Une version quantitative du second principe de la thermodynamique au travers de la matrice information de Fisher

Steeve ZOZOR

Univ. Grenoble Alpes, CNRS, Grenoble INP*, GIPSA-Lab, 38000 Grenoble, France
*Institute of Engineering Univ. Grenoble Alpes
steeve.zozor@cnrs.fr

Résumé – Dans cet article, nous nous intéressons à la formulation informationnelle du second principe de la thermodynamique. Ce principe stipule qu'un système physique isolé soumis à contraintes tend à se désorganiser au cours du temps pour atteindre son état d'incertitude maximum sous les contraintes. Si sa formulation quantitative via l'entropie est connue, à notre connaissance, la littérature fait peu apparaître de contrepartie via l'information de Fisher. Le but de cet article est précisément de mettre en lumière ce principe sous le prisme de la matrice de Fisher.

Abstract – In this paper, we are interested in the informational formulation of the second law of thermodynamics. This principle states that an isolated physical system subject to constraints tends to become more and more disorganized as time goes on, up to achieve its most uncertain state given the constraints. If its quantitative formulation through the entropy is known, up to our knowledge the literature does not show much of a Fisher information counterpart. The goal of this paper is precisely to shed light on this principle via the prism of the Fisher matrix.

1 Introduction

La notion d'information ou d'incertitude attachée à un vecteur aléatoire est source d'étude permanente et dépendante en général du contexte dans lequel on travaille. Nous nous intéressons ici à deux mesures spécifiques, à savoir l'entropie de Shannon et l'information de Fisher, et leur rôle dans l'énoncé du second principe de la thermodynamique.

La notion d'entropie trouve sa source en thermodynamique entre autre sous l'impulsion de Boltzman, Gibbs ou Maxwell dans l'étude de la cinématique des gaz [1–3]. Elle fut introduite dans le contexte de la transmission d'information par Hartley, Nyquist, Wiener, Laplume, Clavier et en particulier par Shannon que l'histoire a mis en exergue [2, 3]. Cette mesure a le sens d'une mesure d'incertitude. À l'inverse, l'information de Fisher trouve sa source dans le monde statistique sous l'impulsion d'Edgeworth, Doob, Pearson, et plus particulièrement Fisher, dont l'histoire a finalement donné la paternité [2–5]. Cette quantité d'information est fondamentale en estimation dans le sens où son inverse borne la covariance de tout estimateur via l'inégalité de Cramér-Rao [2–5]. Cette mesure d'information trouve écho depuis quelques années dans le monde de la physique, par exemple sous l'impulsion de Frieden [6].

Ces deux quantités, issues de mondes différents, sont reliées par des identités ou inégalités : Fisher comme courbure de la divergence de Kullback-Leibler, identité de de Bruijn pour le canal gaussien additif, inégalité de Stam [2,3]. Diverses inégalités entropiques trouvent également leur contrepartie à la Fisher : inégalité de la puissance entropique, théorème du traitement de l'information par exemple [2, 3, 7].

Dans ce papier, nous nous focalisons sur une formulation quantitative du second principe de la thermodynamique, stipulant qu'un système physique isolé sous contraintes tend à se désorganiser au cours du temps pour atteindre son état d'incertitude maximum étant données les contraintes. Sa formulation usuelle fait appel à l'entropie qui, pour un processus de Markov, sous certaines hypothèses, décroît au cours du temps. Nous nous intéressons ici à sa contrepartie exprimée via la matrice information de Fisher qui, à notre connaissance, est assez peu étudiée dans la littérature (voire pas du tout dans le cadre multidimensionel).

2 Quelques définitions

Nous étudions dans cet article l'évolution temporelle des quantités informationnelles précédemment introduites pour les processus Markoviens $\{\boldsymbol{X}_t\}_{t\geq 0}$ où, à chaque instant, le vecteur aléatoire $\boldsymbol{X}_t$ vit sur un ensemble Lebesgue mesurable $\mathcal{X}_t \in \mathbb{R}^d$ et admet une densité de probabilité $p_{\boldsymbol{X}_t}$ par rapport à la mesure de Lebesgue. Nous rappelons ici leur définition et celles des quantités informationnelles considérées.

