CONSTRUCTION DE MESURES DE DIVERGENCE. APPLICATION À LA RÉSOLUTION DE PROBLÈMES INVERSES LINÉAIRES.

J.-F. Bercher et C. Heinrich

Laboratoire des signaux et systèmes (cnrs-ese-ups)
Plateau de Moulon
91192 Gif-sur-Yvette Cedex

e_mail : memm@lss.supelec.fr

Ce document est une version « longue » et nettement plus détaillée d'une communication de même titre présentée au colloque gretsi, *septembre 1995. Il est encore incomplet : nous envisageons encore d'y inclure quelques résultats et parallèles avec la théorie des grandes déviations, de présenter les résultats de limite conditionnelle de* Van Campenhout *et* Cover, *et de développer la partie application aux problèmes inverses.*

RÉSUMÉ

Nous nous proposons dans cet article de définir une mesure de l'information associée aux éléments d'un ensemble convexe $\mathcal{C}$ muni d'une mesure de référence μ. Nous énoncerons quelques propriétés simples que devrait présenter une telle mesure d'information, et nous identifions alors celle-ci à la transformée de Cramér de la mesure de référence μ.

Nous monterons que cette mesure d'information est liée à l'information de Kullback-Leibler et aux familles exponentielles. Nous soulignerons également quelques liens avec les entropies et information de Rényi.

Nous chercherons ensuite à définir des divergences. Nous pourrons en particulier faire apparaître des divergences de Bregman et des f-divergences.

Nombre de mesures connues seront interprétées dans le cadre proposé. Nous examinerons enfin l'application de ces résultats à la résolution de problèmes inverses.

ABSTRACT

Nous nous proposons dans cet article de définir une mesure de l'information associée aux éléments d'un ensemble convexe $\mathcal{C}$ muni d'une mesure de référence μ. Nous énoncerons quelques propriétés simples que devrait présenter une telle mesure d'information, et nous identifions alors celle-ci à la transformée de Cramér de la mesure de référence μ.

Nous monterons que cette mesure d'information est liée à l'information de Kullback-Leibler et aux familles exponentielles. Nous soulignerons également quelques liens avec les entropies et information de Rényi.

Nous chercherons ensuite à définir des divergences. Nous pourrons en particulier faire apparaître des divergences de Bregman et des f-divergences.

Nombre de mesures connues seront interprétées dans le cadre proposé. Nous examinerons enfin l'application de ces résultats à la résolution de problèmes inverses.

1 Introduction

L'utilisation de mesures d'information ou de mesures de gain d'information en traitement du signal et de l'image est multiple, que ce soit en communication, codage et compression, détection, estimation, voir par exemple (Basseville 1989). Nous nous proposons ici de chercher à définir une mesure de l'information associée aux éléments d'un certain ensemble convexe $\mathcal{C}$ muni d'une mesure de référence μ, avec une perpective, — non limitative — d'application pour la résolution de problèmes inverses linéaires. Pour ce faire, nous énoncerons quelques propriétés simples que devrait présenter une telle mesure d'information, et nous identifions alors celle-ci à la transformée de Cramér de la mesure de référence μ. Cette démarche s'inspire des constructions axiomatiques de (Shannon 1948), de (Shore & Johnson 1980) qui permettent de définir l'entropie (de Shannon) comme mesure d'information et critère d'inférence, et des travaux plus récents de (Csiszár 1991).

Nous monterons ensuite comment ces mesures d'information sont liées à l'information de Kullback-Leibler et aux familles exponentielles. Nous soulignerons également quelques liens avec les entropies et information de Rényi.

Disposant de mesures d'information, ou entropies, nous chercherons à définir des gains d'information, ou divergences. En particulier, nous pourrons faire apparaître des divergences de Bregman (Bregman 1967) et nous soulignerons certaines de l eurs propriétés. Nous pourrons également faire apparaître des f-divergences (Csiszár 1967).

Nombre de mesures connues peuvent être interprétées dans le cadre proposé : nous montrerons comment l'on peut déterminer pratiquement ces entropies, à partir de la spécification de $\mathcal{C}$ et μ, et nous retrouvons par exemple les entropies de Shannon, de Burg, de Fermi-Dirac, de Bose-Einstein.

Nous examinerons enfin l'application de ces résultats à la résolution de problèmes inverses. Une annexe collecte enfin quelques résultats d'analyse convexe utiles lors de l'exposé.

2 Quelques propriétés désirables

Soit μ une mesure de référence, et $\mathcal{C}$ l'enveloppe convexe fermée du support de μ. Pour simplifier, on supposera ici que μ est une mesure de probabilité. On note $\overline{\boldsymbol{x}}_\mu$ la moyenne sous cette mesure de probabilité, $\overline{\boldsymbol{x}}_\mu = \mathrm{E}_\mu\{\boldsymbol{x}\}$, et on définit par $I_\mu(\boldsymbol{x})$ « l'information » associée à un élément $\boldsymbol{x}$ de $\mathcal{C}$. Afin de rechercher cette mesure d'information, nous énonçons ci-dessous quelques propriétés intuitives et désirables que devrait vérifier cette mesure d'information.

• 1. Positivité. L'information est une quantité positive : $I_\mu(\boldsymbol{x}) \geq 0$.
• 2. Minimum. L'information est nulle, et minimale, pour la moyenne sous μ. Connaissant $\mathcal{C}$ et μ, $\overline{\boldsymbol{x}}$ est l'objet par défaut dont l'observation n'apporte aucune information, ou dont la sélection ne nécessite aucun apport d'information :

$$\operatorname*{Inf}_{\boldsymbol{x}} I_\mu(\boldsymbol{x}) = I_\mu(\overline{\boldsymbol{x}}_\mu) = 0.$$

• 3. Convexité. L'information est une fonction convexe :

$$I_\mu(\alpha\boldsymbol{x}_1 + (1-\alpha)\boldsymbol{x}_2) \leq \alpha I_\mu(\boldsymbol{x}_1) + (1-\alpha) I_\mu(\boldsymbol{x}_2).$$

Cette propriété de convexité n'est pas immédiatement liée à des notions d'information. Notons cependant que Csiszár (Csiszár 1991, Théorème 1, page 2044) obtient cette propriété de convexité comme conséquence de propriétés de localité et de régularité d'une règle de sélection.

• 4. Indépendance. Si la mesure est séparable en deux blocs, *i.e.* $\mu(\boldsymbol{x}) = \mu_1(\boldsymbol{x}_1)\mu_2(\boldsymbol{x}_2)$, alors l'information globale est la somme des informations définies sur les blocs indépendants :

$$I_\mu(\boldsymbol{x}) = I_{\mu_1}(\boldsymbol{x}_1) + I_{\mu_2}(\boldsymbol{x}_2).$$

• 5. Additivité. Soit $\boldsymbol{x}$ la somme $\boldsymbol{x} = \boldsymbol{x}_1 + \boldsymbol{x}_2$, où $\boldsymbol{x}_1$ et $\boldsymbol{x}_2$ appartiennent respectivement à deux convexes $\mathcal{C}_1$ et $\mathcal{C}_2$ munis des mesures μ_1 et μ_2. L'information apportée par cette somme est inférieure à la somme des informations considérées indépendamment :

$$I_{\mu_1 * \mu_2}(\boldsymbol{x}) \leq I_{\mu_1}(\boldsymbol{x}_1) + I_{\mu_2}(\boldsymbol{x}_2).$$

• 6. Échelle (homogénéité). Le changement d'échelle $\boldsymbol{x} \to a\boldsymbol{x}$, $\mu \to \mu'$ ne modifie pas l'information : $I_{\mu'}(ax) = I_\mu(x)$.

2.1 Conséquences de ces propriétés

La propriété de positivité est induite par les propriétés 2 et 3 : si l'information est minimale et nulle pour la moyenne $\overline{\boldsymbol{x}}_\mu$, et est une fonction convexe, alors elle est nécessairement non négative.

On peut assurer que I_μ soit convexe en l'exprimant comme la convexe conjuguée d'une fonction continue ϕ_μ (voir l'annexe)

$$I_\mu(\boldsymbol{x}) = \operatorname*{Sup}_{\boldsymbol{s}} \left\{ \boldsymbol{s}^{\mathrm{t}}\boldsymbol{x} - \phi_\mu(\boldsymbol{s}) \right\}.$$

L'unicité de $\phi_\mu(\boldsymbol{s})$ est en outre assurée si $\phi_\mu(\boldsymbol{s})$ est convexe fermée, et la relation précédente est réversible : $\phi_\mu(\boldsymbol{s}) = \operatorname*{Sup}_{\boldsymbol{x}} \left\{ \boldsymbol{s}^{\mathrm{t}}\boldsymbol{x} - I_\mu(\boldsymbol{x}) \right\}$.

