AIMÉ FUCHS
GIORGIO LETTA

L'inégalité de Kullback. Application à la théorie de l'estimation

Séminaire de probabilités (Strasbourg), tome 4 (1970), p. 108-131
<http://www.numdam.org/item?id=SPS_1970__4__108_0>

INSTITUT DE RECHERCHE MATHÉMATIQUE AVANCÉE
Laboratoire Associé au C.N.R.S.
Rue René Descartes
STRASBOURG 1969-70

L'INÉGALITÉ DE KULLBACK.
APPLICATION À LA THÉORIE DE L'ESTIMATION.

par A. FUCHS et G. LETTA.

INTRODUCTION

Le but principal de cet exposé est de montrer que la notion de gain d'information (de Shannon) permet de caractériser de façon plus naturelle, et dans un cadre plus général, la notion de résumé exhaustif qui intervient dans la théorie de l'estimation, notion qui est traditionnellement reliée à celle d'information de Fisher.

On montrera la connexion entre cette notion et celle de gain d'information ; il nous sera ensuite facile de retrouver l'inégalité de Cramer-Rao comme conséquence d'une inégalité, due à S. Kullback [6] , fournissant une minoration pour le gain d'information. Nous démontrerons tout d'abord cette inégalité, qui est fondée sur les deux notions de "famille conjuguée d'une loi de probabilité" et de "fonction conjuguée d'une fonction convexe".

Cette dernière notion, introduite par S. Mandelbrojt [7] et W. FENCHEL [2] (voir aussi [5]), ne semble pas être bien connue, en dépit de son importance. C'est pour cette raison qu'il nous a paru opportun de la développer dans toute sa généralité (à savoir pour des fonctions définies sur un espace localement convexe) dans un appendice faisant suite à l'exposé.

1. RAPPEL DE LA NOTION DE GAIN D'INFORMATION.

Nous nous bornerons ici à donner la définition et les principales propriétés de cette notion, renvoyant pour plus de renseignements à [6, chap. 1].

DÉFINITION.

Soit $(E , \mathcal{E})$ un espace mesurable. Pour tout couple μ, μ' de lois de probabilité sur $(E , \mathcal{E})$ équivalentes (c'est-à-dire absolument continues l'une par rapport à l'autre), on pose

$$G(\mu'|\mu) = \int \left(\log \frac{d\mu'}{d\mu}\right) d\mu' \quad ,$$

où $\frac{d\mu'}{d\mu}$ désigne une version de la dérivée de Radon-Nikodym de μ' par rapport à μ. On appelle $G(\mu'|\mu)$ le gain d'information (de Shannon) réalisé en remplaçant μ par μ'.

PROPRIÉTÉ 1.

$0 \leq G(\mu'|\mu) \leq +\infty$, et l'on a $G(\mu'|\mu) = 0$ si et seulement si $\mu = \mu'$.

PROPRIÉTÉ 2.

Pour toute loi de probabilité ν équivalente à μ et à μ', on a

$$G(\mu'|\mu) = G(\mu'|\nu) + \int \left(\log \frac{d\nu}{d\mu}\right) d\mu' \quad ,$$

d'où

$$G(\mu'|\mu) = \int \left(\log \frac{d\mu'}{d\mu}\right) d\mu' \geq \int \left(\log \frac{d\nu}{d\mu}\right) d\mu' \quad .$$

Pour que cette inégalité soit une égalité, il faut et il suffit que l'on ait, ou bien $\nu = \mu'$, ou bien $\int (\log \frac{d\nu}{d\mu})\, d\mu' = +\infty$.

PROPRIÉTÉ 3.

Pour toute application mesurable φ de l'espace mesurable $(E , \mathcal{E})$ dans un espace mesurable $(F , \mathcal{F})$, on a (en désignant par $\varphi(\mu)$, $\varphi(\mu')$ les lois image de μ,μ' par φ)

$$G(\varphi(\mu') \mid \varphi(\mu)) \leq G(\mu' \mid \mu) \quad ,$$

avec égalité si φ est une bijection dont l'inverse est mesurable.

2. FAMILLE CONJUGUÉE D'UNE LOI DE PROBABILITÉ SUR $\mathbb{R}$.

Soit μ une loi de probabilité sur $\mathbb{R}$, admettant un moment du premier ordre m . On appelle fonction génératrice des moments de la loi μ la fonction g définie par

$$g(u) = \int_{\mathbb{R}} e^{ux}\, \mu(dx) \qquad \text{pour tout } u \in \mathbb{R} \ .$$

On a évidemment $0 < g(u) \leq +\infty$ pour tout $u \in \mathbb{R}$, $g(0) = 1$. D'après le lemme de Fatou, la fonction g est semi-continue inférieurement. Nous désignerons par I l'ensemble constitué par les u tels que $g(u) < +\infty$. Il est facile de voir que I est un intervalle (contenant l'origine) , que la restriction de g à I est convexe et que la restriction de g à l'adhérence de I (dans $\mathbb{R}$) est continue.

