Proceedings ISIT 2008, July 6–11, 2008, Toronto, Canada, 2101–2105

On minimization of entropy functionals under moment constraints

Imre Csiszár
A. Rényi Institute of Mathematics
Hungarian Academy of Sciences
H-1364 Budapest, P.O.Box 127, Hungary
Email: csiszar@renyi.hu

František Matúš
Institute of Information Theory and Automation
Academy of Sciences of the Czech Republic
18208 Prague, P.O.Box 18, Czech Republic
Email: matus@utia.cas.cz

***Abstract*— Minimization problems for entropy-like integrals and Bregman distances subject to a finite number of moment constraints are addressed in a general setting. Analogues of the authors' previous results on information projections to families determined by linear constraints, and reverse information projections to exponential families, are established. No constraint qualification is assumed.**

I. Introduction

Let $(X, \mathcal{X}, \mu)$ be a measure space with μ nonzero, σ-finite, and let $\varphi = (\varphi_0, \ldots, \varphi_d)$ be a $(1+d)$-tuple of real-valued measurable functions on X, with φ_0 equal identically to 1. For a point $a = (a_0, \ldots, a_d)$ of $\mathbb{R}^{1+d}$, the set of nonnegative measurable functions g on X that satisfy the *moment constraints*

$$\int_X \varphi\, g\, d\mu = a \tag{1}$$

is denoted by $\mathcal{L}_{\mu,\varphi,a} = \mathcal{L}_a$.

Let γ be a strictly convex and differentiable function on the open interval $(0, +\infty)$, extended to $\mathbb{R}$ by $\gamma(0) = \lim_{t\downarrow 0} \gamma(t)$ and $\gamma(t) = +\infty$ for $t < 0$. Its convex conjugate γ^* is

$$\gamma^*(r) = \sup\nolimits_{t>0} \big[r\, t - \gamma(t) \big]\,, \qquad r \in \mathbb{R}\,.$$

The *entropy functional* $J_{\mu,\gamma} = J_\gamma$ is defined on the nonnegative measurable functions g on X by

$$J_\gamma(g) = \int_X \gamma(g)\, d\mu \tag{2}$$

if the integral exists, finite of not, and by $J_\gamma(g) = +\infty$ otherwise. Minimization of the entropy functional under moment constraints gives rise to the function $H_{\mu,\varphi,\gamma} = H_\gamma$ defined by

$$H_\gamma(a) = \inf\nolimits_{g\in\mathcal{L}_a} J_\gamma(g)\,, \qquad a \in \mathbb{R}^{1+d}\,. \tag{3}$$

Obviously, J_γ and H_γ are convex.

This minimization problem has attracted a lot of attention [2], [6], [14], [15]. For $\gamma(t) = t \ln t$ it reduces to the classical problem of maximizing Shannon (differential) entropy, respectively to minimizing I-divergence if μ is a finite measure. Other frequent choices are $\gamma(t) = -\ln t$ (Burg entropy) and $\gamma(t) = t^2$. The theory is intertwined with that of *Bregman distances* [3], see Section II, that have many applications on their own [13], [4], [11], [17], [12].

When $H_\gamma(a)$ is finite, general results of [5] on minimization of the entropy functional over any convex set imply existence of a function g_a, unique in the μ-a.e. sense, such that

$$J_\gamma(g) \geqslant H_\gamma(a) + B_\gamma(g, g_a)\,, \qquad g \in \mathcal{L}_a\,. \tag{4}$$

Here B_γ denotes Bregman distance. If the infimum in (3) is attained by some function $g \in \mathcal{L}_a$ then (4) implies $g = g_a$, μ-a.e. Otherwise, the function g_a is a *generalized minimizer* in the sense that each sequence g_n in $\mathcal{L}_a$ with $J_\gamma(g_n)$ approaching $H_\gamma(a)$ converges to g_a in Bregman distance.

