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

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 identifierons alors celle-ci à la transformée de CRAMÉR de la mesure de référence μ.

Nous montrerons que cette mesure d'information est liée à l'information de KULLBACK-LEIBLER et aux familles exponentielles. Nous chercherons ensuite à définir des divergences. Nous pourrons en particulier faire apparaître des divergences de BREGMAN et des f-divergences. Nous examinerons enfin l'application de ces résultats à la résolution de problèmes inverses.

ABSTRACT

In this communication, we derive a measure of the information associated to the elements of a convex set $\mathcal{C}$, endowed with a reference measure μ. We state some simple properties that should be verified by such an information measure. These properties lead us to identify this information measure with the CRAMÉR transform of the reference measure μ.

We then give the links between this information measure and the KULLBACK-LEIBLER information, and with exponential families. Then, we try to define divergences measures. In particular, we find out some BREGMAN divergences and some f-divergences. Eventually, we examine the application of these results to the resolution of inverse problems.

1 INTRODUCTION

L'UTILISATION de mesures d'information ou de gain d'information en traitement du signal et de l'image est multiple, que ce soit en communication, codage et compression, détection ou 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, de SHORE & JOHNSON et des travaux plus récents de (CSISZÁR 1991).

Nous monterons ensuite comment ces mesures d'information sont liées à l'information de KULLBACK-LEIBLER. Disposant de mesures d'information, ou entropies, nous chercherons à définir des gains d'information, ou divergences. En particulier, nous retrouverons certaines divergences de BREGMAN et f-divergences.

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 qu'elle devrait vérifier.

• 1. <u>Positivité</u>. L'information est une quantité positive : $I_\mu(\boldsymbol{x}) \geq 0$.
• 2. <u>Minimum</u>. L'information est nulle, et minimale, pour la moyenne sous μ. Connaissant $\mathcal{C}$ et μ, $\overline{\boldsymbol{x}}_\mu$ 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. <u>Convexité</u>. 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 que CSISZÁR (CSISZÁR 1991, Théorème 1, page 2044) obtient cette propriété de convexité comme conséquence de propriétés de localité et de régularité.

• 4. <u>Indépendance</u>. Si la mesure est séparable en deux blocs, *i.e.* $\overline{\mu(\boldsymbol{x}) = \mu_1(\boldsymbol{x}_1)}\mu_2(\boldsymbol{x}_2)$, 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. <u>Additivité</u>. 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. <u>Échelle</u> (homogénéité). Le changement d'échelle $\boldsymbol{x} \to a\boldsymbol{x}$, $\mu \to \mu'$ ne modifie pas l'information : $I_{\mu'}(a\boldsymbol{x}) = I_\mu(\boldsymbol{x})$.

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 entraîne alors $\boldsymbol{s}_{\overline{\boldsymbol{x}}_\mu} = \boldsymbol{0}$. 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 impose

$$\phi_{\mu_1 * \mu_2}(\boldsymbol{s}) \geq \phi_{\mu_1}(\boldsymbol{s}_1) + \phi_{\mu_2}(\boldsymbol{s}_2),$$

où l'égalité est en particulier atteinte pour la fonction génératrice des cumulants.

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

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

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

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 que $g(\boldsymbol{s}, \boldsymbol{x}) = g(\boldsymbol{s}^{\mathrm{t}}\boldsymbol{x})$.

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). La fonction g ne peut alors être que la fonction identité ou la fonction exponentielle, et la fonction f associée l'identité ou 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 positive. Pour $\alpha = 1$, $\phi_\mu(\boldsymbol{s})$ est la fonction génératrice des cumulants de μ.

La mesure d'information recherchée est ainsi la fonctionnelle $I_\mu(\boldsymbol{x})$ définie comme la conjuguée convexe de la fonction génératrice des cumulants de μ :

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

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 $I_{\mu,\alpha}(\boldsymbol{x}) = \alpha I_\mu(\boldsymbol{x})$.