Si l'intervalle I n'est pas réduit à l'origine, alors g est indéfiniment dérivable à l'intérieur de I. Pour tout point u intérieur à I et pour tout entier $n \geq 0$, on a en outre $\int_{\mathbb{R}} |x|^n e^{ux} \mu(dx) < +\infty$ et

$$(1) \qquad g^{(n)}(u) = \int_{\mathbb{R}} x^n e^{ux} \mu(dx) \quad .$$

Notons que, pour que l'intervalle I soit un voisinage de l'origine, il faut et il suffit que la fonction caractéristique associée à la loi μ soit analytique ; la relation (1) donne alors, en particulier :

$$g^{(n)}(0) = \int_{\mathbb{R}} x^n \mu(dx)$$

pour tout entier $n \geq 0$ (ceci justifie le nom de fonction génératrice des moments donné à g).

Si l'intervalle I est fermé à droite (et non réduit à l'origine) et si b est son extrémité droite, on a

$$g'_g(b) = \lim_{\substack{u\to b \\ u<b}} g'(u) = \int_{\mathbb{R}} x\, e^{bx} \mu(dx) \leq +\infty \quad .$$

Ceci résulte immédiatement du théorème de convergence monotone (Beppo-Levi) en remarquant que, pour tout $x \in \mathbb{R}$, la fonction $u \to xe^{ux}$ est croissante.

On a naturellement une propriété analogue si I est fermé à gauche.

Nous désignerons par $\check{I}$ l'intervalle constitué par les $u \in I$ tels que l'on ait $\int_{\mathbb{R}} |x| e^{ux} \mu(dx) < +\infty$. On remarquera que $\check{I}$ contient l'origine et que les seuls points de I qui peuvent ne pas appartenir à $\check{I}$ sont l'extrémité droite de I (dans le cas où I est fermé à droite) et l'extrémité gauche de I (dans le cas où I est fermé à gauche).

Pour tout $u \in \overset{\vee}{I}$ nous désignerons par μ_u la loi de probabilité sur R définie par

$$(2) \qquad \mu_u(A) = \frac{1}{g(u)} \int_A e^{ux} \mu(dx)$$

pour tout ensemble borélien A de R . En d'autres termes, μ_u est la loi obtenue en normalisant la mesure, de base μ, définie par la densité $x \mapsto e^{ux}$.

<u>DÉFINITION</u> (cf. [4, chap. 3].

La famille $(\mu_u)_{u\in \overset{\vee}{I}}$ de lois définies par (2) est appelée la <u>famille conjuguée</u> (<u>au sens de Khintchine</u>)de la loi μ .

On remarquera que l'on a $\mu_o = \mu$.

On appelle <u>fonction génératrice des cumulants</u> de la loi μ la fonction Ψ définie par

$$\Psi(u) = \log g(u) = \log \int_R e^{ux} \mu(dx) \quad \text{pour tout} \quad u \in R .$$

On a $\Psi(0) = 0$. Il est bien connu que la restriction de Ψ à I est convexe (pour qu'elle soit strictement convexe, il faut et il suffit que la loi μ ne soit pas dégénérée).

Pour tout $u \in \overset{\vee}{I}$ nous désignerons par m_u le moment du premier ordre de la loi μ_u . Pour tout point u intérieur à I , on a

$$m_u = \frac{\int_R x\, e^{ux} \mu(dx}{\int_R e^{ux} \mu(dx)} = \frac{g'(u)}{g(u)} = \Psi'(u) \quad .$$

De même, si $u \in \check{I}$ est l'extrémité droite (resp. gauche) de l'intervalle I (non réduit à l'origine), on a $m_u = \Psi'_g(u)$ (resp. $m_u = \Psi'_d(u)$) .
Dans tous les cas on a donc la propriété suivante :

PROPRIÉTÉ (E).

Pour tout $u \in \check{I}$, la droite de $\mathbb{R}^2$, de pente m_u , passant par le point $(u , \Psi(u))$ est une droite d'appui de l'ensemble (convexe) $\{(u,s) \in \mathbb{R}^2 : s \geq \Psi(u)\}$.

Remarque.

Pour tout point u intérieur à I , on a

$$\Psi''(u) = \frac{g''(u)\, g(u) - (g'(u))^2}{g(u)} \quad .$$

En particulier, si I est un voisinage de l'origine, on a $\Psi''(0) = \int x^2\, \mu(dx) - (\int x\mu\,(dx))^2 = \sigma^2$ (moment centré du second ordre de la loi μ) .

Plus en général, si on suppose seulement que l'on ait $\sup I > 0$ (resp. $\inf I < 0$) et que la loi μ soit du second ordre, on a $\lim\limits_{u\downarrow 0} \Psi''(u) = \sigma^2$ (resp. $\lim\limits_{u\uparrow 0} \Psi''(u) = \sigma^2$) .