More refined results are known only under additional standard assumptions, as the (primal) constraint qualification [2] that a strictly positive $g \in \mathcal{L}_a$ with $J_\gamma(g)$ finite exists. This work avoids such restrictions of generality via enhancing techniques of [8] developed for the special case of I-divergence minimization. The enhancements include a new concept of φ-*cone* of the measure μ, defined in Section II, and families of functions $\mathcal{F}_\gamma$ and $exn(\mathcal{F}_\gamma)$, defined in Section IV, that play the roles of exponential families and their extension in [8]. In contrast to I-divergence minimization, attention cannot be restricted here to probability densities, since for some convex functions γ it may happen that $\int_X g_a\, d\mu < 1$ even if $\mathcal{L}_a$ contains probability densities, thus $a_0 = 1$.

A crucial fact is that whenever the function H_γ is not identically $+\infty$, that is, if there exists a nonnegative function g with φg integrable and $J_\gamma(g) < +\infty$, the convex conjugate

$$H^*_\gamma(\vartheta) = \sup\nolimits_{a\in\mathbb{R}^{1+d}} \big[\langle \vartheta, a\rangle - H_\gamma(a) \big]\,, \qquad \vartheta \in \mathbb{R}^{1+d}\,,$$

of H_γ can be represented as the integral

$$H^*_\gamma(\vartheta) = \int_X \gamma^*(\langle \vartheta, \varphi\rangle)\, d\mu\,. \tag{5}$$

This identity follows in many instances from [18] but may be new in this generality. Its proof, as of later auxiliary results, is omitted. Then, the second conjugate of H_γ is given by

$$H^{**}_\gamma(a) = \sup\nolimits_{\vartheta\in\mathbb{R}^{1+d}} \big[\langle \vartheta, a\rangle - \int_X \gamma^*(\langle \vartheta, \varphi\rangle)\, d\mu \big]. \tag{6}$$

The optimization problems (3) and (6), and their values $H_\gamma(a)$ and $H^{**}_\gamma(a)$, are referred to as the *primal* and *dual*, respectively. When $H^{**}_\gamma(a)$ is finite, the nonnegative difference of primal and dual values is called the *duality gap*. By standard properties of convex functions, the gap equals zero, or vanishes, if and only if H_γ is lower semicontinuous at a.

Section II contains preliminaries. In Section III, conditions are provided for H_γ to be finite at $a \in \mathbb{R}^{1+d}$. Theorem 1 identifies the effective domain $dom(H_\gamma)$ of H_γ, thus the set of all $a \in \mathbb{R}^{1+d}$ with $H_\gamma(a) < +\infty$, in terms of the φ-cone of μ.

In Section IV, the inequality (4) is refined to an equality, and the generalized primal minimizer g_a is identified as a specific member of the family $exn(\mathcal{F}_\gamma)$, Theorem 2. The case when a is in the relative interior of $dom(H_\gamma)$ is settled first, for such a dual attainment and zero duality gap are known results that follow by standard arguments of convex analysis. For arbitrary $a \in dom(H_\gamma)$, a measurable set $\hat{F}(a) \subseteq X$ is constructed such that replacing X by $\hat{F}(a)$ the primal problem does not change while the dual problem becomes adjusted to application of results obtained in the previous case.

Section V addresses the dual problem. Theorem 3 shows for each $a \in dom(H_\gamma^{**})$ the existence of a function h_a in $exn(\mathcal{F}_\gamma)$ that satisfies an inequality analogous to ineq. (4), thus admits interpretation as the generalized dual maximizer. When the duality gap vanishes, the generalized primal and dual optimizers coincide.

In Section VI, Theorem 4 revisits the Pythagorean theorem for Bregman distances and details duality between the families of functions $\mathcal{L}_a$ and $exn(\mathcal{F}_\gamma)$. Unlike before, it is necessary to assume that γ is steep, thus $\gamma'(t) \downarrow -\infty$ when $t \downarrow 0$.

Remark 1. Minimization of entropy functionals under perhaps infinitely many moment constraints has been treated, without constraint qualification like here, in [14], [15], [16] via Orlicz space techniques. A comparison of results is not straightforward and is not attempted here. A differential geometric approach [1], [12], [17] provides a geometric intuition, for example on Pythagorean identities. In this paper, no such tools are needed for treating the problem effectively.