Définition 1 (Processus de Markov à temps discret [8])**.** $\{\boldsymbol{X}_t\}_{t\in\mathbb{N}}$ est dit de Markov si les lois conditionnelles vérifient

$$\forall\, t \geq 0, \qquad p_{\boldsymbol{X}_{t+1}|\boldsymbol{X}_t=\boldsymbol{x}_t,\boldsymbol{X}_{t-1}=\boldsymbol{x}_{t-1},\ldots} = p_{\boldsymbol{X}_{t+1}|\boldsymbol{X}_t=\boldsymbol{x}_t},$$

cette dernière étant appelée loi de transition.

Définition 2 (Processus de diffusion [8])**.** $\{\boldsymbol{X}_t\}_{t\geq 0}$ est appelé processus de diffusion (ou de Langevin) s'il satisfait à l'équation différentielle stochastique

$$d\boldsymbol{X}_t = \boldsymbol{b}(\boldsymbol{X}_t, t)\, dt + \boldsymbol{\Sigma}(\boldsymbol{X}_t, t)\, d\boldsymbol{W}_t,$$

avec $\boldsymbol{W}_t$ processus de Wiener [8] d'-dimensionnel et $d' \leq d$, $\boldsymbol{b}$ vecteur de dérive et $\boldsymbol{\Sigma}$ matrice $d \times d'$ de rang plein, paramètre de diffusion ; On appelle $\boldsymbol{D} = \boldsymbol{\Sigma}\,\boldsymbol{\Sigma}^t$ matrice de diffusion, symétrique définie non négative ($\boldsymbol{D} \geq 0$), de rang d'.

Dérive et diffusion caractérisent entièrement le processus, en particulier au travers l'équation de Fokker-Planck :

Théorème 1 (Equation de Fokker-Planck [8])**.** Soit $\{\boldsymbol{X}_t\}_{t\geq 0}$ processus de diffusion, de dérive $\boldsymbol{b}$ et diffusion $\boldsymbol{D}$. Alors la loi de $\boldsymbol{X}_t$ satisfait à l'équation aux dérivées partielles

$$\frac{\partial p}{\partial t} = -\sum_{i=1}^{d} \frac{\partial}{\partial x_i}\big(b_i\, p\big) + \frac{1}{2}\sum_{i,j=1}^{d} \frac{\partial^2}{\partial x_i \partial x_j}\big(D_{i,j}\, p\big).$$

Définition 3 (Entropie différentielle de Shannon [2, 3])**.** Soit p une densité de probabilité sur $\mathcal{X}$. Son entropie différentielle est

$$H(p) = -\int_{\mathcal{X}} p(\boldsymbol{x})\, \log p(\boldsymbol{x})\, d\boldsymbol{x}.$$

H est une mesure d'incertitude attachée à la loi p [2, 3].

Définition 4 (Entropie relative ou divergence de Kullback-Leibler [2, 3])**.** Soit p et q deux distributions telles que le support de q soit inclus dans celui de p, noté $\mathcal{X}$. La divergence de Kullback-Leibler de q relativement à p est définie par

$$D_{\mathrm{kl}}(q\,\|p) = \int_{\mathcal{X}} \log\left(\frac{q(\boldsymbol{x})}{p(\boldsymbol{x})}\right) q(\boldsymbol{x})\, d\boldsymbol{x}$$

(au signe près, entropie de q vue comme densité vis-à-vis de la mesure engendrée par p). On montre alors que $D_{\mathrm{kl}}(q\,\|p) \geq 0$ avec égalité si et seulement si $q = p$ presque partout.

C'est une mesure (informationnelle) d'écart de q vis-à-vis de p. Non symétrique (p sert de référence) et ne satisfaisant pas à l'inégalité triangulaire, ce n'est pas une distance.

Par ailleurs, le principe suivant, dit d'entropie maximale, est intimement lié au second principe de la thermodynamique [1] :

Théorème 2 (Entropie maximale sous contraintes [1–3, 5])**.** Soit $\boldsymbol{S}$ une statistique k-dimensionnelle (ex. $\boldsymbol{S}(x) = [x \quad x^2]^t$), $\mathcal{P}_{\boldsymbol{S}}$ ensemble de lois sur $\mathcal{X}$ de même statistique $\boldsymbol{S}$. Alors la loi d'entropie maximale sous contrainte $\boldsymbol{S}$ est de la forme

$$p_{\mathrm{em}}(\boldsymbol{x}) \equiv \left[\underset{p\in\mathcal{P}_{\boldsymbol{S}}}{\operatorname{argmax}}\, H(p)\right](\boldsymbol{x}) = e^{\boldsymbol{\eta}^t \boldsymbol{S}(\boldsymbol{x}) - \varphi(\boldsymbol{\eta})},$$

loi de la famille exponentielle [2, 4, 5]. $\boldsymbol{\eta}$ et φ sont telles que contrainte et normalisation soient satisfaites.

