MINIMIZATION OF ENTROPY FUNCTIONALS UNDER MOMENT CONSTRAINTS

I. Csiszár (Budapest)

Given a σ-finite measure space $(X, \mathcal{X}, \mu)$ and a d-tuple $\varphi = (\varphi_1, \dots, \varphi_d)$ of measurable functions on X, for $a = (a_1, \dots, a_d) \in \mathbb{R}^d$ let $\mathcal{L}_a$ denote the family of probability density functions g on X satisfying

$$\int \varphi g d\mu = a, \text{ that is, } \int \varphi_i g d\mu = a_i, \quad i = 1, \dots, d.$$

Extensively studied problem: minimize

$$J(g) = \int g \log g d\mu \qquad \text{(negative Shannon entropy)}$$

or

$$K(g, h) = \int g \log \frac{g}{h} d\mu \quad \text{(Kullback-Leibler distance, } I\text{-divergence, relative entropy)}$$

subject to $g \in \mathcal{L}_a$.

First this problem, then its extension to other entropies and distances will be considered.

For $\vartheta = (\vartheta_1, \ldots, \vartheta_d) \in \mathbb{R}^d$ denote

$$\wedge(\vartheta) = \log \int e^{\langle \vartheta, \varphi \rangle} d\mu \qquad \langle \vartheta, \varphi \rangle = \sum_{i=1}^{d} \vartheta_i \varphi_i$$

Assume: $dom(\wedge) = \{\vartheta : \wedge(\vartheta) < +\infty\}$ is nonempty.
Not hard to show: $\wedge(\vartheta)$ is the convex conjugate of the function $H(a) = \inf_{g \in \mathcal{L}_a} J(g)$:

$$\wedge(\vartheta) = H^*(\vartheta) = \sup_{a \in \mathbb{R}^d} [\langle \vartheta, \varphi \rangle - H(a)].$$

Dual problem associated with the primal problem of minimizing $J(g)$ subject to $g \in \mathcal{L}_a$: maximize $\ell_a(\vartheta) = \langle \vartheta, \varphi \rangle - \wedge(\vartheta)$ for $\vartheta \in \mathbb{R}^d$.
The supremum of $\ell_a(\vartheta)$ is the convex conjugate of $\wedge(\vartheta)$, thus the second conjugate $H^{**}(a)$ of $H(a)$. Always $H(a) \geq H^{**}(A)$, the difference is called duality gap.

Exponential family with canonical statistic φ :

$$\mathcal{E} = \{f_\vartheta = e^{\langle \vartheta, \varphi \rangle - \wedge(\vartheta)} : \ \vartheta \in dom(\wedge)\}.$$

When the empirical mean $\frac{1}{n}\sum_{j=1}^{n} \varphi(x_j)$ of φ in a sample $x_1, \ldots, x_n$ drawn from a density in $\mathcal{E}$ is equal to a, the normalized log-likelihood function is $\ell_a(\vartheta)$; for this a the dual problem means ML estimation.

Moreover, if $g \in \mathcal{L}_a$ then

(lik.id) $\quad K(g, f_\vartheta) = J(g) - \ell_a(\vartheta), \quad \vartheta \in dom(\wedge),$

hence providing $J(g)$ is finite, the dual problem is equivalent to minimizing $K(g, f_\vartheta)$ for $f_\vartheta \in \mathcal{E}$.

Note: this interpretation of the dual problem does not apply if $a \notin dom(H)$, in which case $J(g) = +\infty$ for all $g \in \mathcal{L}_a$ even though $H^{**}(a) < H(a) = +\infty$ is possible.

Elementary proposition: If $\mathcal{L}_a \cap \mathcal{E} \neq \emptyset$, it contains a single g_a , and for this

$$J(g) = J(g_a) + K(g, g_a), \quad g \in \mathcal{L}_a;$$

equivalently, the Pythagorean identity holds:

$$K(g, f) = K(g_a, f) + K(g, g_a), \quad g \in \mathcal{L}_a,\ f \in \mathcal{E}$$

In this case, the duality gap is 0:

$$H(a) = H^{**}(a) = J(g_a),$$

and the common member g_a of $\mathcal{L}_a$ and $\mathcal{E}$ is simultaneously the I-projection to $\mathcal{L}_a$ of each $f \in \mathcal{E}$ and the reverse I-projection to $\mathcal{E}$ of each $g \in \mathcal{L}_a$ with $J(g) < +\infty$.