L'information peut alors encore s'écrire

$$I_\mu(\boldsymbol{x}) = \boldsymbol{s}_{\boldsymbol{x}}^{\mathrm{t}}\boldsymbol{x} - \phi_\mu(\boldsymbol{s}_{\boldsymbol{x}}),$$

pour la valeur $\boldsymbol{s}_{\boldsymbol{x}}$ rendant maximum l'argument du Sup, c'est-à-dire vérifiant $\boldsymbol{x} - \phi_\mu'(\boldsymbol{s}_{\boldsymbol{x}}) = \boldsymbol{0}$. Son gradient est alors simplement $I_\mu'(\boldsymbol{x}) = \boldsymbol{s}_{\boldsymbol{x}}$.

La propriété 2, minimum et nullité pour la moyenne, entraîne que $\boldsymbol{s}_{\overline{\boldsymbol{x}}_\mu} = \boldsymbol{0}$. On en déduit alors que $I_\mu(\overline{\boldsymbol{x}}_\mu) = \phi_\mu(\boldsymbol{0}) = \boldsymbol{0}$ d'une part, et d'autre part $\overline{\boldsymbol{x}}_\mu = \phi_\mu'(\boldsymbol{0})$.

Examinons maintenant comment sont modifiées les autres propriétés lorsque l'on exprime I_μ sous la forme d'une convexe conjuguée. Il est facile de voir que ces propriétés deviennent

• 4. Indépendance. $\phi_\mu(\boldsymbol{s}) = \phi_{\mu_1}(\boldsymbol{s}_1) + \phi_{\mu_2}(\boldsymbol{s}_2)$.

• 5. Additivité. La propriété d'additivité entraîne

$$\operatorname*{Sup}_{\boldsymbol{s}} \left\{ \boldsymbol{s}^{\mathrm{t}}(\boldsymbol{x}_1 + \boldsymbol{x}_2) - \phi_{\mu_1 * \mu_2}(\boldsymbol{s}) \right\} \leq \operatorname*{Sup}_{\boldsymbol{s}} \left\{ \boldsymbol{s}^{\mathrm{t}}\boldsymbol{x}_1 - \phi_{\mu_1}(\boldsymbol{s}) \right\} + \operatorname*{Sup}_{\boldsymbol{s}} \left\{ \boldsymbol{s}^{\mathrm{t}}\boldsymbol{x}_2 - \phi_{\mu_2}(\boldsymbol{s}) \right\},$$

et est en particulier vérifiée si $\phi_{\mu_1 * \mu_2}(\boldsymbol{s})$ est séparable en $\phi_{\mu_1 * \mu_2}(\boldsymbol{s}) = \phi_{\mu_1}(\boldsymbol{s}) + \phi_{\mu_2}(\boldsymbol{s})$. Cette restriction pourrait déjà nous permettre de conclure. En effet, on sait que la fonction génératrice des cumulants vérifie cette égalité et transforme ainsi un produit de convolution en une somme. Elle vérifie également la propriété 2, et il n'est pas difficile de voir que les autres propriétés sont également satisfaites. Nous allons cependant tenter d'obtenir cette solution en utilisant uniquement les autres propriétés. Dans le cas général, la propriété d'additivité impose

$$\phi_{\mu_1 * \mu_2}(\boldsymbol{s}) \geq \phi_{\mu_1}(\boldsymbol{s}) + \phi_{\mu_2}(\boldsymbol{s}).$$

• 6. Échelle $\phi_{\mu'}(\boldsymbol{s}) = \phi_\mu(a\boldsymbol{s})$.

Posons maintenant pour $\phi_\mu(\boldsymbol{s})$ la forme générale suivante :

$$\phi_\mu(\boldsymbol{s}) = f(\mathrm{E}_\mu g(\boldsymbol{s}, \boldsymbol{x}))),$$

(la forme la plus générale serait $\phi_\mu(\boldsymbol{s}) = \sum_i f_i(\mathrm{E}_\mu g_i(\boldsymbol{s}, \boldsymbol{x})))$) et recherchons quelles sont les possibilités pour f et g. La condition d'échelle fournit

$$\phi_{\mu'}(\boldsymbol{s}) = f(\mathrm{E}_{\mu'} g(\boldsymbol{s}, \boldsymbol{x}))) = f(\mathrm{E}_\mu g(\boldsymbol{s}, a\boldsymbol{x}))) = \phi_\mu(a\boldsymbol{s}).$$

On en déduit donc que $g(\boldsymbol{s}, a\boldsymbol{x}) = g(a\boldsymbol{s}, \boldsymbol{x})$, et par suite $g(\boldsymbol{s}, \boldsymbol{x}) = g(\boldsymbol{s}^{\mathrm{t}}\boldsymbol{x})$.

Examinons ensuite la condition d'indépendance :

$$\phi_\mu(\boldsymbol{s}) = \phi_{[\mu_1 \mu_2]}([\boldsymbol{s}_1 \boldsymbol{s}_2]) = f(\mathrm{E}_{[\mu_1 \mu_2]}\{g(\boldsymbol{s}_1^{\mathrm{t}}\boldsymbol{x}_1 + \boldsymbol{s}_2^{\mathrm{t}}\boldsymbol{x}_2)\})$$

doit être égale à

$$f(\mathrm{E}_{\mu_1}\{g(\boldsymbol{s}_1^{\mathrm{t}}\boldsymbol{x}_1)\}) + f(\mathrm{E}_{\mu_2}\{g(\boldsymbol{s}_2^{\mathrm{t}}\boldsymbol{x}_2)\}).$$

Comme ceci doit être vrai quels que soient $\boldsymbol{s}_1$ et $\boldsymbol{s}_2$, il faut nécessairement que g soit séparable en ses arguments, ce que l'on notera $g(\boldsymbol{s}_1^{\mathrm{t}}\boldsymbol{x}_1 + \boldsymbol{s}_2^{\mathrm{t}}\boldsymbol{x}_2) = g(\boldsymbol{s}_1^{\mathrm{t}}\boldsymbol{x}_1) \bullet g(\boldsymbol{s}_2^{\mathrm{t}}\boldsymbol{x}_2)$, où l'opération $\bullet$ est soit une addition soit une multiplication (en raison de la symétrie des arguments). Dans ces conditions, g ne peut être que la fonction identité ou la fonction exponentielle, et la fonction f associée soit l'identité soit proportionnelle au logarithme. La propriété 2 permet de rejeter l'identité (condition $\phi_\mu'(\boldsymbol{0}) = \overline{\boldsymbol{x}}_\mu$), et sélectionne

$$\boxed{\phi_{\mu,\alpha}(\boldsymbol{s}) = \alpha \log \mathrm{E}_\mu\{\exp\left(\alpha^{-1}\boldsymbol{s}^{\mathrm{t}}\boldsymbol{x}\right)\},}$$

où α est une constante arbitraire qui définit en quelque sorte « l'unité » de l'information. La propriété d'additivité impose que cette constante soit strictement positive ; par ailleurs, si α est négatif, $\phi_{\mu,\alpha}(\boldsymbol{s})$ est strictement concave, et infinie sur les bornes de son domaine. Le domaine de définition de $I_{\mu,\alpha}(\boldsymbol{x})$ est alors vide. Pour $\alpha = 1$, $\phi_\mu(\boldsymbol{s})$ est la fonction génératrice des cumulants de μ.

La mesure d'information recherchée, la mesure de l'information associée à un élément $\boldsymbol{x}$ d'un convexe $\mathcal{C}$ muni de la mesure de référence μ, est la fonctionnelle $I_\mu(\boldsymbol{x})$ définie comme la conjuguée convexe de la fonction génératrice des cumulants de μ :

$$\boxed{I_\mu(\boldsymbol{x}) = \sup_{\boldsymbol{s} \in D_{\phi_\mu}} \left\{ \boldsymbol{s}^{\mathrm{t}} \boldsymbol{x} - \phi_\mu(\boldsymbol{s}) \right\},}$$

où D_{ϕ_μ} est le domaine de définition de $\phi_\mu(\boldsymbol{s})$. Cette fonction est aussi, par définition, la transformée de Cramér de μ.

Pour α positif quelconque, il est facile de vérifier que $I_{\mu,\alpha}(\boldsymbol{x}) = \alpha I_\mu(\boldsymbol{x})$ (α doit être positif pour garantir la positivité et la convexité de $I_{\mu,\alpha}(\boldsymbol{x})$).