3. LA FONCTION Ψ^* .

On a vu dans le paragraphe précédent que la restriction de la fonction Ψ à l'intervalle I où Ψ est finie est une fonction convexe. La fonction Ψ est en outre semi-continue inférieurement. On peut donc lui associer sa fonction conjuguée Ψ^* au sens précisé dans l'Appendice. Cette fonction peut être représentée de façon explicite sous la forme suivante :

$$\Psi^*(v) = \sup_{u \in \mathbb{R}} [uv - \Psi(u)] \quad \text{pour tout} \quad v \in \mathbb{R} \ .$$

L'ensemble constitué par les v vérifiant $\Psi^*(v) < +\infty$ est un intervalle et la restriction de Ψ^* à cet intervalle est une fonction convexe. En outre la propriété (E) de la fonction Ψ se traduit (cf. Appendice, prop. (2.9)) en la propriété (E^*) suivante de la fonction Ψ^* :

PROPRIÉTÉ (E^*).

Pour tout $u \in \check{I}$ on a $\Psi^*(m_u) = u\, m_u - \Psi(u)$ et la droite de $\mathbb{R}^2$, de pente u, passant par le point $(m_u , \Psi(m_u))$ est une droite d'appui de l'ensemble (convexe) $\{(v,t) \in \mathbb{R}^2 : t \geq \Psi^*(v)\}$.

Remarque.

En particulier il en résulte que la fonction Ψ^* est positive et qu'elle s'annule au point $m_0 = m$. Si l'intervalle I est réduit à l'origine, la fonction Ψ^* est identiquement nulle ; si au contraire on a $\sup I > 0$ (resp. $\inf I < 0$) et si la loi μ admet un moment centré du second ordre $\sigma^2 > 0$, alors on a $\Psi^*(m + h) = \dfrac{h^2}{2\sigma^2} + o(h^2)$ lorsque $h \downarrow 0$ (resp. $h \uparrow 0$) .

En effet, si $\sup I > 0$, on a, d'après la remarque à la fin du paragraphe précédent, $\Psi''(u) > 0$ pour tout $u > 0$ assez petit. Il en résulte (cf. Appendice, prop. (3.2))

$$\Psi^{*''}(m + h) = \frac{1}{\Psi''(\Psi'(m + h))}$$

pour tout $h > 0$ assez petit, et par suite

$$\lim_{h \downarrow 0} \Psi^{*''}(m + h) = \frac{1}{\sigma^2} \quad .$$

4. L'INÉGALITÉ DE KULLBACK.

Soient maintenant μ, μ' deux lois de probabilité sur $\mathbb{R}$ équivalentes, admettant des moments du premier ordre m, m' , et désignons par Ψ , comme dans le paragraphe précédent, la fonction génératrice des cumulants de la loi μ .

THÉORÈME.

On a l'inégalité suivante ("inégalité de Kullback" ; cf. [6, p. 38]) :

$$G(\mu'|\mu) \geq \Psi^*(m') \quad .$$

En outre, si le moment m' coïncide avec le moment du premier ordre d'un élément μ_u de la famille conjuguée de μ , alors, pour que l'inégalité précédente soit une égalité, il faut et il suffit que $\mu' = \mu_u$.

DÉMONSTRATION.

Nous utiliserons les notations du paragraphe précédent.

a) Pour tout $u \in \check{I}$ on a, d'après la propriété 2 du § 1 ,

$$G(\mu'|\mu) \geq \int_{\mathbb{R}} \left(\log \frac{d\mu_u}{d\mu}\right) d\mu' = \int_{\mathbb{R}} \left(\log \frac{e^{ux}}{g(u)}\right) \mu'(dx) \tag{1}$$

$$= \int_{\mathbb{R}} (ux - \Psi(u))\, \mu'(dx) = um' - \Psi(u) \quad .$$

Il en résulte

$$G(\mu'|\mu) \geq \sup_{u \in \check{I}} [um' - \Psi(u)] = \Psi^*(m') \quad .$$

b) Supposons à présent qu'il existe un élément u de $\check{I}$ tel que $m' = m_u$. D'après la propriété (E^*) du paragraphe précédent, la relation (1) donne

$$\int \left(\log \frac{d\mu_u}{d\mu}\right) d\mu' = \Psi^*(m') \quad ,$$

de sorte que l'égalité $G(\mu'|\mu) = \Psi^*(m')$ équivaut à l'égalité

$$G(\mu'|\mu) = \int_R \left(\log \frac{d\mu_u}{d\mu}\right) d\mu' \quad ,$$

c'est-à-dire (d'après la propriété 2 du § 1) à $\mu' = \mu_n$.

<u>Remarque.</u>

On notera que si I est réduit à l'origine, alors Ψ^* est identiquement nulle et l'inégalité de Kullback n'affirme rien d'autre que la positivité du gain.

5. LA NOTION DE RÉSUMÉ EXHAUSTIF.

Soit $(P_\theta)_{\theta \in \Theta}$ une famille de lois de probabilité équivalentes sur un espace mesurable $(\Omega , \mathfrak{F})$.