HISTORY HINTS

Boltzmann, Gibbs: in 19. century
Jaynes, Kullback: in the fifties
Čencov 1972: information projections, diff. geom. approach
Barndorff - Nielsen 1977: convex analysis approach to MLE for exponential families
Csiszár 1975, Topsøe 1979: generalized minimizer when minimum not attained (Shannon case)
Csiszár 1991: axiomatic approach
Borwein and Lewis 1991: convex analysis approach for general "entropies"
Csiszár 1995: generalized minimizer, general case

Several recent works employ advanced Orlicz space techniques (Léonard 2001-2007) or diff. geom. (Amari and Nagaoka 2000, etc.)

This talk is based on works of Csiszár and Matúš 2001-2008 and hopefully will show that classical tools suffice for treating the problem efficiently.

Convex core of a finite measure Q on $\mathbb{R}^d$ (Csiszár - Matúš 2001):

$cc(Q) =$ intersection of all convex Borel sets with full Q-measure

$=$ set of means of all probability measures $P \ll Q$ that have mean

For the measure μ on X, define

$cc_\varphi(\mu) = \{ \int \varphi g d\mu : g$ prob.density, φg integrable $\}$

$= \{a \in \mathbb{R}^d : \mathcal{L}_a \neq \emptyset\}.$

If μ is finite then

$$cc_\varphi(\mu) = cc(\mu_\varphi), \quad \mu_\varphi \text{ image of } \mu \text{ on } \mathbb{R}^d.$$

Lemma: If $a \in cc_\varphi(\mu)$, there exists $g \in \mathcal{L}_a$ with $\mu(\{x : g(x) > 0\}) < +\infty$, g bounded.

Corollary: $dom(H) = cc_\varphi(\mu)$, that is, the necessary condition $\mathcal{L}_a \neq \emptyset$ for $H(a) = \inf_{g \in \mathcal{L}_a} J(g) < +\infty$ is sufficient, as well.

Face of a convex set $C \subset \mathbb{R}^d$: Nonempty convex subset $F \subset C$ such that a convex combination $tx + (1-t)y$ of $x \in C$ and $y \in C$ (with $0 < t < 1$) belongs to F only if $x, y \in F$

For a face F of $cc_\varphi(\mu)$, denote

$$\tilde{F} = \{x : \varphi(x) \in c\ell(F)\}$$

Lemma: For a in a face F of $cc_\varphi(\mu)$, each $g \in \mathcal{L}_a$ vanishes outside $\tilde{F}$ (μ-a.e.)

Extended exponential family $ext\mathcal{E}$:
The union of the families $\mathcal{E}_F$ for all faces F of $cc_\varphi(\mu)$, where

$$\mathcal{E}_F = \{f_{F,\vartheta} = e^{\langle\vartheta,\varphi\rangle - \wedge_f(\vartheta)} \mathbf{1}_{\tilde{F}} \ : \vartheta \in dom(\wedge_F)\}$$

$$\wedge_F(\vartheta) = \log \int_{\tilde{F}} e^{\langle\vartheta,\varphi\rangle} d\mu$$

Theorem 1 (Csiszár - Matúš 2003): Whenever $\mathcal{L}_a \neq \emptyset$ thus $a \in cc_\varphi(\mu)$, there exists a unique g_a, perhaps not in $\mathcal{L}_a$, such that

$$J(g) = H(a) + K(g, g_a), \ \ \forall g \in \mathcal{L}_a$$

Moreover $g_a \in \mathcal{E}_F$, for the face F of $cc_\varphi(\mu)$ whose relative interior contains a.

Clearly, if $g_a \in \mathcal{L}_a$ then it minimizes $J(g)$ subject to $g \in \mathcal{L}_a$. Otherwise, it is a generalized minimizer: every sequence g_n in $\mathcal{L}_a$ with $J(g_n) \to H(a)$ satisfies $K(g_n, g_a) \to 0$, in particular, $g_n \to g_a$ in $L_1(\mu)$.