Notons que la minimisation sous contrainte linéaire de toute fonction croissante de $I_\mu(\boldsymbol{x})$, $h(I_\mu(\boldsymbol{x}))$ conduit à la même solution que la minimisation de $I_\mu(\boldsymbol{x})$. Cependant, on ne pourra pas alors conserver toutes les propriétés que nous avons énoncées. La convexité imposera que h soit elle-même convexe, l'annulation pour la moyenne imposera $h(\boldsymbol{0}) = 0$, les propriétés d'échelle et d'additivité seront conservées, mais la propriété d'indépendance ne pourra être respectée.

Notons encore une propriété d'invariance de cette mesure d'information. Supposons que l'on observe une moyenne empirique $\bar{\boldsymbol{x}}_n = 1/n \sum_{i=1}^n \boldsymbol{x}_i$ de n variables aléatoires indépendantes de même loi μ, et notons $\bar{\mu}_n$ la mesure de référence associée à cette moyenne. Il est facile de vérifier que l'on a alors

$$I_{\bar{\mu}_n}(\boldsymbol{x}) = n I_\mu(\boldsymbol{x}).$$

Il est intéressant de noter que (Schützenberger, 1954, page 65), cité par (Kullback 1959, page 41) définit la fonction génératrice des cumulants comme une pseudo-information (car elle ne possède pas toutes les propriétés d'une mesure d'information). Nous retrouverons cette interprétation de la fonction génératrice des cumulants comme mesure d'information à propos de l'information de Rényi.

On peut étendre le raisonnement précédent au cas où l'on observe une fonction $\boldsymbol{y} = \boldsymbol{T}(\boldsymbol{x})$ de $\boldsymbol{x}$. Soit ν la mesure de référence associée aux variables aléatoires $\boldsymbol{y}$. L'information associée est alors

$$I_\nu(\boldsymbol{y}) = \sup_{\boldsymbol{s} \in D_{\phi_\nu}} \left\{ \boldsymbol{s}^{\mathrm{t}} \boldsymbol{y} - \log \left(\mathrm{E}_\nu \left\{ \exp \left(\boldsymbol{s}^{\mathrm{t}} \boldsymbol{y} \right) \right\} \right) \right\}$$

On peut utiliser le lemme suivant :

$$\int g(y)\, \mathrm{d}\nu(y) = \int g(T(\boldsymbol{x}))\, \mathrm{d}\mu(\boldsymbol{x}).$$

Il s'en suit que $\mathrm{E}_\nu \left\{ \exp \left(\boldsymbol{s}^{\mathrm{t}} y \right) \right\} = \mathrm{E}_\mu \left\{ \exp \left(\boldsymbol{s}^{\mathrm{t}} T(\boldsymbol{x}) \right) \right\}$. Dans ces conditions,

$$\begin{aligned} I_\nu(\boldsymbol{y}) &= I_\mu(\boldsymbol{T}(\boldsymbol{x})) \\ &= \sup_{\boldsymbol{s} \in D_{\phi_\mu}} \left\{ \boldsymbol{s}^{\mathrm{t}} \boldsymbol{T}(\boldsymbol{x}) - \log \left(\mathrm{E}_\mu \left\{ \exp \left(\boldsymbol{s}^{\mathrm{t}} \boldsymbol{T}(\boldsymbol{x}) \right) \right\} \right) \right\}. \end{aligned}$$

Lien avec l'information de Kullback-Leibler

Il est possible de relier la notion d'information que nous avons introduite ci-dessus à l'information de Kullback-Leibler, et de lui donner ainsi un sens plus précis. Rappelons l'expression de l'information de Kullback : celle-ci est définie entre deux distributions de probabilité, P et Q, par

$$\begin{aligned} D(P||Q) &= \int \mathrm{d}P(\boldsymbol{x}) \log \frac{\mathrm{d}P}{\mathrm{d}Q}(\boldsymbol{x}) \quad \text{si } P \ll Q \\ &= +\infty \text{ sinon.} \end{aligned}$$

Notons que si P et Q admettent deux densités p et q par rapport à une mesure de référence λ, alors

$$D(P||Q) = \int p(\boldsymbol{x}) \log \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\, \mathrm{d}\lambda(\boldsymbol{x}).$$

L'information de Kullback admet la description variationelle suivante, voir (Gray 1990, page 104) pour une démonstration (le résultat original est dû à Donsker-Varadhan) :

$$D(P||Q) = \sup_\phi \{ \mathrm{E}_P \{\phi\} - \log \mathrm{E}_Q \{\exp \phi\} \}.$$

On a donc en particulier

$$D(P||Q) \geq \mathrm{E}_P \{\phi\} - \log \mathrm{E}_Q \{\exp \phi\} \qquad \forall \phi$$

Si l'on pose maintenant $\phi = s\psi$, l'inégalité s'entendra alors $\forall s, \forall \psi$. En fixant maintenant ψ, on obtient

$$D(P||Q) \geq s\mathrm{E}_P \{\psi\} - \log \mathrm{E}_Q \{\exp s\psi\} \qquad \forall s.$$

Par conséquent,

$$D(P||Q) \geq \sup_s \{ s\mathrm{E}_P \{\psi\} - \log \mathrm{E}_Q \{\exp s\psi\} \}.$$

Notons maintenant P^* la distribution qui réalise l'égalité, c'est-à-dire

$$D(P^*||Q) = \sup_s \{ s\mathrm{E}_{P^*} \{\psi\} - \log \mathrm{E}_Q \{\exp s\psi\} \}.$$

Dans l'ensemble des distributions de moyenne fixée (à y pour fixer les idées)

$$\mathcal{P}_y = \{ P \mid \mathrm{E}_P \{\psi\} = \mathrm{E}_{P^*} \{\psi\} = y \},$$

P^* réalise le minimum de l'information de Kullback :

$$D(P^*||Q) = \inf_{P \in \mathcal{P}_y} D(P||Q).$$

On obtient ainsi

$$\boxed{\inf_{P \in \mathcal{P}_y} D(P||Q) = \sup_s \{ sy - \log \mathrm{E}_Q \{\exp s\psi\} \} = I_Q(y).}$$

Ceci nous permet donc d'interpréter l'information I_μ comme une forme contractée de l'information de Kullback $D(P||\mu)$, agissant sur un ensemble de moyennes possibles $\mathcal{C}$. On notera dans la suite $\phi_{\mu,\psi}(s) = \log \mathrm{E}_\mu \{\exp s\psi\}$, en omettant l'indice lorsque ψ est l'identité.

Il est facile de vérifier que la distribution P^* s'écrit

$$\mathrm{d}P^*(x) = \exp \left(s\psi(x) - \phi_{\mu,\psi}(s) \right)\, \mathrm{d}\mu(x),$$

où s vérifie $y = \phi'_{\mu,\psi}(s)$. Au sein de la famille exponentielle engendrée par μ, s est le paramètre naturel, et $\phi_{\mu,\psi}(s)$ le potentiel. Notons (voir annexe) que la transformée de LEGENDRE-FENCHEL (conjuguée convexe) du potentiel

$$I_\mu(\psi) = \sup_s \{s\psi - \phi_{\mu,\psi}(s)\}$$

associe à ψ le paramètre naturel s tel que $\psi = \phi'_{\mu,\psi}(s)$ (s est ici une fonction de x). Réciproquement, la conjuguée convexe de $I_\mu(\psi)$ est $\phi_{\mu,\psi}(s)$ et associe s à ψ selon $s = I'_\mu(\psi)$. On a alors la relation de YOUNG

$$I_\mu(\psi) + \phi_{\mu,\psi}(s) \geq s\psi,$$

avec égalité lorsque s et ψ sont liées par une des relations de dérivée précédentes.

Lien avec les entropie et information de RÉNYI

Remarquons tout d'abord que si l'on s'intéresse à $\psi(x) = \log(p(x))$, alors

$$\phi_{\mu,\psi}(s) = \log \mathrm{E}_\mu \{\exp s \log p(x)\} = \log \mathrm{E}_\mu \{p(x)^s\},$$

qui est, à un facteur $1/(s-1)$, l'entropie de RÉNYI (RÉNYI 1966) d'ordre s : $H_s(P) = \frac{1}{s-1}\phi_{\mu,\psi}(s)$. On remarque ainsi que l'entropie définie par RÉNYI s'identifie à $\phi_{\mu,\psi}(s)$ et non à sa fonction duale. Considérons maintenant le logarithme du rapport de vraisemblance, en prenant $\psi(x) = \log\left(\frac{p(x)}{q(x)}\right)$, où p et q sont deux densités par rapport à μ. Le potentiel $\phi_{Q,\psi}(s)$ s'écrit alors

$$\begin{aligned}\phi_{Q,\psi}(s) &= \log \int q(x) \exp\left(s \log \tfrac{p(x)}{q(x)}\right) \mathrm{d}\mu(x) \\ &= \log \int p(x)^s q(x)^{1-s} \,\mathrm{d}\mu(x),\end{aligned}$$

qui est cette fois ci liée à l'information de RÉNYI d'ordre s, $R_s(P,Q) = \frac{1}{s-1}\phi_{Q,\psi}(s)$.