Pour toute variable aléatoire X sur $(\Omega , \mathfrak{F})$, à valeurs dans un espace mesurable $(E , \mathcal{E})$, nous désignerons par $G_X(\theta'|\theta)$ le gain d'information $G(X(\theta') \mid X(\theta))$ (où $X(P_\theta)$, $X(P_{\theta'})$ désignent les lois image de P_θ, $P_{\theta'}$ par X).

Soit maintenant X une variable aléatoire réelle sur $(\Omega , \mathfrak{F})$ et soit $X_1, \ldots, X_n$ un système de n variables aléatoires réelles sur $(\Omega , \mathfrak{F})$ tel que, pour tout $\theta \in \Theta$ $(X_1, \ldots, X_n)$ soit un n-échantillon issu de X sur l'espace probabilisé $(\Omega , \mathfrak{F} , P_\theta)$.

Pour tout couple θ,θ' d'éléments de Θ , on a alors

(1) $$G_{(X_1, \ldots, X_n)}(\theta'|\theta) = n\, G_X(\theta'|\theta)$$

("propriété d'additivité"). Si φ est une fonction réelle borélienne sur R^n , si $T = \varphi(X_1, \ldots, X_n)$, on a aussi (d'après la propriété 3 du § 1)

(2) $$G_T(\theta'|\theta) \leq G_{(X_1, \ldots, X_n)}(\theta'|\theta) \quad .$$

On dira que l'estimateur $T = \varphi(X_1, \ldots, X_n)$ est un <u>résumé exhaustif</u> de $(X_1, \ldots, X_n)$ si, pour tout couple θ,θ' d'éléments de Θ , on a égalité dans la relation (2).

Supposons maintenant que Θ soit un intervalle ouvert de R , que, pour tout $\theta \in \Theta$, la loi image $X(P_\theta)$ admette une densité $f(\cdot\, , \theta)$ par rapport à la mesure de Lebesgue sur R et que la fonction φ , ainsi que les densités $f(\cdot\, , \theta)$, vérifient les conditions de régularité précisées dans [3, chap. 13].

Il est facile de voir que la notion de résumé exhaustif donnée ci-dessus coïncide alors avec la notion classique.

6. CONNEXION ENTRE LE GAIN D'INFORMATION DE SHANNON ET L'INFORMATION DE FISHER.

Dans ses tentatives pour étudier l'efficacité d'un estimateur, R.A. Fisher a été conduit à introduire une quantité qui avait certaines propriétés d'une information ; il avait appelé cette quantité <u>quantité d'information que X apporte sur la paramètre θ</u> , et l'avait notée $I_X(\theta)$. L'étude de cette "information de Fisher" est faite dans les traités classiques de

Statistique en liaison avec la notion d'exhaustivité (voir, p.ex., [3, p. 181]) et nous n'y reviendrons pas ici. En revanche nous voudrions mettre en lumière la connexion entre cette information de Fisher et le gain d'information de Shannon, connexion qui avait été entrevue par Savage [8] et Kullback [6], mais qui ne semble pas être très connue. Voici en quoi elle consiste. Plaçons nous toujours dans les hypothèses précisées à la fin du paragraphe précédent et considérons, pour tout couple $\theta, \theta + h$ d'éléments de Θ, le gain $G_X(\theta + h|\theta)$. Il est intuitif qu'à cause des hypothèses de régularité portant sur les densités $f(\cdot, \theta)$, cette quantité tende vers 0 lorsque $h \to 0$; mais il est remarquable qu'en fait ce soit <u>un infiniment petit du second ordre en h</u>. De façon précise, on a

$$(1) \qquad G_X(\theta + h|\theta) = I_X(\theta)\,\frac{h^2}{2} + o(h^2) \quad .$$

En d'autres termes, $G_X(\theta + h|\theta)$ est un infiniment petit du second ordre en h et l'information de Fisher $I_X(\theta)$ intervient comme coefficient dans la partie principale de $G_X(\theta + h|\theta)$ lorsque h est petit.

La connexion établie ci-dessus permet de déduire, à partir des propriétés du gain d'information exprimées par les relations (1) et (2) du § 5, les propriétés correspondantes de l'information de Fisher :

$$(2) \qquad I_{(X_1, \ldots, X_n)}(\theta) = n\, I_X(\theta) \qquad \text{pour tout } \theta \in \Theta \quad ,$$

$$(3) \qquad I_T(\theta) \leq I_{(X_1, \ldots, X_n)}(\theta) \qquad \text{pour tout } \theta \in \Theta$$

(avec égalité si et seulement si T est un résumé exhaustif).

7. L'INÉGALITÉ DE CRAMER-RAO.

Nous allons voir aussi que, moyennant la connexion ci-dessus, l'inégalité de Cramer-Rao résulte immédiatement de celle de Kullback.