Generalized Pythagorean identity:

$$K(g,f) = K(\mathcal{L}_a, f) + K(g, g_a) \;\; \forall g \in \mathcal{L}_a, f \in \mathcal{E}$$

where $K(\mathcal{L}_a, f) = \inf_{g \in \mathcal{L}_a} K(g,f) \geq K(g_a, f)$

Thus, g_a is the generalized I-projection to $\mathcal{L}_a$ of each $f \in \mathcal{E}$. If $a \in ri(cc_\varphi(\mu))$ thus $g_a \in \mathcal{E}$, then g_a is also the reverse I-projection to $\mathcal{E}$ of each $g \in \mathcal{L}_a$ with $J(g) < +\infty$, and the duality gap is zero.

$g_a \notin \mathcal{L}_a$ can happen if $g_a = f_\vartheta$ with ϑ on boundary of $dom(\wedge)$; g_a may be the same for several vectors a.

Existence of minimizer (I-projection): $g_a \in \mathcal{L}_a$ holds for all $a \in ri(cc_\varphi(\mu))$ if and only if $\wedge$ is steep, and for all $a \in ri(F)$, if and only if $\wedge_F$ is steep.

Theorem 2. (Csiszár - Matúš 2003. 2008): If $H^{**}(a) = \sup_{\vartheta\in\mathbb{R}^d} l_a(\vartheta)$ is finite, there exists a unique density h_a such that

$$H^{**}(a) - \ell_a(\vartheta) \geq K(h_a, f_\vartheta), \quad \vartheta \in dom(\wedge).$$

Moreover, $h_a \in \mathcal{E}_F$ where F is the largest face of $cc_\varphi(\mu)$ with $a \in ri(F) + barr(dom(\wedge))$.

Here $barr$ denotes barrier cone: for any convex set $C \subset \mathbb{R}^d$, $barr(C) = \{b : \sup_{c\in C}\langle b, c\rangle < +\infty\}$.

Supplement: $dom(H^{**}) = cc_\varphi(\mu) + barr(dom(\wedge))$.

The maximum of $\ell_a(\vartheta)$ is attained (MLE exists) if and any only if $h_a \in \mathcal{E}$. Otherwise, h_a is a generalized MLE: every sequence ϑ_n in $dom(\wedge)$ with $\ell_a(\vartheta_n) \to H^{**}(a)$ satisfies $K(h_a, f_{\vartheta_n}) \to 0$, in particular, $f_{\vartheta_n} \to h_a$ in $L_1(\mu)$.

GENERAL ENTROPY FUNCTIONALS

In the sequel, γ is a given strictly convex, differentiable function on $(0, +\infty)$, $\gamma(0)$ is defined as $\lim_{t \downarrow 0} \gamma(t)$; later, $\gamma'(0)$, $\gamma'(+\infty)$ are also defined limiting.

γ-entropy of a nonnegative function g on X:

$$J_\gamma(g) = \int \gamma(g) d\mu$$

Familiar choices of γ, in addition to $t \log t$:

$\gamma(t) = -\log t$	Burg entropy
$\gamma(t) = sign(\alpha - 1) t^\alpha$	Rényi (Tsallis) entropy

Problem: minimize $J_\gamma(g)$ subject to $g \in \mathcal{L}_a$, where $\mathcal{L}_a$ is defined slightly differently than before: attention is not restricted to probability densities, accordingly we set $\varphi = (\varphi_0, \varphi_1, \ldots, \varphi_d)$ with φ_0 identically 1, and for $a = (a_0, \ldots, a_d) \in \mathbb{R}^{1+d}$,

$$\mathcal{L}_a = \{g \geq 0 \ : \ \int \varphi g d\mu = a\}$$

BASIC TOOLS

The convex conjugate of γ,

$$\gamma^*(r) = \sup_{t>0} \, [rt - \gamma(t)]$$

is a nondecreasing convex function, finite and differentiable in $(-\infty, \gamma'(+\infty))$, and its derivative goes to $+\infty$ as $r \uparrow \gamma'(+\infty)$.
$\gamma^*(\gamma'(+\infty))$ may or may not be finite.