On peut donc relier l'information de RÉNYI et l'information de KULLBACK en considérant le problème de minimisation de l'information de KULLBACK entre P et Q sous la contrainte $y = \mathrm{E}_P\{\psi(x)\}$:

$$\begin{aligned}\inf_{P \in \mathcal{P}_y} D(P||Q) &= D(P^*||Q) \\ &= \sup_s \{sy - \phi_{Q,\psi}(s)\} \\ &= s^* y - \phi_{Q,\psi}(s^*), \\ &= s^* y - (s^*-1) R_{s^*}(P,Q)\end{aligned}$$

avec $y = \phi'_{Q,\psi}(s^*)$, et $p^* \propto p(x)^{s^*} q(x)^{1-s^*}$. En prenant alors $s^* = 0$, on obtient $p^* = q, y = -D(Q||P)$ et $D(P^*||Q) = 0$. Pour $s^* = 1$, on obtient $p^* = p, y = D(P||Q)$ et $D(P^*||Q) = D(P||Q)$.

L'expression précédente donne ainsi un sens au paramètre s de l'information de RÉNYI : il s'agit du paramètre naturel de la loi exponentielle de moyenne $y = \mathrm{E}_P\{\psi(x)\}$.

L'entropie et l'information de RÉNYI sont souvent utilisées pour borner des probabilités d'erreur, par exemple dans les problèmes de codage de source ou de test d'hypothèse. Ces bonnes d'erreur peuvent en général être obtenues en appliquant des résultats de grandes déviations au logarithme de la distribution empirique ou au log rapport de vraisemblance empirique. Les bornes s'expriment alors en fonction du minimum de l'information de KULLBACK sous l'information apportée par la moyenne empirique, et il est ainsi possible de faire apparaître l'entropie ou l'information de RÉNYI. Comme le note (CSISZÁR 1993), l'adéquation de ces bornes requiert un choix pertinent du paramètre s. Le lien donné ci-dessus entre les informations de KULLBACK et RÉNYI peut permettre d'éclairer ce choix.

3 DIVERGENCES

À partir des mesures d'information que nous avons définies, on peut rechercher à définir des divergences entre deux éléments du convexe $\mathcal{C}$. Nous examinons dans ce paragraphe comment on peut retrouver plusieurs méthodes de construction de divergences et leurs liens avec la forme d'information étudiée ici.

3.1 Divergences de BREGMAN

Si on considère deux fonctions convexes conjuguées l'une de l'autre $f(x)$ et $f^*(x^*)$, où x et x^* sont les variables conjuguées, on a toujours $f(x) + f^*(x^*) \geq xx^*$, avec égalité lorsque $x^* = f'(x)$. Rappelons que la mesure d'information obtenue $I_\mu(x)$ est définie comme la convexe conjuguée de $\phi_\mu(s)$. On a par conséquent

$$I_\mu(x) + \phi_\mu(s) \geq xs,$$

avec égalité lorsque $x = \phi'_\mu(s)$. On peut alors définir une « distance » entre x et s par

$$\mathcal{B}_\mu(x,s) = I_\mu(x) + \phi_\mu(s) - xs.$$

Considérons deux valeurs particulières x_1 et s_2, auxquelles sont respectivement associées les variables duales par

$$\begin{aligned} &I_\mu(x_i) + \phi_\mu(s_i) - x_i s_i \qquad i = 1,2. \\ &x_i = \phi'_\mu(s_i), \qquad s_i = I'_\mu(x_i)\end{aligned}$$

En utilisant cette relation, la distance de BREGMAN (BREGMAN 1967), appelée également divergence dirigée, peut alors s'exprimer comme

$$\begin{aligned}\mathcal{B}_\mu(x_1,x_2) &= I_\mu(x_1) - I_\mu(x_2) - I'_\mu(x_2)(x_1 - x_2), \\ &= \phi_\mu(s_1) - \phi_\mu(s_2) - \phi'_\mu(s_2)(s_1 - s_2).\end{aligned}$$

Remarquons que si l'on considère deux membres de la famille exponentielle engendrée par μ, P_1 et P_2, respectivement de moyenne et paramètre naturel (x_1, s_1) et (x_2, s_2), alors la distance de BREGMAN n'est autre que l'information de KULLBACK entre les deux distributions, $D(P_1||P_2)$. L'utilisation de la divergence dirigée correspond donc ici à modifier la mesure de référence μ, en prenant comme référence le membre de la famille exponentielle de moyenne x_2 ; l'objet par défaut sur $\mathcal{C}$ devient ainsi x_2. Notons que l'on a bien entendu $\mathcal{B}(x,x) = 0$, et $\mathcal{B}_\mu(x,x_\mu) = I_\mu(x)$.

Il est intéressant de noter que la famille des distances de BREGMAN présente une propriété de projectivité, et que la « projection » vérifie un théorème de PYTHAGORE, voir également (JONES & BYRNE 1990). Si on considère une contrainte convexe *linéaire* $x \in \mathcal{C}$, et la projection x^* définie par $x^* = \arg\min_{x \in \mathcal{C}} \mathcal{B}(x, x_2)$ alors les distances de BREGMAN vérifient [1]

$$\mathcal{B}(x_1,x_2) = \mathcal{B}(x_1,x^*) + \mathcal{B}(x^*,x_2).$$

1. La distance de KULLBACK vérifie bien entendu cette égalité triangulaire, mais il n'est pas absolument nécessaire que $\mathcal{C}$ soit convexe.

3.2 *Différence de* JENSEN

Une manière classique de construire des divergences à partir d'une entropie $H(x)$ est d'utiliser la différence de JENSEN

$$J_{\alpha,H}(x_1,x_2) = H(\alpha x_1+(1-\alpha)x_2)-\alpha H(x_1)-(1-\alpha)H(x_2).$$

Celle-ci peut être comprise comme une mesure de distance construite à partir d'une propriété de concavité. En effet, les mesures d'information $I_\mu(x)$ que nous avons définies (qui sont l'opposé d'entropies) sont convexes par construction, et

$$-J_{\alpha,H}(x_1,x_2) = \alpha I_\mu(x_1)+(1-\alpha)I_\mu(x_2)-I_\mu(\alpha x_1+(1-\alpha)x_2) \geq 0.$$

L'utilisation de la différence de JENSEN peut ainsi permettre de générer une classe de mesures de divergence entre x_1 et x_2, indexée par le paramètre α. On peut noter que les divergences de BREGMAN correspondent à la dérivée de la différence de JENSEN pour $\alpha = 0$.

3.3 *f-divergences*

Les f-divergences sont une famille de divergences introduites et étudiées par CSISZÁR, et indépendemment par ALI et SILVEY, (CSISZÁR 1967, ALI & SILVEY 1966). Ces divergences s'expriment sous la forme

$$\mathcal{D}_f(x_1,x_2) = x_2 f\left(\frac{x_1}{x_2}\right),$$

où f est une fonction convexe vérifiant $f(1) = f'(1) = 0$.

Nous avons vu qu'une manière de modifier l'objet par défaut $\bar{\boldsymbol{x}}_\mu$ consiste à prendre une distribution exponentielle par rapport à μ comme mesure de référence, la moyenne de cette distribution exponentielle devenant alors le nouvel objet par défaut. Ceci mène alors aux divergences de BREGMAN.