Démonstration. La preuve fait appel à la technique d'optimisation variationnelle sous containtes [2, 9] et ne sera par détaillée ici. Une preuve informationnelle part de la solution et s'appuie sur la positivité de $D_{\mathrm{kl}}(p\,\|p_{\mathrm{em}})$ [2]. □

Définition 5 (Matrice information de Fisher)**.** Soit $\{p(\cdot;\boldsymbol{\theta})\}_{\boldsymbol{\theta}\in\Theta}$ une famille de loi paramétrée par $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^m$ telle que p soit différentiable en $\boldsymbol{\theta}$. La matrice de Fisher paramétrique, de dimension $m \times m$, est définie par

$$\boldsymbol{J}_{\boldsymbol{\theta}}(p) = \int_{\mathcal{X}} \frac{\partial \log p(\boldsymbol{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\, \frac{\partial \log p(\boldsymbol{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^t}\, p(\boldsymbol{x};\boldsymbol{\theta})\, d\boldsymbol{x},$$

matrice de covariance de la fonction score $\boldsymbol{\mathcal{S}}(\boldsymbol{\theta}) = \frac{\partial}{\partial\boldsymbol{\theta}} \log p(\boldsymbol{X};\boldsymbol{\theta})$, dérivée de la log-vraisemblance $\log p$. Par construction, $\boldsymbol{J}_{\boldsymbol{\theta}}(p) \geq 0$ symétrique définie non négative.

Quand $\mathcal{X}$ ne dépend pas de $\boldsymbol{\theta}$, sous conditions de regularité,

$$\boldsymbol{J}_{\boldsymbol{\theta}}(p) = -\int_{\mathcal{X}} \frac{\partial^2 \log p(\boldsymbol{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}^t}\, p(\boldsymbol{x};\boldsymbol{\theta})\, d\boldsymbol{x},$$

moyenne de la Hessienne de la log-vraisemblance.

Si $\boldsymbol{\theta}$ est un paramètre de position, $p(\boldsymbol{x};\boldsymbol{\theta}) \equiv p(\boldsymbol{x} - \boldsymbol{\theta})$, $\boldsymbol{J}_{\boldsymbol{\theta}}$ est indépendante de $\boldsymbol{\theta}$ et, par ailleurs, différentier en $\boldsymbol{x}$ ou en $\boldsymbol{\theta}$ est équivalent. On notera $\boldsymbol{J}(p)$ la matrice obtenue, dite Fisher non paramétrique.

$\boldsymbol{J}_{\boldsymbol{\theta}}$ est une mesure d'information sur $\boldsymbol{\theta}$, ce qui peut se voir au travers de l'inégalité de Cramér-Rao qui borne la covariance de tout estimateur (non biaisé) de $\boldsymbol{\theta}$ par $\boldsymbol{J}_{\boldsymbol{\theta}}^{-1}$ [2–4, 6].

3 Version entropique : rappels

Dans le cadre de processus de Markov à temps discret, on trouve dans la littérature l'équivalent du fameux second principe de la thermodynamique [1, 2] :

Théorème 3 (Version entropique – temps discret)**.** Soit $\{\boldsymbol{X}_t\}_{t\in\mathbb{N}}$ processus de Markov de loi de transition donnée, independante de la condition initiale $\boldsymbol{X}_0$. Soient deux distributions initiales q_0 et p_0, conduisant à q_t et p_t pour $\boldsymbol{X}_t$. Alors :

- $\forall\, t \geq 0, \qquad D_{\mathrm{kl}}(q_{t+1}\,\|p_{t+1}) \leq D_{\mathrm{kl}}(q_t\,\|p_t)$:
  q_t et p_t tendent à se rapprocher au cours du temps.
- Si p est distribution stationnaire,
  $$D_{\mathrm{kl}}(q_{t+1}\,\|p) \leq D_{\mathrm{kl}}(q_t\,\|p) \ :$$
  q_t tend à se rapprocher d'une loi stationnaire.
- Si les $\boldsymbol{X}_t$ ont une statistique $\boldsymbol{S}$ indépendante de t et la loi d'entropie maximale sous contrainte $\boldsymbol{S}$ est stationnaire,
  $$H(q_{t+1}) \geq H(q_t) \ :$$
  le système tend à se désorganiser (et atteindre celui d'entropie maximale sous contraintes $\boldsymbol{S}$).