Gardons les hypothèses formulées à la fin du § 5 et supposons en outre que, pour tout $\theta \in \Theta$, la variable aléatoire X sur l'espace probabilisé $(\Omega , \mathfrak{F} , P_\theta)$ admette une variance finie $\sigma_\theta^2 > 0$ et une espérance égale à θ ("X sans biais"). Nous désignerons par Ψ_θ la fonction génératrice des cumulants de la loi $X(P_\theta)$ et nous supposerons que, pour tout $\theta \in \Theta$, l'intervalle $I_\theta = \{v : \Psi_\theta(v) < + \infty\}$ soit un voisinage de l'origine. On a alors, d'après l'inégalité de Kullback,

$$G_X(\theta + h|\theta) \geq \Psi_\theta^*(\theta + h)$$

pour tout couple $\theta, \theta + h$ d'éléments de Θ.

Or, d'après la relation (1), le premier membre est égal à $I_X(\theta)\frac{h^2}{2} + o(h^2)$ et, d'après la remarque finale du § 3, le second membre vaut $\frac{h^2}{2\sigma_\theta^2} + o(h^2)$. Il en résulte l'inégalité de Cramer-Rao :

$$I_X(\theta) \geq \frac{1}{\sigma_\theta^2} .$$

Remarque.

La démonstration précédente est encore valable si l'on suppose seulement que, pour tout $\theta \in \Theta$, l'intervalle I_θ ne se réduit pas à l'origine.

APPENDICE

LA CONJUGAISON DES FONCTIONS CONVEXES.

1. RAPPEL SUR LA NOTION DE CONNEXION DE GALOIS.

Soient A,B deux ensembles, G une partie de $A \times B$.

Pour toute partie X de A , désignons par σX la partie de B constituée par les éléments y de B tels que l'on ait $(x,y) \in G$ pour tout élément x de X .

De façon analogue, pour toute partie Y de B , désignons par τY la partie de A constituée par les éléments x de A tels que l'on ait $(x,y) \in G$ pour tout élément y de Y .

Le couple (σ,τ) , constitué par les applications σ (de $\mathcal{P}(A)$ dans $\mathcal{P}(B)$) et τ (de $\mathcal{P}(B)$ dans $\mathcal{P}(A)$), est appelé la <u>connexion de Galois</u> associée à la correspondance (A , B , G) [cf. 9, pp. 70-74] .

(1.1.) Les applications σ et τ sont décroissantes (lorsqu'on considère les ensembles $\mathcal{P}(A)$ et $\mathcal{P}(B)$ ordonnés par l'inclusion). On a $\sigma\emptyset = B$, $\tau\emptyset = A$ et

$$\sigma(\bigcup_i X_i) = \bigcap_i \sigma X_i \qquad , \qquad \tau(\bigcup_i Y_i) = \bigcap_i \tau Y_i$$

pour toute famille non vide (X_i) de parties de A et pour toute famille non vide (Y_i) de parties de B .

(1.2.) L'application $\tau \circ \sigma$ est une opération d'enveloppe dans $\mathcal{P}(A)$ (c'est-à-dire une application croissante, extensive et idempotente de $\mathcal{P}(A)$ dans lui-même). De même, l'application $\sigma \circ \tau$ est une opération d'enveloppe dans $\mathcal{P}(B)$.

Les parties X de A qui sont fermées pour l'opération d'enveloppe $\tau \circ \sigma$ (c'est-à-dire telles que l'on ait $\tau\sigma X = X$) seront appelées les ensembles <u>$\tau\sigma$-fermés</u>.

De même, les parties Y de B qui sont fermées pour l'opération d'enveloppe $\sigma \circ \tau$ seront appelées les ensembles <u>$\sigma\tau$-fermés.</u>

(1.3.) Pour qu'une partie X de A soit $\tau\sigma$-fermée, il faut et il suffit qu'elle soit de la forme τY , avec $Y \subset B$, ou (ce qui revient au même) qu'elle soit l'intersection, dans A , d'une famille (éventuellement vide) de parties de A de la fórme $\tau\{y\}$, avec $y \in B$. (En particulier, l'ensemble A et l'ensemble vide sont $\tau\sigma$-fermés). On a une caractérisation analogue pour les ensembles $\sigma\tau$-fermés.

(1.4.) Si à tout ensemble $\tau\sigma$-fermé X on associe l'ensemble $\sigma\tau$-fermé σX , on obtient une bijection de la classe des ensembles $\tau\sigma$-fermés sur la classe des ensembles $\sigma\tau$-fermés (la bijection réciproque étant celle qui, à tout ensemble $\sigma\tau$-fermé Y , associe l'ensemble $\tau\sigma$-fermé τY). Un ensemble $\tau\sigma$-fermé X et un ensemble $\sigma\tau$-fermé Y , qui se correspondent dans cette bijection, c'est-à-dire tels que l'on ait $\sigma X = Y$ (et par suite $\tau Y = X$) , seront dits <u>conjugués</u> dans la connexion de Galois (σ,τ) .

