j$ est la marginale de f dans la direction $\bar{a} = (a_1, \dots, a_{n+1})$. On peut donc appliquer l'inégalité (9.2) pour déduite

$$I\left(\sum_{j=1}^{n+1} a_j X_j\right) \leq \int \frac{\operatorname{div}(pf)}{f}$$

pour tout champ de vecteur p tel que $\langle p, \bar{a} \rangle = 1$. En prenant un champ de vecteurs p bien construit (cf [1]) il est possible de déduire

$$I\left(\sum_{j=1}^{n+1} a_j X_j\right) \leq n \sum_{j=1}^{n+1} b_j^2 I\left(\frac{1}{\sqrt{1-a_j^2}} \sum_{i \neq j} a_i X_i\right) \tag{10.1}$$

pour tout b_j tel que $\sum_{j=1}^{n+1} b_j \sqrt{1-a_j^2} = 1$. En particulier on peut prendre $a_i = 1/\sqrt{n+1}$ et $b_i = 1/\sqrt{n(n+1)}$. En plongeant ces choix dans (10.1) on déduit

$$I\left(\frac{1}{\sqrt{n+1}} \sum_{j=1}^{n+1} X_j\right) \leq I\left(\frac{1}{\sqrt{n}} I(X_i)\right). \tag{10.2}$$

Soit maintenant $Y_{n+1} = \frac{1}{\sqrt{n+1}} \sum_{j=1}^{n+1} X_j$ et définissons $Y_{n+1}^{(t)} = \sqrt{e^{-t}} Y_{n+1} + \sqrt{1-e^{-t}} Z$. On peut toujours écrire $Z = \frac{1}{\sqrt{n}} \sum_{j=1}^{n+1} Z_j$ pour $Z_1, \dots, Z_{n+1}$ des Gaussiennes indépendantes. On déduit

$$Y_{n+1}^{(t)} = \sum_{j=1}^{n+1} \frac{1}{\sqrt{n+1}} X_j^{(t)}$$

avec $X_j^{(t)} = \sqrt{e^{-t}} X_j + \sqrt{1-e^{-t}} Z_j$. Reprenant la formule de de Bruijn (8.3) on déduit

$$\begin{aligned}
H(Z) - H(Y_{n+1}) &= \int_0^\infty \left(I\left(Y_{n+1}^{(t)}\right) - 1\right) dt \\
&= \int_0^\infty \left(I\left(\frac{1}{\sqrt{n+1}} \sum_{j=1}^{n+1} X_j^{(t)}\right) - 1\right) dt \\
&\leq \int_0^\infty \left(I\left(\frac{1}{\sqrt{n}} \sum_{j=1}^{n} X_j^{(t)}\right) - 1\right) dt \\
&= \int_0^\infty \left(I\left(Y_n^{(t)}\right) - 1\right) dt = H(Z) - H(Y_n).
\end{aligned}$$

L'entropie croît donc de façon monotone le long des convolutions standardisées.

References

[1] S. Artstein, K. Ball, F. Barthe, and A. Naor. Solution of Shannon's problem on the monotonicity of entropy. *Journal of the American Mathematical Society*, 17(4):975–982, 2004.

[2] S. Artstein, K. M. Ball, F. Barthe, and A. Naor. On the rate of convergence in the entropic central limit theorem. *Probability Theory and Related Fields*, 129(3):381–390, 2004.

[3] D. Bakry and M. Émery. Diffusions hypercontractives. In *Lecture Notes in Math.*, number 1123 in Séminaire de Probabilités XIX, pages 179–206. Springer, 1985.

[4] K. Ball, F. Barthe, and A. Naor. Entropy jumps in the presence of a spectral gap. *Duke Mathematical Journal*, 119(1):41–63, 2003.

[5] K. Ball and V. Nguyen. Entropy jumps for random vectors with log-concave density and spectral gap. *Preprint arxiv:1206.5098v3*, 2012.

[6] A. Barron. Monotonic central limit theorem for densities. Technical report, Technical Report, 1984.

[7] A. R. Barron. Entropy and the central limit theorem. *The Annals of Probability*, 14(1):336–342, 1986.

[8] E. Carlen and A. Soffer. Entropy production by block variable summation and central limit theorems. *Commun. Math. Phys.*, 140(2):339–371, 1991.

[9] T. Cover and J. Thomas. *Elements of Information Theory*, volume Second Edition. Wiley & Sons, New York, 2006.

[10] H. Fischer. *A history of the central limit theorem: from classical to modern probability theory*. Springer, 2010.

[11] D. Guo, S. Shamai, and S. Verdú. Mutual information and minimum mean-square error in gaussian channels. *IEEE Transactions on Information Theory*, 51(4):1261–1282, 2005.

[12] O. Johnson. *Information theory and the central limit theorem*. Imperial College Press, London, 2004.

[13] O. Johnson and A. Barron. Fisher information inequalities and the central limit theorem. *Probability Theory and Related Fields*, 129(3):391–409, 2004.

[14] S. M. Ross. *Introduction to probability models*. Academic press, 2014.

[15] C. E. Shannon. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(1):3–55, 2001.

[16] A. Stam. Some inequalities satisfied by the quantities of information of fisher and shannon. *Information and Control*, 2(2):101–112, 1959.

[17] S. Verdú and D. Guo. A simple proof of the entropy-power inequality. *IEEE Transactions on Information Theory*, 52(5):2165–2166, 2006.