Démonstration. On trouvera une preuve dans [2]. Il s'agit (i) d'écrire la divergence des lois jointes, $D_{\mathrm{kl}}\big(q_{\boldsymbol{X}_{t+1},\boldsymbol{X}_t}\,\|p_{\boldsymbol{X}_{t+1},\boldsymbol{X}_t}\big)$ en écrivant respectivement $p_{\boldsymbol{X}_{t+1},\boldsymbol{X}_t}(\boldsymbol{x}_{t+1},\boldsymbol{x}_t) = p_{\boldsymbol{X}_{t+1}|\boldsymbol{X}_t=\boldsymbol{x}_t}(\boldsymbol{x}_{t+1}) p_{\boldsymbol{X}_t}(\boldsymbol{x}_t) = p_{\boldsymbol{X}_t|\boldsymbol{X}_{t+1}=\boldsymbol{x}_{t+1}}(\boldsymbol{x}_t) p_{\boldsymbol{X}_{t+1}}(\boldsymbol{x}_{t+1})$ (et de même pour q), (ii) d'utiliser le fait que $p_{\boldsymbol{X}_{t+1}|\boldsymbol{X}_t=\boldsymbol{x}_t} = q_{\boldsymbol{X}_{t+1}|\boldsymbol{X}_t=\boldsymbol{x}_t}$ (ce qui n'est en général pas vrai pour la transition de $t+1$ vers t), pour obtenir $D_{\mathrm{kl}}(q_t\,\|p_t) = D_{\mathrm{kl}}(q_{t+1}\,\|p_{t+1}) + \int_{\mathcal{X}_{t+1}} D_{\mathrm{kl}}\big(q_{\boldsymbol{X}_t|\boldsymbol{X}_{t+1}=\boldsymbol{x}}\,\big\|p_{\boldsymbol{X}_t|\boldsymbol{X}_{t+1}=\boldsymbol{x}}\big)\, q_{t+1}(\boldsymbol{x})\, d\boldsymbol{x}$. On conclut de $D_{\mathrm{kl}}\big(q_{\boldsymbol{X}_t|\boldsymbol{X}_{t+1}=\boldsymbol{x}}\,\big\|p_{\boldsymbol{X}_t|\boldsymbol{X}_{t+1}=\boldsymbol{x}}\big) \geq 0$. Le second item s'obtient de $p_t = p$ et le dernier en remplaçant $p = p_{\mathrm{em}}$, par son expression théorème 2. □

Historiquement, la paternité de la version entropique du principe revient à Boltzman et de son étude de la repartition des vitesses dans un gaz, via le théorème H [6]. Dans le cadre qui nous intéresse, elle repose sur le résultat qui suit :

Théorème 4 (Variations temporelles entropiques)**.** Soit $\{\boldsymbol{X}_t\}_{t\geq 0}$ processus de diffusion, deux conditions initiales q_0, p_0 et q_t, p_t les deux distributions correspondantes pour $\boldsymbol{X}_t$. Supposons de plus $\mathcal{X}_t = \mathcal{X}$ indépendant du temps. Alors

$$\frac{d}{dt} D_{\mathrm{kl}}(q_t \,\|p_t\,) = -\frac{1}{2} \operatorname{Tr} \int_{\mathcal{X}} \boldsymbol{D}(\boldsymbol{x},t)\, \frac{\partial \log \frac{q_t(\boldsymbol{x})}{p_t(\boldsymbol{x})}}{\partial \boldsymbol{x}}\, \frac{\partial \log \frac{q_t(\boldsymbol{x})}{p_t(\boldsymbol{x})}}{\partial \boldsymbol{x}^t}\, q_t(\boldsymbol{x})\, d\boldsymbol{x}$$

où l'argument de la trace Tr peut être interprété comme une matrice (divergence) de Fisher "de noyau $\boldsymbol{D}$" [10]. Par ailleurs,

$$\frac{d}{dt} H(p_t) = \operatorname{Tr} \int_{\mathcal{X}} \frac{\partial \boldsymbol{b}(\boldsymbol{x},t)}{\partial \boldsymbol{x}^t}\, p_t(\boldsymbol{x})\, d\boldsymbol{x} - \frac{1}{2} \operatorname{Tr} \int_{\mathcal{X}} \boldsymbol{D}(\boldsymbol{x},t)\, \frac{\partial^2 \log p_t(\boldsymbol{x})}{\partial \boldsymbol{x}\, \partial \boldsymbol{x}^t}\, p_t(\boldsymbol{x})\, d\boldsymbol{x}$$