Denote by u the function on $\mathbb{R}$ equal to $(\gamma^*)'$ in $(-\infty, \gamma'(+\infty))$ and $+\infty$ outside.
Then $u(r) = 0$ if $r \leq \gamma'(0)$, and u is strictly increasing from 0 to $+\infty$ in the interval

$$(\gamma'(0), \gamma'(+\infty)).$$

Lemma 1. For $r < \gamma'(+\infty)$

$$\gamma'(u(r)) = \max\left[\gamma'(0), r\right] = r + |\gamma'(0) - r|_+$$
$$\gamma(u(r)) + \gamma^*(r) = ru(r).$$

For non-negative numbers t, s define

$$\Delta_\gamma(t.s) = \gamma(t) = [\gamma(s) + \gamma'(s)(t-s)]$$

(not meaningful for $s = 0$ if $\gamma'(0) = -\infty$; then we set $\Delta_\gamma(0,0) = 0, \Delta_\gamma(t,0) = +\infty$ if $t > 0$)

Bregman distance of nonnegative functions $g,\ h$ on X :

$$B_\gamma(g,h) = \int \Delta_\gamma(g,h)d\mu$$

Clearly, $B_\gamma(g,h) \geq 0$, equality iff $g = h\ [\mu]$

KEY IDENTITY

Denote:

$\mathcal{L}_a$: the family of nonnegative (measurable) functions g on X satisfying the constraints $\int g\varphi d\mu = a \qquad (a = (a_0, \ldots, s_d) \in \mathbb{R}^{1+d})$

$\mathcal{F}_\gamma$: the family of functions $f_\vartheta = u(\langle \vartheta, \varphi \rangle)$ with $\vartheta \in \mathbb{R}^{1+d}$ such that $\int \gamma^*(\langle \vartheta, \varphi \rangle) d\mu$ is finite, and $\langle \vartheta, \varphi \rangle < \gamma'(+\infty)$ $[\mu]$

Key identity: For $g \in \mathcal{L}_a$ and $f_\vartheta \in \mathcal{F}_\gamma$

$$J_\gamma(g) - \left[\langle \vartheta, a \rangle - \int \gamma^*(\langle \vartheta, \varphi \rangle) d\mu\right] =$$

$$= B_\gamma(g, f_\vartheta) + \int g|\gamma'(0) - \langle \vartheta, \varphi \rangle|_+ d\mu$$

Proof: Immediate, using Lemma 1.

Proposition: If $\mathcal{L}_a \cap \mathcal{F}_\gamma \neq \emptyset$, it consists of a single function $g = f_\vartheta$, this g minimizes $J_\gamma(g)$ subject to $g \in \mathcal{L}_a$, and ϑ maximizes $\langle \vartheta, a \rangle - \int \gamma^*(\langle \vartheta, \varphi \rangle) d\mu$; these minimum and maximum are equal. [But ϑ need not be unique, only f_ϑ is.]

Proof: Immediate from the key identity.

The family $\mathcal{F}_\gamma$ is the γ-analogue of an exponential family in the theory of Shannon entropy maximization. While the functions in $\mathcal{F}_\gamma$ need not be probability densities, in the case $\gamma(t) = t \log t$ they are exactly the constant multiples of the probability densities in $\mathcal{F}_\gamma$, which form an exponential family in the familiar statistical sense. For other γ however, no simple way is apparent to identify the probability densities in $\mathcal{F}_\gamma$.

Convex conjugate of $H_\gamma(a) = \inf_{g \in \mathcal{L}_a} J_\gamma(g)$:

$$H^*_\gamma(\vartheta) = \sup_{a \in \mathbb{R}^{1+d}} \left[\langle \vartheta, a \rangle - H_\gamma(a) \right], \quad \vartheta \in \mathbb{R}^{1+d}.$$

Lemma 2: If $dom(H_\gamma) \neq \emptyset$, thus there exists some g with $\int \gamma(g) d\mu < +\infty$ and $g\varphi_i$ integrable for $i = 0, \ldots, d$, then

$$H^*_\gamma(\vartheta) = \int \gamma^*(\langle \vartheta, \varphi \rangle) d\mu$$

Dual problem: find the dual value

$$H^{**}_\gamma(a) = \sup_{\vartheta \in \mathbb{R}^{1+d}} \left[\langle \vartheta, a \rangle - H^*_\gamma(\vartheta) \right],$$

and if it is finite, find $\vartheta \in \mathbb{R}^{1+d}$ that attains the maximum, if such ϑ exists (dual attainment). In the latter case, the function $f_\vartheta = u(\langle \vartheta, \varphi \rangle)$ will be called dual solution, rather than ϑ itself.

Lemma 3: for $\vartheta \in dom(H^*_\gamma)$, the directional derivative of H^*_γ at ϑ, in a direction τ, exists and equals $\int f_\vartheta \langle \tau, \varphi \rangle d\mu < +\infty$ whenever $\vartheta + t\tau \in dom(H^*_\gamma)$ for some $t > 0$. In particular, H^*_γ is differentiable in the interior of its essential domain, with the gradient equal to $\int f_\vartheta \varphi d\mu$.