(1.5.) Pour qu'un ensemble $\tau\sigma$-fermé X et un ensemble $\sigma\tau$-fermé Y soient conjugués dans la connexion de Galois (σ,τ) , il faut et il suffit que l'on ait $X \times Y \subset G$.

2. LA CONJUGAISON DES FONCTIONS CONVEXES.

Si f est une fonction numérique, définie dans un ensemble E , nous désignerons par $D(f)$ la partie $E \times R$ constituée par les points $(x,t) \in E \times R$ qui sont "au-dessus du graphe de f" :

$$D(f) = \{(x,t) \in E \times R : f(x) \leq t\} \quad .$$

On a $f_1 \leq f_2 \Leftrightarrow D(f_1) \supset D(f_2)$ pour tout couple f_1 , f_2 de fonctions numériques définies dans E. L'ensemble $D(f)$ **détermine** donc la fonction f . On a en outre $D(+\infty) = \emptyset$, $D(-\infty) = E \times R$ et

$$D(\sup_i f_i) = \bigcap_i D(f_i)$$

pour toute famille non vide (f_i) de fonctions numériques définies dans E .

(2.1.) PROPOSITION.

Soient E un espace localement convexe réel, f une fonction définie dans E , à valeurs dans $]-\infty , +\infty]$, distincte de la constante $+\infty$. Les conditions suivantes sont alors équivalentes :

a) f est l'enveloppe supérieure d'une famille de fonctions linéaires affines continues sur E ;

b) f est semi-continue inférieurement, l'ensemble où f est finie est convexe et la restriction de f à cet ensemble est une fonction convexe ;

c) l'ensemble $D(f)$ est fermé et convexe dans $E \times R$.

DÉMONSTRATION.

Les implications $(a) \Rightarrow (b)$ et $(b) \Rightarrow (c)$ sont immédiates. Il reste à démontrer l'implication $(c) \Rightarrow (a)$.

Soient x_o un élément de E et t_o un nombre réel tel que $t_o < f(x_o)$. Nous montrerons qu'il existe dans $E \times R$ un hyperplan fermé qui sépare strictement le point (x_o , t_o) et l'ensemble $D(f)$ et qui est le graphe d'une fonction linéaire affine continue sur E.

(1) Supposons d'abord que $f(x_o)$ soit fini. D'après un théorème bien connu concernant la séparation des ensembles convexes dans un espace localement convexe (cf. [1, § 5, n° 3, Prop. 4]), on sait qu'il existe dans $E \times R$ un hyperplan fermé H qui sépare strictement (x_o , t_o) et $D(f)$. L'hyperplan H a une équation de la forme

$$H = \{(x,t) : g(x) + \alpha t = 0 \}$$

où g est une fonction linéaire affine continue sur E et α est un scalaire. On ne peut avoir $\alpha = 0$, car alors les deux points $(x_o , f(x_o)) \in D(f)$ et $(x_o , t_o) \notin D(f)$ seraient d'un même côté de H.

On a donc $\alpha \neq 0$, de sorte que l'hyperplan $H = \{(x,t) : t = - \alpha^{-1} g(x)\}$ est bien le graphe d'une fonction linéaire affine continue sur E.

(2) Supposons maintenant $f(x_o) = + \infty$. Comme f n'est pas la constante $+ \infty$, il existe un point x_1 tel que $f(x_1)$ soit fini. On peut choisir un scalaire t_1 assez petit pour que le segment K d'extrémités (x_o , t_o) et (x_1 , t_1) ne rencontre pas l'ensemble $D(f)$. [En effet, pour tout scalaire t, l'ensemble

$$L(t) = \{\lambda \in [0,1] : f(x_o + \lambda(x_1-x_o)) \leq t_o + \lambda(t-t_o)\}$$

est compact, et on a $L(t') \subset L(t)$ si $t' < t$, $\bigcap_{t\in R} L(t) = \emptyset$: donc $L(t) = \emptyset$ dès que t est assez petit].

D'après le théorème de séparation cité plus haut, il existe un hyperplan fermé H qui sépare strictement K et $D(f)$. D'après (1) , H est nécessairement le graphe d'une fonction linéaire affine continue sur E .

Si E est un espace localement convexe réel, nous désignerons par $\mathcal{H}(E)$ l'ensemble constitué par les fonctions f_α définies dans E , à valeurs dans $]-\infty , +\infty]$, et distinctes de la constante $+\infty$ et satisfaisant aux conditions équivalentes (a) , (b) , (c) de (2.1.).

Soient maintenant E et F deux espaces vectoriels réels, mis en dualité par la forme bilinéaire $(x,y) \mapsto \langle x,y \rangle$, et munissons E et F des topologies faibles, $\sigma(E , F)$ et $\sigma(F , E)$, définies par cette dualité.

Les formes linéaires continues sur E sont alors les formes du type $x \mapsto \langle x,y \rangle$ avec y élément de F . De même, les formes linéaires continues sur F sont les formes du type $y \mapsto \langle x,y \rangle$, avec x élément de E .

(2.2.) <u>DÉFINITION</u>. (cf. [2] , [5]).