En particulier, quand la dérive ne dépend pas des états, $\boldsymbol{b}(\boldsymbol{x},t) \equiv \boldsymbol{b}(t)$, le premier terme s'annule et si de plus la diffusion ne dépend pas non plus des états, $\boldsymbol{D}(\boldsymbol{x},t) \equiv \boldsymbol{D}(t)$,

$$\frac{d}{dt} H(p_t) = \frac{1}{2} \operatorname{Tr} \left(\boldsymbol{D}(t) \boldsymbol{J}(p_t) \right)$$

qui n'est autre qu'une version de l'identité de de Bruijn [2, 3].

Démonstration. Ces résutats sont les versions multidimensionnelles de [10]. La version en divergence se prouve tout d'abord par permutation de l'intégration et dérivation temporelle (on suppose les conditions requises). Puis, remarquant que $\displaystyle\int_{\mathcal{X}} \frac{\partial q_t(\boldsymbol{x})}{\partial t}\, d\boldsymbol{x} = \frac{d}{dt} \int_{\mathcal{X}} q_t(\boldsymbol{x})\, d\boldsymbol{x} = \frac{d}{dt} 1 = 0$, on obtient $\displaystyle\frac{d}{dt} D_{\mathrm{kl}}(q_t \,\|p_t\,) = \int_{\mathcal{X}} \left(\frac{\partial q_t(\boldsymbol{x})}{\partial t} \log v_t(\boldsymbol{x}) - \frac{\partial p_t(\boldsymbol{x})}{\partial t}\, v_t(\boldsymbol{x}) \right) d\boldsymbol{x}$ avec $v_t = \frac{q_t}{p_t}$, rapport de vraisemblances. On remplace alors les dérivées temporelles par leur expression issue de l'équation de Fokker-Planck, ce qui donne une expression de la forme $\frac{d}{dt} D_{\mathrm{kl}}(q_t \,\|p_t\,) = B_t + \Delta_t$ où le premier terme est issu de la dérive, et le second de la diffusion. On montre alors que l'intégrande de B_t s'écrit sous la forme $\sum_i \frac{\partial}{\partial x_i} \big(b_i\, q_t (\log v_t - 1) \big)$ et, par le théorème de la divergence, on en déduit que $B_t = 0$. Le terme de diffusion fait apparaître la double somme de $\frac{\partial^2 (D_{i,j}\, q_t)}{\partial x_i \partial x_j} \log v_t - \frac{\partial^2 (D_{i,j}\, p_t)}{\partial x_i \partial x_j} v_t$, qui s'écrit aussi $\displaystyle\frac{\partial^2 (D_{i,j}\, q_t\, (\log v_t - 1))}{\partial x_i \partial x_j} - \frac{\partial v_t}{\partial x_i} \left(\frac{1}{v_t} \frac{\partial (D_{i,j}\, q_t)}{\partial x_j} - \frac{\partial (D_{i,j}\, p_t)}{\partial x_j} \right)$. De par le théorème de la divergence, l'intégrale impliquant le permier terme différentiel s'annule. On aboutit alors au résultat en constatant que $\frac{1}{v_t} \frac{\partial (D_{i,j}\, q_t)}{\partial x_j} - \frac{\partial (D_{i,j}\, p_t)}{\partial x_j} = \frac{\partial v_t}{\partial x_j} D_{i,j} \frac{p_t}{v_t}$ et de $\frac{\partial v_t}{\partial x_i} = v_t \frac{\partial \log v_t}{\partial x_i}$.