Il est possible de procéder différemment pour modifier l'objet par défaut. On peut en effet vérifier que si l'on change l'objet par défaut dans la propriété 2, en prenant un nouvel objet m dans $\mathcal{C}$, on aboutit à la fonction (on se place à nouveau ici dans le cas scalaire ; le raisonnement s'étend facilement au cas vectoriel pour une mesure de référence séparable)

$$\phi_\mu^{(m)}(s) = \frac{m}{\bar{x}_\mu}\log \mathrm{E}_\mu\{\exp sx\},$$

qui vérifie bien $\phi_\mu^{(m)'}(0) = m$. La convexe conjuguée de cette fonction, c'est-à-dire la nouvelle mesure d'information est alors

$$I_\mu^{(m)}(x) = \frac{m}{\bar{x}_\mu} I_\mu(\bar{x}_\mu \frac{x}{m}).$$

Il nous suffit alors maintenant de poser $f_\mu(x) = I_\mu(\bar{x}_\mu x)/\bar{x}_\mu$, pour obtenir une f-divergence :

$$\begin{aligned}\mathcal{D}_{f_\mu}(x_1,x_2) &= x_2 f_\mu\left(\frac{x_1}{x_2}\right),\\ &= \frac{x_2}{\bar{x}_\mu} I_\mu\left(\bar{x}_\mu\frac{x_1}{x_2}\right).\end{aligned}$$

On peut vérifier facilement que $f(1) = \bar{x}_\mu^{-1} I_\mu(\bar{x}_\mu = 0$, et que $f'(1) = \bar{x}_\mu^{-1} I'_\mu(\bar{x}_\mu = 0$ (à l'aide par exemple de la relation $I'_\mu = \phi_\mu^{'-1}$, et en tenant compte de $x_\mu = \phi'_\mu(0)$).

Dans le cas vectoriel, avec une mesure séparable, ceci devient

$$\mathcal{D}_{f_\mu}(\boldsymbol{a},\boldsymbol{b}) = \sum_{i=1}^N \frac{b_i}{\bar{x}_i} I_\mu\left(\bar{x}_i\frac{a_i}{b_i}\right),$$

où a_i, b_i et x_i notent respectivement les i^{e} composantes de $\boldsymbol{a}$, $\boldsymbol{b}$ et $\bar{\boldsymbol{x}}_\mu$.

Plaçons nous dans un ensemble $\mathcal{C}$ dont tous les éléments sont de moyenne unité (relativement à la mesure μ), et *a fortiori* tel que $\bar{x}_\mu = 1$ (ce sera par exemple l'ensemble des distributions de probabilité), et prenons pour μ la mesure de LEBESGUE sur $\mathbb{R}^+$. On obtient alors $I_\mu(x) = -\log x$, et

$$\mathcal{D}_{f_\mu}(x_1,x_2) = x_2\log\left(\frac{x_2}{x_1}\right),$$

c'est-à-dire l'information de KULLBACK $D(x_2||x_1)$.

En changeant ainsi l'objet par défaut, par deux procédures différentes, nous pouvons retrouver les deux grandes classes de divergences — divergences de BREGMAN et f-divergences — et leur donner ici la signification d'un changement de mesure d'information lorsque l'on fait évoluer l'objet par défaut. Notons toutefois que ces constructions ne permettent pas de retrouver *toutes* les divergences de BREGMAN et f-divergences, mais simplement deux sous-classes de ces divergences, cohérentes avec la mesure d'information que nous avons définie.[2]

4 QUELQUES EXEMPLES

Nous donnons ici quelques exemples de fonctions $I_\mu(x)$ obtenues par ce procédé, ainsi que l'expression de certaines divergences associées. Nous examinerons tout d'abord comment on peut construire pratiquement les fonctions $I_\mu(x)$ en fonction de la mesure de référence μ.

4.1 *Calcul pratique de* $I_\mu(x)$

Rappelons que les fonctions I_μ et $\phi_m u$ sont conjuguées convexes l'une de l'autre. Le calcul pratique de $I_\mu(x)$ s'appuie donc sur les propriétés des convexes conjuguées (voir l'annexe A, relation A-4). On obtient ainsi, avec $I_\mu^{'} = \phi_\mu^{'-1}$,

$$I_\mu(x) = xI_\mu^{'}(x) - \phi_\mu(I_\mu^{'}(x)).$$

La difficulté principale réside dans la détermination de $I_\mu(x)^{'}$, qui nécessite le calcul de la réciproque de $\phi_\mu^{'}(s)$. Pour beaucoup de mesures, ce calcul ne conduit pas à une expression analytique de $I_\mu^{'}(x)$, et la mesure d'information ne peut alors être définie qu'implicitement comme la convexe conjuguée de $\phi_\mu(s)$. Nous verrons cependant que pour l'application à la résolution de problèmes inverses, on peut se contenter de l'expression de $\phi_\mu(s)$.

2. **f-divergence et distributions indéfiniment divisibles.** Nous reprenons ici une remarque de CSISZÁR (journée maximum d'entropie sur la mouenne et grandes déviations, Évry, 30 mai 1995). Prenons comme mesure de référence μ une distribution indéfiniment divisible, et considérons la distribution μ_n telle que $\mathrm{E}_{\mu_n}\{\exp tx\} = \mathrm{E}_\mu\{\exp tx\}^n$. On obtient alors $I_{\mu_n}(x) = nI_\mu(x/n)$, qui présente la forme générique des f-divergences.

4.2 Exemples de critères

4.2.1 Mesure de référence gaussienne

L'utilisation d'une mesure de référence gaussienne $\mu = \mathcal{N}(\boldsymbol{m}, \boldsymbol{\Gamma})$, mène facilement au critère quadratique

$$I_\mu(\boldsymbol{x}) = (\boldsymbol{x} - \boldsymbol{m})^{\mathrm{t}} \boldsymbol{\Gamma}^{-1} (\boldsymbol{x} - \boldsymbol{m}).$$

Observons que le minimum est bien obtenu pour la moyenne de la mesure de référence $\boldsymbol{m}$.

4.2.2 Mesure de référence gamma et critère d'Itakura-Saïto

On peut s'intéresser à des variables positives, c'est-à-dire définies sur le convexe $\mathcal{C} = [0, \infty[^N$ en choisissant comme mesure de référence un produit de lois gamma, $\Gamma(\beta_i, \alpha_i)$, $i = 1..N$. La tranformée de Cramér de cette mesure fournit alors

$$I_\mu(\boldsymbol{x}) = \sum_{i=1}^{N} (\alpha_i x_i - \beta_i) + \beta_i \log\left(\frac{\beta_i}{\alpha_i x_i}\right).$$

Il est facile de vérifier que ce dernier critère prend son minimum et s'annule pour la moyenne de la mesure de référence, $\bar{x}_i = \beta_i/\alpha_i$. Pour $\beta_i = 1$, on obtient la divergence d'Itakura-Saïto, et en prenant enfin $\alpha_i = 1$, on obtient l'entropie de Burg.

4.2.3 Mesure de Poisson et information de Kullback

On a considéré dans le paragraphe précédent le cas où le convexe $\mathcal{C}$ est $\mathbb{R}_+^N$, et la mesure de référence une loi Γ. Toujours pour des variables positives, on peut également envisager d'utiliser une mesure de Poisson, de paramètre m_i, $i = 1..N$.[3]

Le calcul de la transformée de Cramér de μ conduit alors à

$$I_\mu(\boldsymbol{x}) = \sum_{i=1}^{N} \left[x_i \log\left(\frac{x_i}{m_i}\right) + m_i - x_i \right],$$

qui est l'expression de l'information de Kullback (entropie croisée), dans laquelle le terme correctif $(m_i - x_i)$ assure la positivité de I_μ dans le cas où $\boldsymbol{x}$, ou $\boldsymbol{m}$ ne serait pas normalisé à l'unité. Pour $m_i = 1/N$ et $\boldsymbol{x}$ de somme 1, on retrouve l'entropie de Shannon (au signe près).

4.2.4 Lois composées de Poisson

Une famille intéressante de mesures de référence peut être obtenue en utilisant des sommes poissonnisées, voir par exemple (Gamboa & Lavielle 1994). On considère donc la somme de N variables aléatoires, où N suit lui-même une distribution de Poisson.[4] En notant ψ la fonction génératrice des variables aléatoires, et f le paramètre de Poisson, la fonction génératrice de la somme est $\exp(-f + f\psi(s))$. La fonction génératrice des cumulants est alors simplement $-f + f\psi(s)$, et pour un vecteur complet $\boldsymbol{x}$, on obtient alors

$$\phi(\boldsymbol{s}) = \sum_{i=1}^{N} -f_i + f_i \psi_i(s_i).$$

On peut par exemple appliquer cette construction pour une distribution gamma, $\Gamma(\beta_i, \alpha_i)$ $i = 1..N$ pour $\boldsymbol{x}$ (le critère correspondant a été exhibé en (Gamboa & Lavielle 1994, Bercher 1995)) :

$$I_\mu(\boldsymbol{x}) = \sum_{i=1}^{N} f_i \left\{ \beta_i \frac{x_i}{m_i} + 1 - (\beta_i + 1) \left(\frac{x_i}{m_i}\right)^{\frac{\beta_i}{\beta_i+1}} \right\},$$

dans lequel on a fait apparaître la moyenne $m_i = \beta_i/\alpha_i$.