Corollary: For $a \in \mathbb{R}^{1+d}$ with $H^{**}_\gamma(a)$ finite, a dual solution satisfies

$$\int f_\vartheta d\mu \leq a_0$$

Proof: Straightforward calculus. Differentiation within the integral is justified by monotone convergence.

Proposition 2: If $H_\gamma(a)$ is finite and a is in the relative interior of $dom(H_\gamma)$ then the primal and dual values are equal, and dual attainment holds. Moreover, the dual solution f_ϑ satisfies for each $g \in \mathcal{L}_a$

$$J_\gamma(g) = H_\gamma(a) + B_\gamma(g, f_\vartheta) +$$

$$+ \int g|\gamma'(0) - \langle \vartheta, \varphi \rangle|_+ d\mu.$$

If, in addition, $H^*_\gamma(\vartheta) = \int \gamma^*(\langle \vartheta, \varphi \rangle) d\mu$ is essentially smooth then the dual solution f_ϑ belongs to $\mathcal{L}_a$, hence it is a primal solution, too.

Proof: The first assertion is a general convex analysis result. The second assertion follows from it, by the key identity. The last assertion follows by Lemma 3, since if H^*_γ is essentially smooth, the maximizing ϑ has to be in the interior of $dom(H^*_\gamma)$.

Theorem 1: If $\gamma(0) = 0$ then

$$dom(H_\gamma) = \{ta : t \geq 0, a \in cc_\varphi(\mu)\}$$

If $\gamma(0) = +\infty$ (and $\mu(X) < +\infty$) then $dom(H_\gamma)$ is either empty, or

$$dom(H_\gamma) = \{ta : t > 0,\ a \in ri(cc_\varphi(\mu))\} = dom(H_\gamma^{**}).$$

In both cases, $dom(H_\gamma)$ is a cone that has a simple description in terms of $cc_\varphi(\mu)$.
In the case $\gamma(0) = +\infty$, no general criterion appears available to determine whether $dom(H_\gamma)$ is empty or not, but if nonempty, then Proposition 2 tells the story, as $dom(H_\gamma)$ is relatively open. In the sequel, we concentrate on the case $\gamma(0) = 0$.

Given nonzero $a \in dom(H_\gamma)$, equivalently $a_0 > 0, a_0^{-1}a \in cc_\varphi(\mu)$, let $F(a)$ denote the face of $cc_\varphi(\mu)$ whose relative interior contains $a_0^{-1}a$. Let ν denote the restiction of μ to $\tilde{F}(a) = \varphi^{-1}(e\ell(F/a))$.

Each $g \in \mathcal{L}_a$ vanishes outside $\tilde{F}(a)$

$$\Longrightarrow H_\gamma(a) = \inf_{g \in \mathcal{L}_a} \int \gamma(g) d\nu.$$

Proposition 2 applies to this a and the measure ν in the role of μ because $a_0^{-1}a$ is in the relative interior of the face $F(a)$ equal to $cc_\varphi(\nu)$.

It follows, provided $H_\gamma(a) > -\infty$, that the maximum of $\langle \vartheta, a\rangle - \int \gamma^*(\langle \vartheta, \varphi\rangle) d\nu$ is attained, and with a maximizing ϑ, the function

$$f_{F,\vartheta} = u\left(\langle \vartheta, \varphi\rangle\right) \mathbf{1}_{\tilde{F}(a)}$$

satisfies for all $g \in \mathcal{L}_a$

$$J_\gamma(g) =$$

$$= H_\gamma(a) + B_\gamma(g, f_{F,\vartheta}) + \int g|\gamma'(0) - \langle \vartheta, \varphi\rangle|_+ d\mu$$

Theorem 2: To every $a \neq 0$ with $H_\gamma(a)$ finite, there exists a (unique) generalized primal solution, of form $f_{F,\vartheta} = u(\langle \vartheta, \varphi\rangle)\mathbf{1}_{\tilde{F}(a)}$ and it satisfies the above identity.

Essential smoothness of $\int_{\tilde{F}(a)} \gamma^*(\langle \vartheta, \varphi\rangle) d\mu$ is a sufficient condition for primal attainment.