II. Preliminaries

2.1. The convex conjugate γ^* is clearly nondecreasing. Its value $\gamma^*(r)$ is finite for r less than, and $+\infty$ for r greater than $\gamma'(+\infty)$. Further, $\gamma^*(r) \downarrow -\gamma(0)$ as $r \downarrow -\infty$. Here and elsewhere, notations as $\gamma'(0)$, $\gamma'(+\infty)$, etc., are defined by limiting as $t \downarrow 0$ or $t \uparrow +\infty$, etc.

The strict convexity of γ implies that γ^* is essentially smooth [19, Theorem 26.3], in particular differentiable in $(-\infty, \gamma'(+\infty))$. Denote by u the function on $\mathbb{R}$ with $u(r)$ equal to $\gamma^{*\prime}(r)$ if $r < \gamma'(+\infty)$ and to $+\infty$ if $r \geqslant \gamma'(+\infty)$.

Lemma 1. *The function u equals 0 on $(-\infty, \gamma'(0)]$, and strictly increases in $(\gamma'(0), \gamma'(+\infty))$, mapping this interval onto $(0, +\infty)$.*

Lemma 2. *For $r < \gamma'(+\infty)$*

$$\gamma'(u(r)) = r + |\gamma'(0) - r|_+ \quad \text{and} \quad \gamma(u(r)) = ru(r) - \gamma^*(r)\,.$$

2.2. For $t \geqslant 0$ and $s > 0$ let

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

This formula defines also $\Delta_\gamma(t,0)$ when $\gamma'(0)$ is finite. Otherwise, let $\Delta_\gamma(t,0)$ equal to 0 or $+\infty$ according as $t = 0$ or $t > 0$. The strict convexity of γ implies that $\Delta_\gamma(t,s)$ is nonnegative, equal to zero if and only if $t = s$.

Lemma 3. *For $t \geqslant 0$ and $r < \gamma'(+\infty)$*

$$\gamma(t) + \gamma^*(r) = tr + \Delta_\gamma(t, u(r)) + t\,|\gamma'(0) - r|_+\,.$$

For nonnegative finite-valued measurable functions g, h on X the *Bregman distance* $B_{\mu,\gamma} = B_\gamma$ is defined by

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

It is nonnegative, equal to zero if and only if $g = h$, μ-a.e. For a sequence of functions g_n, either kind of convergence in Bregman distance, $B_\gamma(g_n, h) \to 0$ or $B_\gamma(h, g_n) \to 0$, implies convergence of g_n to h in measure on each set $A \subseteq \mathcal{X}$ of finite measure [5]. Hence, a subsequence of g_n converges to h, μ-a.e. For a condition on γ sufficient for convergence in Bregman distance to imply L_1-convergence see [5, Lemma 3].

Lemma 4. *For $g \in \mathcal{L}_a$ and $\vartheta \in \mathbb{R}^{1+d}$ such that $\langle\vartheta,\varphi\rangle$ is μ-a.e. less than $\gamma'(+\infty)$*

$$J_\gamma(g) + \int_X \gamma^*(\langle\vartheta,\varphi\rangle)\,d\mu = \langle\vartheta, a\rangle + B_\gamma(g, u(\langle\vartheta,\varphi\rangle)) + \int_X g\,|\gamma'(0) - \langle\vartheta,\varphi\rangle|_+\,d\mu\,.$$

The above key identity is proved by an integration in Lemma 3.