Cette expression peut être reliée au critère entropique introduit par (Jones & Byrne 1990).

En posant $\gamma = \frac{\beta}{\beta+1}$ (on se place ici dans la cas scalaire pour alléger les notations), le critère précédent peut s'écrire, à un facteur près

$$I_\mu(x) = -\left(\frac{x}{m}\right)^\gamma + \gamma \frac{x}{m} + 1 - \gamma.$$

Il est particulièrement intéressant d'observer que pour $\gamma = 1/2$, on fait apparaître un critère en racine carrée, pour lequel a longtemps été cherchée une justification, voir par exemple (Narayan & Nityananda 1986).

On peut maintenant chercher à construire une f-divergence à l'aide de l'expression de $I_\mu(x)$. Posons m=1. La divergence s'écrit alors

$$D_\gamma(x_1, x_2) = x_2 I_\mu(\frac{x_1}{x_2}) = -{x_1}^\gamma {x_2}^{1-\gamma} + \gamma x_1 + (1-\gamma) x_2.$$

Il est remarquable que l'on obtienne alors, toujours pour $\gamma = 1/2$, la distance d'Hellinger : $\left(\sqrt{(x_1)} - \sqrt{(x_2)}\right)^2$.

La divergence de Bregman peut également être calculée :

$$\mathcal{B}(x_1, x_2) = (x_2^\gamma - x_1^\gamma) - \gamma x_2^{\gamma-1}(x_2 - x_1).$$

4.2.5 Mesure de référence de Bernoulli

Si l'on s'intéresse au convexe $\mathcal{C}$ défini par $\mathcal{C} = \{\boldsymbol{x} | x_i \in [a_i, b_i] \quad i = 1..N\}$, on peut utiliser une mesure uniforme sur chacun des intervalles. Cependant une mesure uniforme ne conduit pas à une expression explicite de $I_\mu(\boldsymbol{x})$. On peut également utiliser une mesure de Bernoulli sur chacun des intervalles $[a_i, b_i]$; le convexe est alors construit comme l'enveloppe convexe du support de μ, avec $\mathrm{d}\mu_i(x_i) = \alpha_i \delta(x_i - a_i) + (1 - \alpha_i)\delta(b_i - x_i)$. Notons que le même type de résultat peut être obtenu en utilisant une loi binomiale. La mesure d'information prend la forme

$$I_\mu(\boldsymbol{x}) = \sum_{i=1}^{N} \left(\frac{x_i - a_i}{b_i - a_i}\right) \log\left(\frac{x_i - a_i}{1 - \alpha_i}\right) + \frac{b_i - x_i}{b_i - a_i} \log\left(\frac{b_i - x_i}{\alpha_i}\right) - \log(b_i - a_i).$$

Il est facile de vérifier ici encore que cette expression s'annule pour la solution par défaut $x_i = \alpha_i a_i + (1 - \alpha_i) b_i$, $i = 1..N$. Pour $a_i = 0$, $b_i = 1$, et $\alpha_i = 1/2$, on obtient l'entropie de Fermi-Dirac :

$$I_\mu(\boldsymbol{x}) = \sum_{i=1}^{N} (x_i \log(x_i) + (1 - x_i) \log(1 - x_i)) + N \log(2).$$

3. Ce modèle peut être un modèle de formation d'image, qui représente l'accumulation de grains d'énergie (par exemple des photons), selon un processus de Poisson, et de telle sorte que la moyenne au site i soit m_i. Un tel modèle peut par exemple être rencontré en astronomie.

4. Ce modèle peut être considéré comme une extension du modèle de formation d'image développé dans le paragraphe précédent. Chaque composante de $\boldsymbol{Z}$ représente toujours un site ou un pixel de l'image, et plutôt que d'accumuler sur les différents sites (selon une loi de Poisson) des « grains d'intensité » de valeur constante, on considère que l'intensité de chacun des grains est variable et gouvernée, au site i, par une distribution μ_i.

4.2.6 Loi de PASCAL

Nous terminerons ces quelques exemples avec la loi de PASCAL (binomiale négative). Celle-ci décrit le nombre d'épreuves de Bernoulli pour obtenir 1 r fois. Notons un lien avec la loi de POISSON : si on considère le paramètre de la loi de POISSON aléatoire, et de loi $\Gamma(p/(1-p), r)$, alors on retrouve la loi de PASCAL. La mesure d'information associée est ici

$$I_\mu(\boldsymbol{x}) = \sum_{i=1}^{N} (x_i \log(x_i) - (x_i + r_i)\log(x_i + r_i) \\ -x_i \log(1-p_i) - r_i \log(p_i/r_i)) .$$

Pour la loi géométrique, ($r_i = 1$), on obtient $I_{\mu_i}(x_i) = x_i \log(x_i) - (x_i+1)\log(x_i+1) - x_i \log(1-p_i) - \log(p_i)$, dans laquelle le terme $x_i \log(x_i) - (x_i+1)\log(x_i+1)$ est l'entropie de BOSE-EINSTEIN.

Remarquons qu'il est possible d'obtenir l'entropie de BOSE-EINSTEIN en utilisant la construction développée ici en prenant pour μ la mesure discrète $\mu(k) = 1,\ k = 0..\infty$ (mesure de comptage). Notons encore qu'il ne s'agit plus ici d'une mesure de probabilité.

5 APPLICATION À LA RÉSOLUTION DE PROBLÈMES INVERSES

On considère le problème inverse, qui consiste à déterminer, restaurer, un objet $\boldsymbol{x}$ lorsqu'on dispose de l'équation d'observation suivante :

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b},$$

où les composantes de $\boldsymbol{y}$ sont des observations indirectes et bruitées d'un objet $\boldsymbol{x}$, à l'aide d'un procédé expérimental caractérisé par la matrice de transfert $\boldsymbol{A}$. Cette matrice peut être mal-conditionnée ou incomplète. Dans ces conditions, l'inversion directe du problème est impossible, ou conduit à une solution de peu d'intérêt. Les données doivent être complétées par une information supplémentaire qui permette de sélectionner une solution acceptable.

On peut souvent disposer d'une connaissance sur la distribution du bruit d'observation. D'autre part, on peut également disposer de conaissances sur un procédé de formation de l'objet $\boldsymbol{x}$ (modèle de formation d'image par exemple), ou se donner un modèle (probabiliste) de l'objet. Enfin, on peut chercher à respecter certaines contraintes convexes, comme une contrainte de positivité ou l'appartenance à un gabarit, en se donnant une mesure de référence sur cet ensemble convexe. Sans aller plus avant dans cette discussion, nous supposerons ici disposer d'une mesure de référence μ sur l'objet, et d'une mesure ν pour le bruit. On pourra consulter l'article de C. HEINRICH *et al.* (HEINRICH *et al.* 1995) dans ces actes pour une modélisation de signaux impulsionnels et l'application de cette technique en échographie appliquée au contrôle non destructif, et en spectroscopie.

À l'aide de ces deux mesures de références, on peut déterminer les informations associées $I_\mu(\boldsymbol{x})$ et $I_\nu(\boldsymbol{b})$, et rechercher alors la solution $\boldsymbol{x}$ la moins informative, c'est-à-dire la plus proche de la solution par défaut $\bar{\boldsymbol{x}}_\mu$, tout en respectant une contrainte « informationnelle » liée à l'observation. On recherchera par exemple à préserver $I_\nu(\boldsymbol{b} = \boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}) \leq \rho$. Ce problème d'optimisation peut donc se formuler comme

$$\begin{cases} \inf\limits_{\boldsymbol{x} \in \mathcal{C} \cap \mathcal{D}} I_\mu(\boldsymbol{x}) \\ I_\nu(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}) \leq \rho, \end{cases}$$

où $\mathcal{C}$ et $\mathcal{D}$ sont les domaines de définition des deux fonctions.

On peut montrer que ce problème est équivalent à rechercher l'argument du minimum de

$$\inf_{\boldsymbol{x} \in \mathcal{C} \cap \mathcal{D}} I_\mu(\boldsymbol{x}) + \alpha I_\nu(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}),$$

où α est une constante positive (c'est le paramètre de LAGRANGE associé à la contrainte $I_\nu(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}) \leq \rho$). Ce critère prend la forme classique des critères régularisés, et α (son inverse) joue alors le rôle d'un paramètre de régularisation.