Soient f une fonction de la classe $\mathcal{H}(E)$, g une fonction de la classe $\mathcal{H}(F)$. On dit que f et g sont <u>conjuguées</u>, si on a l'"inégalité de W.H. Young" :

$$\langle x,y \rangle \leq f(x) + g(y) \quad \text{pour tout} \quad (x,y) \in E \times F \ .$$

Nous nous proposons de montrer que la notion de conjugaison ainsi introduite peut être ramenée à la notion générale de connexion de Galois.

Considérons, à cet effet, la correspondance entre les ensembles $A = E \times R$ et $B = F \times R$, dont le graphe est l'ensemble G constitué par les couples $((x,t), (y,u)) \in A \times B$ tels que

$$< x,y > \leq t + u \quad ;$$

et désignons par (σ,τ) la connexion de Galois associée à cette correspondance.

(2.3.) PROPOSITION.

Soit X une partie non vide de $E \times R$, distincte de $E \times R$. Pour que X soit un ensemble $\tau\sigma$-fermé, il faut et il suffit qu'il existe une fonction $f \in \mathcal{H}(E)$ telle que l'on ait $X = D(f)$, (On a une caractérisation analogue pour les ensembles $\sigma\tau$-fermés).

DÉMONSTRATION.

En effet, pour que X soit un ensemble $\tau\sigma$-fermé, il faut et il suffit, d'après (1.3.), qu'il existe une famille (y_i , u_i) d'éléments de $F \times R$, telle que l'on ait

$$X = \bigcap_i \tau \{(y_i , u_i)\} \quad ,$$

c'est-à-dire

$$X = \bigcap_i D(f_i) = D(\sup_i f_i) \quad ,$$

où, pour tout i, on a désigné par f_i la fonction linéaire affine sur E définie par

$$f_i(x) = < x,y_i > - u_i \quad .$$

(2.4.) PROPOSITION.

Soient f une fonction de la classe $\mathcal{H}(E)$, g une fonction de la classe $\mathcal{H}(F)$. Pour que f et g soient conjuguées au sens de la définition (2.2.), il faut et il suffit que les ensembles $D(f)$ et $D(g)$ soient conjugués dans la connexion de Galois (σ , τ) .

DÉMONSTRATION.

Pour que les ensembles $D(f)$ et $D(g)$ soient conjugués dans la connexion de Galois (σ , τ) , il faut et il suffit (d'après (1.5.)) que l'on ait

$$< x,y > \leq t + u$$

pour tout $(x,t) \in D(f)$ et pour tout $(y,u) \in D(g)$.

Or cette condition équivaut évidemment à la relation

$$< x,y > \leq f(x) + g(y) \quad \text{pour tout} \quad (x,y) \in E \times F \ .$$

(2.5.) COROLLAIRE.

Pour toute fonction $f \in \mathcal{H}(E)$, il existe une fonction unique $g \in \mathcal{H}(F)$ telle que f et g soient conjuguées. Pour toute fonction $g \in \mathcal{H}(F)$, il existe une fonction unique $f \in \mathcal{H}(E)$ telle que f et g soient conjuguées.

(2.6.) PROPOSITION.

Soient $f \in \mathcal{H}(E)$ et $g \in \mathcal{H}(F)$ deux fonctions conjuguées. On a alors

$$g(y) = \sup_{x \in E} [< x,y > - f(x)] \quad \text{pour tout} \quad y \in F \ , \tag{2.7.}$$

$$f(x) = \sup_{y \in F} [< x,y > - g(y)] \quad \text{pour tout} \quad x \in E \ . \tag{2.8.}$$

DÉMONSTRATION.

Il suffit de démontrer la relation (2.7.) . Soit $y \in F$. On a évidemment $g(y) \geq \sup_{x \in E} [< x,y > - f(x)]$. Soit u un nombre réel tel que l'on ait

$$u \geq < x,y > - f(x) \qquad \text{pour tout} \quad x \in E \quad ,$$

et montrons qu'on a nécessairement $u \geq g(y)$ (ce qui achèvera la démonstration).

Pour tout $(x,t) \in D(f)$, on a $< x,y > \leq f(x) + u \leq t + u$.

Cela entraîne $(y,u) \in \sigma\, D(f) = D(g)$, c'est-à-dire $u \geq g(y)$.

(2.9.) PROPOSITION.

Soient $f \in \mathcal{H}(E)$ et $g \in \mathcal{H}(F)$ deux fonctions conjuguées. Pour tout couple $(x_o , y_o) \in E \times F$, les trois propriétés suivantes sont alors équivalentes :

(a) $\qquad < x_o , y_o > = f(x_o) + g(y_o)$;

(b) $\qquad f(x_o)$ est fini et l'hyperplan

$$H = \{(x,t) \in E \times R : t = f(x_o) + < x - x_o , y_o >\}$$

est un hyperplan d'appui de $D(f)$ (dans $E \times R$) ;

(c) $\qquad g(y_o)$ est fini et l'hyperplan

$$K = \{(y,u) \in F \times R : u = g(y_o) + < x_o , y - y_o >\}$$

est un hyperplan d'appui de $D(g)$ (dans $F \times R$) .