2.2 *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. Celle-ci est définie par

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

L'information de KULLBACK admet la description variationnelle suivante, voir (GRAY 1990, page 104) :

$$D(P||Q) = \operatorname*{Sup}_{\phi} \left\{ \mathrm{E}_P\left\{\phi\right\} - \log \mathrm{E}_Q\left\{\exp \phi\right\} \right\}.$$

En posant $\phi = s\psi$, et en fixant ψ, on a alors

$$D(P||Q) \geq \operatorname*{Sup}_{s} \left\{ s\mathrm{E}_P\left\{\psi\right\} - \log \mathrm{E}_Q\left\{\exp s\psi\right\} \right\}.$$

Notons maintenant P^* la distribution qui réalise l'égalité dans l'expression précédente.

Dans l'ensemble des distributions de moyenne fixée y, $\mathcal{P}_y = \{P \mid \mathrm{E}_P\left\{\psi\right\} = \mathrm{E}_{P^*}\left\{\psi\right\} = y\}$, P^* réalise le minimum de l'information de KULLBACK :

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

On obtient ainsi

$$\boxed{\operatorname*{Inf}_{P \in \mathcal{P}_y} D(P||Q) = \operatorname*{Sup}_{s} \left\{ sy - \log \mathrm{E}_Q\left\{\exp s\psi\right\} \right\} = I_Q(y).}$$

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 $\mathcal{C}$. Il est facile de vérifier que la distribution P^* appartient à la famille exponentielle naturelle engendrée par μ :

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

où s vérifie $y = \phi'_{\mu,\psi}(s)$.

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 les liens entre la forme d'information étudiée ici et deux grandes classes de divergences. Sauf indication contraire, nous nous placerons dans ce paragraphe dans le cas scalaire.

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 $x^*=f'(x)$. La mesure d'information $I_\mu(x)$ étant construite comme la convexe conjuguée de $\phi_\mu(s)$, on a

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

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

$$I_\mu(x_i)+\phi_\mu(s_i)-x_is_i=0 \qquad i=1,2,$$

avec $x_i=\phi'_\mu(s_i)$ et $s_i=I'_\mu(x_i)$. En utilisant cette relation, la distance de Bregman, appelée également divergence dirigée, peut alors s'exprimer comme

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

Si on considère deux membres P_1 et P_2 de la famille exponentielle engendrée par μ, respectivement de moyennes et paramètres naturels (x_1,s_1) et (x_2,s_2), alors la distance de Bregman n'est autre que l'information de Kullback $D(P_1||P_2)$ entre les deux distributions. 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 $\mathcal{B}_\mu(x,x_\mu)=I_\mu(x)$.

3.2 *f-divergences*

Les f-divergences sont une famille de divergences introduites et étudiées par Csiszár, et indépendemment par Ali et Silvey. Si f est une fonction convexe vérifiant $f(1)=f'(1)=0$, ces divergences s'expriment sous la forme

$$\mathcal{D}_f(x_1,x_2)=x_2f\left(\frac{x_1}{x_2}\right).$$

Nous avons vu dans le paragraphe précédent qu'une manière de modifier l'objet par défaut $\bar{\boldsymbol{x}}_\mu$ consiste à prendre une distribution exponentielle par rapport à μ comme nouvelle mesure de référence. 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

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

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 :

$$\mathcal{D}_{f_\mu}(x_1,x_2)=x_2f_\mu\left(\frac{x_1}{x_2}\right)=\frac{x_2}{\bar{x}_\mu}I_\mu\left(\bar{x}_\mu\frac{x_1}{x_2}\right).$$

On peut vérifier facilement que $f(1)=\bar{x}_\mu^{-1}I_\mu(\bar{x}_\mu)=0$, et que $f'(1)=\bar{x}_\mu^{-1}I'_\mu(\bar{x}_\mu)=0$.

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.

4 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. Nous supposerons ici disposer d'une mesure de référence μ sur l'objet, et d'une mesure ν pour le bruit. On pourra consulter (Heinrich *et al.* 1995), ainsi que (Bercher 1995) pour des exemples de modélisation et d'application aux problèmes inverses.