On peut obtenir une formulation duale de ce critère en utilisant le théorème de dualité de FENCHEL avec $\mathcal{F}(\boldsymbol{x}) = I_\mu(\boldsymbol{x})$ et $\mathcal{G}(\boldsymbol{x}) = \alpha I_\nu(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x})$. On peut vérifier que $\mathcal{G}^*(\boldsymbol{x}^*) = \boldsymbol{s}^{\mathrm{t}}\boldsymbol{y} - \alpha I_\nu^*(\alpha^{-1}\boldsymbol{s}^{\mathrm{t}})$, où $\boldsymbol{x}^*$ est de la forme $\boldsymbol{x}^* = \boldsymbol{s}^{\mathrm{t}}\boldsymbol{A}$. Le problème équivalent est alors

$$\sup_{x^* = \boldsymbol{s}^{\mathrm{t}}\boldsymbol{A} \in \mathcal{C}^* \cap \mathcal{D}^*} \left\{ \boldsymbol{s}^{\mathrm{t}}\boldsymbol{y} - \alpha I_\nu^*(\alpha^{-1}\boldsymbol{s}^{\mathrm{t}}) - I_\mu^*(\boldsymbol{s}^{\mathrm{t}}\boldsymbol{A}) \right\}.$$

En se souvenant que I_μ et $\phi_\mu(\boldsymbol{\lambda}) = \log \mathrm{E}_\mu \left\{ \exp\left(\boldsymbol{\lambda}^{\mathrm{t}}\boldsymbol{x}\right) \right\}$ forment une paire de conjuguées convexes, on peut réécrire le problème dual, qui est souvent non contraint, comme

$$\sup_{x^* = \boldsymbol{s}^{\mathrm{t}}\boldsymbol{A} \in \mathcal{C}^* \cap \mathcal{D}^*} \left\{ \boldsymbol{s}^{\mathrm{t}}\boldsymbol{y} - \alpha \phi_\nu(\alpha^{-1}\boldsymbol{s}^{\mathrm{t}}) - \phi_\mu(\boldsymbol{s}^{\mathrm{t}}\boldsymbol{A}) \right\},$$

et, pour $\boldsymbol{s}^*$ solution du problème dual précédent, la solution du problème primal est donnée par

$$\boldsymbol{x} = \phi_\mu^{'}(\boldsymbol{s}^{\mathrm{t}}\boldsymbol{A})\Big|_{\boldsymbol{s}=\boldsymbol{s}^*} .$$

Notons que $\boldsymbol{x}$, comme moyenne d'une distribution exponentielle par rapport à μ, appartient automatiquement à l'enveloppe convexe $\mathcal{C}$ du support de μ.

Des propriétés analogues peuvent être obtenues pour les classes de divergences que nous avons définies.

ANNEXE A : DUALITÉ DE FENCHEL

Nous regroupons dans ce paragraphe plusieurs résultats et outils importants, issus de l'analyse convexe et de la théorie de la dualité de FENCHEL. Des références importantes sur le sujet sont les ouvrages (LUENBERGER 1969) et (ROCKAFELLAR 1970) ; on peut également consulter la synthèse récente (ROCKAFELLAR 1993).

1.1 Fonctions convexes conjuguées

Nous débuterons tout d'abord par la notion d'ensemble conjugué et de fonctionnelles convexes conjuguées. Considérons $\mathcal{F}$ une fonctionnelle convexe définie sur un sous-ensemble $\mathcal{C}$ d'un espace normé $\mathcal{X}$. On définit la fonctionnelle

convexe conjuguée par

$$\mathcal{F}^*(x^*) = \sup_{x\in\mathcal{C}} \{< x, x^* > -\mathcal{F}(x)\}, \qquad \text{(A-1)}$$

et l'ensemble conjugué $\mathcal{C}^*$ comme l'ensemble des x^* tels que $\mathcal{F}^*(x^*)$ soit finie.

En tant que supremum de fonctions continues, $\mathcal{F}^*$ est elle-même convexe. Plaçons nous dans le cas scalaire, et considérons à nouveau la définition de la conjuguée convexe, avec $t = x$, et $s = x^* : \mathcal{F}^*(s) = \sup_t \{st - \mathcal{F}(t)\}$. Lorsque $\mathcal{F}(t)$ est différentiable sur son domaine de définition, et en notant t_s la valeur de t rendant maximum $\{st - \mathcal{F}(t)\}$, on a

$$\begin{cases} \mathcal{F}^*(s) = st_s - \mathcal{F}(t_s), \\ \text{avec } s - \mathcal{F}'(t_s) = 0. \end{cases} \qquad \text{(A-2)}$$

Examinons maintenant l'expression de la dérivée de $\mathcal{F}^*(s)$ par rapport à s : $\mathcal{F}^{*'}(s) = t_s + t'_s[s - \mathcal{F}'(t_s)] = t_s$; on obtient ainsi la relation

$$\mathcal{F}^{*'}(s) = \mathcal{F}^{'-1}(s), \qquad \text{(A-3)}$$

c'est-à-dire que les dérivées de $\mathcal{F}$ et de sa conjuguée $\mathcal{F}^*$ sont réciproques l'une de l'autre. Ce résultat peut se généraliser pour des fonctions convexes fermées (semi-continues inférieurement) définies sur $\mathbb{R}^M$, à l'aide de la notion de sous-différentielle, voir (Ellis 1985, chapitre 6).

Afin de déterminer l'expression de $\mathcal{F}^*$, il faut donc dériver $\mathcal{F}$, puis calculer sa réciproque $\mathcal{F}^{'-1} = \mathcal{F}^{*'}$. Il faut enfin intégrer l'expression obtenue. Il est possible d'éviter cette intégration en utilisant la relation $\mathcal{F}^*(s) = st_s - \mathcal{F}(t_s)$, ainsi que $\mathcal{F}^{*'}(s) = t_s$. On obtient alors

$$\mathcal{F}^*(s) = s\mathcal{F}^{*'}(s) - \mathcal{F}(\mathcal{F}^{*'}(s)). \qquad \text{(A-4)}$$

En dérivant à nouveau les relations $\mathcal{F}^{*'}(s) = t_s$ et $s = \mathcal{F}'(t_s)$ par rapport à s, on obtient respectivement $\mathcal{F}^{*''}(s) = t'_s$ et $t'_s\mathcal{F}''(t_s) = 1$. On en déduit que

$$\mathcal{F}^{*''}(s)\mathcal{F}''(t_s) = 1. \qquad \text{(A-5)}$$

1.2 Fonctions concaves conjuguées

Soit $\mathcal{G}$ une fonction concave définie sur un ensemble $\mathcal{D}$. La fonction concave conjuguée est définie par

$$\mathcal{G}^*(x^*) = \inf_{x\in\mathcal{D}} < x, x^* > -\mathcal{G}(x),$$

sur l'ensemble conjugué $\mathcal{D}^*$. Notons que la fonction concave conjuguée ainsi définie *n'est pas* l'opposé de la convexe conjuguée de $-\mathcal{G}$. C'est avec cette définition que le théorème de dualité de Fenchel est valide.

1.2.1 Propriétés, inégalités de Young

Nous donnons ici les propriétés essentielles des fonctions convexes conjuguées. Nous supposerons à nouveau que $\mathcal{F}$ et $\mathcal{F}^*$ sont différentiables sur l'intérieur de leur domaine. On notera $\partial\mathcal{F}(\boldsymbol{t})$ la sous-différentielle de $\mathcal{F}$ en $\boldsymbol{t}$ (l'ensemble des sous-gradients de $\mathcal{F}$ en $\boldsymbol{t}$). Pour des fonctions convexes fermées sur $\mathbb{R}^M$, on a les propriétés suivantes (Ellis 1985, théorème VI.5.3, page 221)

(a) $\mathcal{F}^*$ est une fonction convexe fermée sur $\mathbb{R}^M$,

(b) $< \boldsymbol{s}, \boldsymbol{t} > \leq \mathcal{F}(\boldsymbol{t}) + \mathcal{F}^*(\boldsymbol{s})$ pour tout $\boldsymbol{s}, \boldsymbol{t} \in \mathbb{R}^M$,

(c) $< \boldsymbol{s}, \boldsymbol{t} > = \mathcal{F}(\boldsymbol{t}) + \mathcal{F}^*(\boldsymbol{s})$ ssi $\boldsymbol{s} \in \partial\mathcal{F}(\boldsymbol{t})$ (condition analogue à $s - \mathcal{F}'(t_s) = 0$),

(d) $\boldsymbol{s} \in \partial\mathcal{F}(\boldsymbol{t})$ ssi $\boldsymbol{t} \in \partial\mathcal{F}^*(\boldsymbol{s})$ (condition analogue à la réciprocité des dérivées $\mathcal{F}^{*'}(s) = \mathcal{F}^{'-1}(s)$)

(e) $\mathcal{F}^{**} = \mathcal{F}$, ce qui signifie que $\mathcal{F}$ est aussi la convexe conjuguée de $\mathcal{F}^*$.