Le cas de l'entropie se traite de la même manière. □

De ce théorème on en déduit le fameux théorème H :

Corollaire 1 (Théorème H)**.** Sous les conditions du théorème 4,

— $\forall t \geq 0, \qquad \dfrac{d}{dt} D_{\mathrm{kl}}(q_t \,\|p_t\,) \leq 0.$

— S'il existe une distribution p stationnaire pour l'équation de Fokker-Planck,
$$\frac{d}{dt} D_{\mathrm{kl}}(q_t \,\|p\,) \leq 0.$$

— Si les $\boldsymbol{X}_t$ ont une statistique $\boldsymbol{S}$ indépendante du temps et si p_{em}, loi d'entropie maximale sous contrainte $\boldsymbol{S}$ est stationnaire, alors
$$\frac{d}{dt} H(q_t) \geq 0.$$

Démonstration. Le premier item est conséquence de la forme de la variation temporelle de la divergence, théorème 4, en notant que $\boldsymbol{D} \geq 0$. Le second s'obtient avec $p_t = p$, et le dernier de par la forme de $p = p_{\mathrm{em}}$, théorème 2. □

Si ces formulations entropiques sont bien connues, il se trouve que le second principe de la thermodynamique peut aussi s'exprimer quantitativement via la matrice information de Fisher. Nous étendons ici les versions scalaires de [6, 11] au cas multidimensionnel, donnant par ailleurs une expression explicite du taux de variation de la matrice information de Fisher.

4 Une version via la matrice de Fisher

Théorème 5 (Version à la Fisher – temps discret)**.** Soit $\{\boldsymbol{X}_t\}_{t\in\mathbb{N}}$ processus de Markov, avec la loi de transition donnée, indépendante de la condition initiale $\boldsymbol{X}_0$. Considérons une famille $\{p_0(\cdot;\boldsymbol{\theta})\}_{\boldsymbol{\theta}\in\Theta}$ de conditions initiales, paramétrée par $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^m$ et supposons que pour tout t, $p_t(\cdot;\boldsymbol{\theta})$ soit différentiable en $\boldsymbol{\theta}$. Alors

$$\forall\, t \geq 0, \quad \boldsymbol{J}_{\boldsymbol{\theta}}(p_{t+1}) \leq \boldsymbol{J}_{\boldsymbol{\theta}}(p_t) \qquad (A \geq B \Leftrightarrow A - B \geq 0)$$

Démonstration. On peut par exemple partir de la forme de $\boldsymbol{J}_{\boldsymbol{\theta}}(p_{\boldsymbol{X}_{t+1},\boldsymbol{X}_t})$ via la hessienne de la log-vraisemblance de la loi jointe, en écrivant respectivement $p_{\boldsymbol{X}_{t+1},\boldsymbol{X}_t}(\boldsymbol{x}_{t+1},\boldsymbol{x}_t) = p_{\boldsymbol{X}_{t+1}|\boldsymbol{X}_t=\boldsymbol{x}_t}(\boldsymbol{x}_{t+1}) p_{\boldsymbol{X}_t}(\boldsymbol{x}_t) = p_{\boldsymbol{X}_t|\boldsymbol{X}_{t+1}=\boldsymbol{x}_{t+1}}(\boldsymbol{x}_t) p_{\boldsymbol{X}_{t+1}}(\boldsymbol{x}_{t+1})$. De la non dépendance de la transition $p_{\boldsymbol{X}_{t+1}|\boldsymbol{X}_t=\boldsymbol{x}_t}$ de la condition initiale, elle est également indépendante de $\boldsymbol{\theta}$, de sorte que $\frac{\partial^2 \log p_{\boldsymbol{X}_{t+1}|\boldsymbol{X}_t=\boldsymbol{x}_t}(\boldsymbol{x}_{t+1})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^t} = 0$. On aboutit alors à, $\displaystyle\boldsymbol{J}_{\boldsymbol{\theta}}(p_t) = \boldsymbol{J}_{\boldsymbol{\theta}}(p_{t+1}) + \int_{\mathcal{X}_{t+1}} \boldsymbol{J}_{\boldsymbol{\theta}}(p_{\boldsymbol{X}_t|\boldsymbol{X}_{t+1}=\boldsymbol{x}})\, p_{n+1}(\boldsymbol{x};\boldsymbol{\theta})\, d\boldsymbol{x}$ et on conclut de $\boldsymbol{J}_{\boldsymbol{\theta}}(p_{X_n|X_{n+1}=x_{n+1}}) \geq 0$. □

À l'image du théorème 4, on peut également caractériser les variations temporelles de la matrice de Fisher pour un processus de diffusion. On trouvera des versions pour des contextes particuliers (dérive et/ou diffusion spécifiques) dans [12, § 5].