À l'aide de ces deux mesures de référence, 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

$$\left\{\begin{array}{l}\displaystyle\operatorname*{Inf}_{\boldsymbol{x}\in\mathcal{C}\cap\mathcal{D}} I_\mu(\boldsymbol{x})\\ I_\nu(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x})\leq\rho,\end{array}\right.$$

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

$$\operatorname*{Inf}_{\boldsymbol{x}\in\mathcal{C}\cap\mathcal{D}}\left\{I_\mu(\boldsymbol{x})\right\}+\alpha I_\nu(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}),$$

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.

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 $\boldsymbol{x}^* = \boldsymbol{s}^{\mathrm{t}}\boldsymbol{A}$. En se souvenant que I et ϕ forment une paire de conjuguées convexes, le problème équivalent, qui est souvent non contraint, s'écrit

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

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

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

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

ANNEXE : 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, voir par exemple (LUENBERGER 1969).

1.1 Fonctions conjuguées

On considère une fonctionnelle convexe $\mathcal{F}$ définie sur un sous-ensemble $\mathcal{C}$ d'un espace normé $\mathcal{X}$. La fonctionnelle *convexe conjuguée* est définie par

$$\mathcal{F}^*(x^*) = \operatorname*{Sup}_{x\in\mathcal{C}}\{< x, x^* > - \mathcal{F}(x)\}, \tag{A-1}$$

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 $s = x^*$: $\mathcal{F}^*(s) = \mathrm{Sup}_t\{st - \mathcal{F}(t)\}$. Lorsque $\mathcal{F}(t)$ est différentiable sur son domaine de définition, et en notant t_s la valeur de t rendant maximum $\{st - \mathcal{F}(t)\}$, $s = \mathcal{F}'(t_s)$, on a

$$\mathcal{F}^*(s) = st_s - \mathcal{F}(t_s). \tag{A-2}$$

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 de réciprocité des dérivées :

$$\mathcal{F}^{*'}(s) = \mathcal{F}'^{-1}(s). \tag{A-3}$$

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

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

Fonctions concaves conjuguées : Soit $\mathcal{G}$ une fonction concave définie sur un ensemble $\mathcal{D}$. La fonction concave conjuguée est définie sur l'ensemble conjugué $\mathcal{D}^*$ par

$$\mathcal{G}^*(x^*) = \operatorname*{Inf}_{x\in\mathcal{D}}\{< x, x^* > - \mathcal{G}(x)\}.$$

1.2 Propriétés, inégalités de YOUNG

On suppose 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 $\boldsymbol{t}$. Pour des fonctions convexes fermées sur $\mathbb{R}^M$, on a (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 $\mathcal{F}^*$: la conjugaison convexe est « réversible » si la fonction initiale $\mathcal{F}$ est convexe fermée.

1.3 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, sous des conditions d'existence très larges,

$$\mu = \operatorname*{Inf}_{x\in\mathcal{C}\cap\mathcal{D}}\{\mathcal{F}(x) - \mathcal{G}(x)\} = \operatorname*{Sup}_{x^*\in\mathcal{C}^*\cap\mathcal{D}^*}\{\mathcal{G}^*(x^*) - \mathcal{F}^*(x^*)\}.$$

Sous quelques conditions peu restrictives, les solutions optimales sont liées par une simple opération de dérivation : si $\boldsymbol{x}_0$ et $\boldsymbol{x}_0^*$ sont respectivement les solutions des problèmes primal et dual, alors

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

RÉFÉRENCES

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.

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.

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

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

C. HEINRICH, J.-F. BERCHER, G. LE BESNERAIS & G. DEMOMENT (1995), « Méthode du maximum d'entropie sur la moyenne et mélanges de distributions », dans *Actes du 15e colloque* GRETSI, septembre 1995.

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

Plusieurs articles des auteurs en rapport avec ce travail, dont une version longue de cette communication, peuvent être obtenus sur le ftp anonyme ftp.supelec.fr, dans le répertoire /lss/Papers/Bercher.