Parmi ces propriétés, la propriété (e), indique que la conjugaison convexe est « réversible » si la fonction initiale $\mathcal{F}$ est convexe fermée.

1.2.2 Théorème de dualité de Fenchel

Nous pouvons maintenant énoncer le théorème de dualité de Fenchel. Si $\mathcal{F}$ et $\mathcal{G}$ sont deux fonctionnelles, respectivement convexes et concaves définies sur $\mathcal{C}$ et $\mathcal{D}$, alors

$$\mu = \inf_{x\in\mathcal{C}\cap\mathcal{D}} \{\mathcal{F}(x) - \mathcal{G}(x)\} = \sup_{x^*\in\mathcal{C}^*\cap\mathcal{D}^*} \{\mathcal{G}^*(x^*) - \mathcal{F}^*(x^*)\}. \qquad \text{(A-6)}$$

sous les conditions suivantes : $\mathcal{C}\cap\mathcal{D}$ contient des points dans l'intérieur relatif de $\mathcal{C}$ et $\mathcal{D}$, l'un au moins des ensembles $[\mathcal{F}, \mathcal{C}]$ et $[\mathcal{G}, \mathcal{D}]$ est d'intérieur non vide, et μ est fini.

Les résultats de la théorie de la dualité de Fenchel possèdent des extensions dans le cas d'une formulation intégrale — voir par exemple (Borwein & Lewis 1991) et (Rockafellar 1970). On peut en particulier en déduire que $\int_{\mathcal{C}} p(\boldsymbol{x}) \log(p(\boldsymbol{x}))\, \mathrm{d}\mu(\boldsymbol{x})$ et $\log \int \exp(\boldsymbol{s}^{\mathrm{t}}\boldsymbol{x})\, \mathrm{d}\mu(\boldsymbol{x})$ forment une paire de convexes conjuguées.

1.2.3 Relation entre les solutions des problèmes primal et dual

Lorsqu'un problème d'optimisation est résolu en utilisant sa formulation duale, il est nécessaire d'exprimer la solution du problème primal en fonction de celle du problème dual. Sous des conditions peu restrictives, les solutions optimales sont liées par une simple opération de dérivation (voir à nouveau (Borwein & Lewis 1991) ou (Decarreau *et al.* 1992)) : si $\boldsymbol{x}_0$ et $\boldsymbol{x}_0^*$ sont respectivement les solutions des problèmes primal et dual, alors

$$\begin{cases} \boldsymbol{x}_0 = \mathcal{F}^{*'}(\boldsymbol{x}_0^*) = \mathcal{G}^{*'}(\boldsymbol{x}_0^*), \\ \boldsymbol{x}_0^* = \mathcal{F}'(\boldsymbol{x}_0) = \mathcal{G}'(\boldsymbol{x}_0). \end{cases}$$

Annexe B : propriétés de la transformée de Cramér

La conjuguée convexe de cette fonction génératrice des cumulants est aussi appelée transformée de Cramér de μ (Azencott 1978, Dacunha-Castelle & Duflo 1982). Cette transformée possède de nombreuses propriétés, qui sont particulièrement intéressantes, lorsque la transformée de Cramér $I_\mu(\boldsymbol{x})$ est utilisée comme critère pour la résolution d'un problème inverse. On note $\mathcal{C}$ l'enveloppe convexe fermée du support de μ. Les propriétés essentielles sont les suivantes :

- I_μ est continûment différentiable et strictement convexe sur $\mathcal{C}$,
- $I_\mu(\boldsymbol{x}) = +\infty$ pour $\boldsymbol{x} \notin \mathcal{C}$, et sa dérivée croît à l'infini sur la frontière de $\mathcal{C}$,
- $I_\mu(\boldsymbol{x}) \geq 0$, avec égalité pour $\boldsymbol{x} = \bar{\boldsymbol{x}}_\mu$.

La stricte convexité permet une implémentation simple, et surtout garantit l'unicité de la solution. Le fait que $I_\mu(\boldsymbol{x})$ soit infinie en dehors de $\mathcal{C}$ et que sa dérivée soit elle-même infinie à la frontière de $\mathcal{C}$ interdit à une méthode d'optimisation locale de s'évader de l'ensemble $\mathcal{C}$, et fournit de ce fait une solution dans cet ensemble, sans qu'il soit nécessaire de spécifier explicitement cette contrainte au cours de l'optimisation.

Références

S. M. Ali & D. Silvey (1966), « A general class of coefficients of divergence of one distribution to another », *Journal of Royal Statistical Society B*, **28**, n° 1, pages 131–142.

R. Azencott (1978), « Grandes déviations et applications », dans P. L. Hennequin, éditeur, *École D'été de Probabilité de Saint Flour VIII-1978*, pages 2–172, Springer-Verlag, 1978.

M. Basseville (1989), « Distance measures for signal processing and pattern recognition », *Signal Processing*, **18**, pages 349–369.

J.-F. Bercher (1995), *Développement de critères de nature entropique pour la résolution de problèmes inverses linéaires.*, Thèse de doctorat, Université de Paris-Sud.

J. Borwein & A. Lewis (1991), « Duality Relationships for Entropy-like Minimization Problems », *SIAM Journal of Control and Optimization*, **29**, n° 2, pages 325–338.

L. M. Bregman (1967), « The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming », *U.S.S.R. Comput. Math. and Math. Phys.*, **7**, pages 200–217.

I. Csiszár (1967), « Information-type measures of difference of probability distributions and indirect observations », *Studia Scientiarum Mathematicarum Hungarica*, **2**, pages 299–318.

I. Csiszár (1991), « Why least-squares and maximum entropy? An axiomatic approach to inference for linear inverse problems », *The Annals of Statistics*, **19**, n° 4, pages 2032–2066.

I. Csiszár (1993), « Generalized Cutoff Rates ans Rényi Information Measures », dans IEEE *International Symposium on Information Theory*, San Antonio.

D. Dacunha-Castelle & M. Duflo (1982), *Probabilités et statistiques, 1. Problème à temps fixe*, Masson, Paris.

A. Decarreau, D. Hilhorst, C. Lemaréchal & J. Navaza (1992), « Dual Methods in Entropy Maximization. Application to some Problems in Crystallography », *SIAM Journal of Optimization*, **2**, n° 2, pages 173–197.

R. S. Ellis (1985), *Entropy, Large Deviations, and Statistical Mechanics*, Springer-Verlag, New York.

F. Gamboa & M. Lavielle (1994), « On Two–Dimensional Spectral Realization », *IEEE Transactions on Information Theory*, **40**, n° 5, pages 1603–1608.

R. M. Gray (1990), *Entropy and information theory*, Springer-Verlag, New York.

C. Heinrich, J.-F. Bercher, G. Le Besnerais & G. Demoment (1995), « Maximum d'entropie sur la moyenne et mélanges de distributions », dans *Actes Du 15^e colloque* Gretsi, septembre 1995.

L. K. Jones & C. L. Byrne (1990), « General Entropy Criteria for Inverse Problems, with Applications to Data Compression, Pattern Classification and Cluster Analysis », *IEEE Transactions on Information Theory*, **36**, n° 1, pages 23–30.

S. Kullback (1959), *Information Theory and Statistics*, Wiley, New York.

D. Luenberger (1969), *Optimization by Vector Space Methods*, Wiley, J., New York, 1 edition.

R. Narayan & R. Nityananda (1986), « Maximum entropy image restoration in astronomy », *Ann. Rev. Astron. Astrophys.*, **24**, pages 127–170.

A. Rényi (1966), *Calcul des probabilités*, Dunod, Paris.

R. T. Rockafellar (1970), *Convex Analysis*, Princeton University Press.

R. T. Rockafellar (1993), « Lagrange Multipliers and Optimality », *SIAM Review*, **35**, n° 2, pages 183-238.

C. Shannon (1948), « The Mathematical Theory of Communication », *Bell System Technical Journal*, **27**, pages 379–423, 623–656.

J. Shore & R. Johnson (1980), « Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy », *IEEE Transactions on Information Theory*, **IT-26**, pages 26–37.

Plusieurs articles des auteurs en rapport avec ce travail, peuvent être obtenus sur le ftp anonyme ftp.supelec.fr, dans le répertoire /lss/Papers/Bercher.