Théorème 6 (Variation temporelle de la matrice de Fisher)**.** Soit $\{\boldsymbol{X}_t\}_{t\geq 0}$ processus de diffusion et $\{p_0(\cdot;\boldsymbol{\theta})\}_{\boldsymbol{\theta}\in\Theta}$ famille paramétrique de conditions initiales, et supposons que pour tout t la loi p_t est différentiable en $\boldsymbol{\theta}$. Alors

$$\frac{d}{dt} \boldsymbol{J}_{\boldsymbol{\theta}}(p_t) = -\int_{\mathcal{X}} \frac{\partial^2 \log p_t(\boldsymbol{x};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{x}^t} \boldsymbol{D}(\boldsymbol{x},t) \frac{\partial^2 \log p_t(\boldsymbol{x};\boldsymbol{\theta})}{\partial\boldsymbol{x}\partial\boldsymbol{\theta}^t} p_t(\boldsymbol{x}) d\boldsymbol{x}$$

Démonstration. La preuve étend celle de [11] au cas multidimensionnel (non immédiat) et consiste à dériver l'expression $\boldsymbol{J}_{\boldsymbol{\theta}}(p_t) = \int_{\mathcal{X}} \frac{\frac{\partial p_t(\boldsymbol{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial p_t(\boldsymbol{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^t}}{p_t(\boldsymbol{x};\boldsymbol{\theta})} d\boldsymbol{x}$, permuter dérivée en t et intégration, puis dérivée en t et en $\boldsymbol{\theta}$, et remplacer les dérivées temporelles par leur expression issue de l'équation de Fokker-Planck. On obtient alors une expression de la forme $\frac{d}{dt}\boldsymbol{J}_{\boldsymbol{\theta}}\left(p_t\right) = \boldsymbol{B}_t(\boldsymbol{\theta}) + \boldsymbol{\Delta}_t(\boldsymbol{\theta})$ où le premier terme résulte de la dérive, et le second de la diffusion. En explicitant l'intégrande du terme issue de la dérive, se servant du fait que la dérive ne dépend pas du paramètre (la famille satisfaisant à la même équation de Fokker-Planck), on l'écrit comme un terme de divergence et on évoque le théorème de la divergence pour montrer que $\boldsymbol{B}_t = 0$. Puis, en manipulant les termes issus de la diffusion, et en utilisant la symétrie de $\boldsymbol{D}$, sa non dépendance en $\boldsymbol{\theta}$ et des permutations d'indices, on montre que $\boldsymbol{\Delta}_t(\boldsymbol{\theta}) = \frac{1}{2}\int_{\mathcal{X}} \sum_{i,j=1}^{d} \left(\frac{\partial^2 (D_{i,j}\, p_t\, \boldsymbol{\mathcal{S}}_{\boldsymbol{\theta}} \boldsymbol{\mathcal{S}}_{\boldsymbol{\theta}}^t)}{\partial x_i\, \partial x_j} - 2\, D_{i,j}\, p_t\, \frac{\partial \boldsymbol{\mathcal{S}}_{\boldsymbol{\theta}}}{\partial x_i} \frac{\partial \boldsymbol{\mathcal{S}}_{\boldsymbol{\theta}}^t}{\partial x_j} \right) d\boldsymbol{x}$ avec $\boldsymbol{\mathcal{S}}_{\boldsymbol{\theta}}$ fonction score. Composante à composante, on reconnait dans le premier intégrande la trace d'une hessienne, de sorte que le premier terme dans $\boldsymbol{\Delta}_t$ s'annule et ainsi $\boldsymbol{\Delta}_t(\boldsymbol{\theta}) = -\int_{\mathcal{X}} \sum_{i,j=1}^{d} \frac{\partial \boldsymbol{\mathcal{S}}_{\boldsymbol{\theta}}}{\partial x_i}\, D_{i,j}\, \frac{\partial \boldsymbol{\mathcal{S}}_{\boldsymbol{\theta}}^t}{\partial x_j}\, p_t(\boldsymbol{x})\, d\boldsymbol{x}$: c'est précisement la formulation du théorème. □

On notera que pour $\boldsymbol{\theta}$ paramètre de position, si toutefois dérive et diffusion ne dépendent pas de l'état (sinon dépendants du paramètre), on obtient le même résultat pour la matrice non-paramétrique, avec $\boldsymbol{D}(\boldsymbol{x},t) \equiv \boldsymbol{D}(t)$ et la hessienne de la log-vraisemblance en lieu et place de la dérivé double en $\boldsymbol{\theta}$ et $\boldsymbol{x}^t$.