DÉMONSTRATION.

Il suffit de démontrer l'équivalence des conditions (a) et (b). D'après (2.6.) on a

$$g(y_o) = \sup_{x\in E} [< x,y_o > - f(x)] \quad .$$

La condition (a) équivaut donc à l'inégalité

$$< x,y_o > - f(x) \leq < x_o , y_o > - f(x_o) \qquad \text{pour tout} \quad x \in E \quad ,$$

et cette inégalité exprime précisément la condition (b).

3. LE CAS DES FONCTIONS D'UNE VARIABLE RÉELLE.

Plaçons nous maintenant dans le cas particulier où les espaces vectoriels E et F coïncident avec R (et où la forme bilinéaire est donnée par $< x,y > = xy$).

Pour toute fonction f de $\mathcal{H}(R)$, nous désignerons par f^* la fonction conjuguée de f (c'est-à-dire la seule fonction g de $\mathcal{H}(R)$ telle que f et g soient conjuguées).

La proposition (2.9.) donne lieu aux deux propositions suivantes

(3.1.) PROPOSITION.

Soit f une fonction de $\mathcal{H}(R)$. Pour que la fonction f^* soit strictement convexe, il faut et il suffit que f possède les propriétés suivantes :

(a) l'intervalle $I = \{x : f(x) < +\infty\}$ n'est pas réduit à un point ;

(b) la restriction de f à l'intérieur de I est dérivable, et sa dérivée n'est pas majorée (resp. minorée) si I est fermé à droite (resp. à gauche).

(3.2.) PROPOSITION.

Soit f une fonction de $\mathcal{H}(\bar{R})$ et soit $g = \dot{f}^*$ **la fonction conjuguée.** Soit U un intervalle ouvert de R, tel que la restriction de f à U soit finie, dérivable et strictement convexe. La restriction de g à l'intervalle ouvert $V = \{f'(x) : x \in U\}$ est alors finie, dérivable et strictement convexe. En outre, pour tout $x \in U$ et pour tout $y \in V$, les trois relations suivantes sont équivalentes :

$$y = f'(x) \qquad , \qquad x = g'(y) \qquad , \qquad xy = f(x) + g(y) \quad .$$

(En d'autres termes, l'application $x \mapsto f'(x)$ est un homéomorphisme de U sur V, dont l'homéomorphisme inverse est l'application $y \mapsto g'(y)$ de V sur U et dont le graphe est l'ensemble des couples $(x,y) \in U \times V$ tels que $xy = f(x) + g(y)$). Par conséquent, si la fonction f possède en tout point x de U une dérivée seconde $f''(x) > 0$, alors la fonction g possède en tout point y de V une dérivée seconde $g''(y) > 0$, et l'on a

$$f''(x) = \frac{1}{g''(f'(x))} \text{ pour } x \in U \quad , \quad g''(y) = \frac{1}{f''(g'(y))} \text{ pour } y \in V \, .$$

DÉMONSTRATION.

Soit y un élément de V et soit x l'élément (unique) de U tel que $f'(x) = y$. D'après (2.9.), on a $xy = f(x) + g(y)$; ceci prouve que la restriction de g à V est finie.

Soit encore y un élément de V et x l'unique élément de U tel que $f'(x) = y$. Les droites de pentes $x_1 = g'_g(y)$ et $x_2 = g'_d(y)$ passant par le point $(y,g(y))$ sont des droites d'appui de l'ensemble $D(g)$.

D'après (2.9.) il en résulte $f'(x_1) = y = f'(x_2)$, c'est-à-dire $x_1 = x_2 = x$. En d'autres termes, la fonction g est dérivable au point y et sa dérivée $g'(y)$ coïncide avec le seul élément x de U tel que $f'(x) = y$.

APPENDICE

BIBLIOGRAPHIE

[1] N. BOURBAKI Espaces vectoricls topologiques. Chap. II. Hermann (1966).

[2] W. FENCHEL On conjugate convex functions. Canad. J. of Math. - vol. 1 (1949) p. 73-77.

[3] C. FOURGEAUD & A. FUCHS - Statistique. Dunod (1967).

[4] J. KEILSON Green's function methods in Probability Theory. Griffin's Statistical Monographs (1965).

[5] M.A.KRASNOSELSKY & Y.B. RUTITSTY - Convex functions and Orlicz spaces. Hindustan publ. corp. (1962).

[6] S. KULLBACK Information Theory ans Statistics. J. Wiley (1959).

[7] S. MANDELBROJT Sur les fonctions convexes. C.R. Ac. Sc. Paris, vol. 209 (1939) 977-978.

[8] L.J. SAVAGE The foundations of Statistics. J. Wiley (1954).

[9] G. SZÁSZ Introduction to lattice theory. Ac. Press (1963).