On déduit immédiatement du théorème 6 et de $\boldsymbol{D} \geq 0$ l'équivalent Fisher du théorème H :

Corollaire 2 (Théorème $\boldsymbol{J}$)**.** Sous les conditions du théorème 6,

$$\frac{d}{dt}\boldsymbol{J}_{\boldsymbol{\theta}}(p_t) \leq 0.$$

5 Discussions

Dans cet article, nous avons revisité la formulation du second principe de la thermodynamique sous le prisme de la matrice de Fisher. Contrairement à l'entropie, la matrice de Fisher est multidimensionnelle, plus riche informationnellement que l'entropie. Contrairement à l'entropie, cette mesure trouve sa source dans le monde des statistiques ou de l'estimation. L'expression du second principe de la thermodynamique via cette mesure lui confère a posteriori un sens physique, ce que tend à montrer également les travaux de Frieden [6].

Une question sous-jacente à ce principe, perspective directe de ce papier, est dès lors la contrepartie à la Fisher du principe d'entropie maximale (matrice de Fisher minimale au sens défini non négatif). Ce principe est étudié depuis longtemps dans le cadre scalaire [13, 14] mais comme problème sans solution, sauf cas particuliers. Le problème est revenu d'actualité sous l'impulsion de physiciens intéressés par le développement d'une physique statistique basée sur l'information de Fisher [6, 15], mais aussi dans le cadre des statistiques ou du traitement du signal [16, 17]. Toutefois, à notre connaissance, seules des versions scalaires sont étudiées et l'étude dans un cadre général matriciel semble rester question ouverte.

Références

[1] N. Merhav. *Statistical Physics for Electrical Engineering*. Springer, 2018.

[2] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, 2nd edition, 2006.

[3] O. Rioul. *Théorie de l'information et du codage*. Lavoisier, 2007.

[4] S. M. Kay. *Fundamentals for Statistical Signal Processing : Estimation Theory*. vol. 1. Prentice Hall, 1993.

[5] C. P. Robert. *The Bayesian Choice. From Decision-Theoretic Foundations to Computational Implementation*. Springer, 2nd edition, 2007.

[6] B. R. Frieden. *Science from Fisher Information : A Unification*. Cambridge University Press, 2004.

[7] O. Rioul. Information theoretic proofs of entropy power inequalities. *IEEE Trans. on IT*, 57(1) :33–55, Jan. 2011.

[8] H. Risken. *The Fokker-Planck Equation*. Springer, 1996.

[9] B. van Brunt. *The Calculus of Variations*. Springer, 2004.

[10] A. Wibisono, V. Jog, and P.-L. Loh. Information and estimation in Fokker-Planck channels. In *IEEE Int. Symp. on Inf. Th.*, pages 2673–2677, Aachen, Germany, 2017.

[11] A. R. Plastino and A. Plastino. Symmetries of the Fokker-Planck equation and the Fisher-Frieden arrow of time. *Phys. Rev. E*, 54(4) :4423–4426, Oct. 1996.

[12] P. Garbaczewski. Differential entropy and dynamics of uncertainty. *J. Stat. Phys.*, 123(2) :315–355, Apr. 2006.

[13] P. J. Huber. Robust estimation of a location parameter. *Annals of Math Stat.*, 35(1) :73–101, Mar. 1964.

[14] A. M. Kagan, Y. V. Linnik, and C. R. Rao. *Characterization problems in mathematical statistics*. John Wiley & Sons, 1973.

[15] B.R. Frieden, A. Plastino, A.R. Plastino, and B.H. Soffer. Non-equilibrium thermodynamics and Fisher information : An illustrative example. *Phys. Let. A*, 304(3-4) :73–78, Nov. 2002.

[16] J. M. Borwein, A. S. Lewis, and D. Noll. Maximum entropy reconstruction using derivative information, part 1 : Fisher information and convex duality. *Math. Op. Res.*, 21(2) :442–468, May 1996.

[17] V. Živojnović. Minimum Fisher information of moment-constrained distributions with application to robust blind identification. *Sig. Proc.*, 65(2) :297–13, Mar. 1998.