2.3. The *convex core* $cc(Q)$ of a σ-finite Borel measure Q on $\mathbb{R}^d$ is the intersection of those convex Borel sets in $\mathbb{R}^d$ whose complements have Q-measure zero. Though originally introduced only for finite measures [7], the concept and results have natural extensions to σ-finite ones [10], such as:

The closure of $cc(Q)$ equals the convex support $cs(Q)$, the smallest closed convex set whose complement has Q-measure zero. Every face F of $cc(Q)$ equals the convex core of the restriction of Q to $cl(F)$, the closure of F. For probability measures P dominated by Q whose means $\int_{\mathbb{R}^d} z\,P(dz) \in \mathbb{R}^d$ exist, these means exhaust $cc(Q)$, and exactly those of them belong to a face F of $cc(Q)$ for which $cs(P) \subseteq cl(F)$.

The relative interior $ri(A)$ of a set $A \subseteq \mathbb{R}^d$ is its interior in the relative topology of the affine hull of A.

2.4. Let μ_φ denote the φ-image of μ on $\mathbb{R}^{1+d}$. Since μ_φ is not necessarily σ-finite, the concept of convex core may not apply directly. Therefore, $cc_\varphi(\mu)$ is defined as equal to $cc(\nu_\varphi)$, for any finite measure ν equivalent to μ. The arbitrary choice of ν causes no ambiguity, and $cc_\varphi(\mu) = cc(\mu_\varphi)$ if μ_φ is σ-finite. The assumption $\varphi_0 = 1$ implies that $cc_\varphi(\mu)$ is contained in the hyperplane of points a with $a_0 = 1$.

The conical hull of the convex set $cc_\varphi(\mu)$

$$cn_\varphi(\mu) = \{ta \colon t \geqslant 0\,,\ a \in cc_\varphi(\mu)\}$$

is called the *φ-cone* of μ, and plays a key role in the sequel.

Lemma 5. *The set $\mathcal{L}_a$ is nonempty if and only if $a \in cn_\varphi(\mu)$.*

Lemma 6. *If $a \in cn_\varphi(\mu)$ then $\mathcal{L}_a$ contains a bounded function $g \geqslant 0$ that is positive only on a set of finite μ-measure.*

Lemma 7. *For $a \in \mathbb{R}^{1+d}$ with $a_0 = t > 0$ and $\frac{1}{t}\,a$ in a face F of $cc_\varphi(\mu)$, each $g \in \mathcal{L}_a$ vanishes μ-a.e. outside $\varphi^{-1}(cl(F))$.*

III. Finiteness of H_γ

This section identifies situations in which the function H_γ takes finite values or attains $\pm\infty$. Outside $cn_\varphi(\mu)$ this function is equal to $+\infty$ by Lemma 5.

Lemma 8. *If μ is a finite measure then the function H_γ never takes the value $-\infty$.*

The results of this section do depend on the standing assumption that φ_0 is a positive constant. The following example illustrates that Lemma 8 can fail otherwise.

Example 1. Let μ be Lebesgue measure on the unit interval and $\gamma(t) = -t^{3/4}$. The infimum of $\int_0^1 \gamma(g)\,dx$ subject to $g \geqslant 0$ and the moment constraint $\int_0^1 x\,g(x)\,dx = a$ equals $-\infty$ if $a > 0$. In fact, for $n \geqslant 1$ the functions

$$g_n(x) = \begin{cases} 2n^2 \cdot a\,, & x \leqslant \frac{1}{n}\,, \\ 0\,, & x > \frac{1}{n}\,, \end{cases}$$

satisfy the constraint, and $\int_0^1 \gamma(g_n)\,dx = -(2n^2 \cdot a)^{3/4}\frac{1}{n}$ goes to $-\infty$. The infimum is 0 if $a = 0$, and $+\infty$ if $a < 0$.

Lemma 9. *If μ is infinite and g nonnegative and integrable such that $\int_X \gamma(g)\,d\mu$ exists, finite or not, then $J_\gamma(g)$ equals $+\infty$ when $\gamma(0) > 0$ and $-\infty$ when $\gamma(0) < 0$.*

In the sequel, when μ is infinite it is assumed that $\gamma(0) = 0$, for otherwise $H_\gamma(a)$ does note take finite values, by Lemma 9. As the following example shows, this assumption does not exclude $H_\gamma(a) = -\infty$ for some $a \in \mathbb{R}^{1+d}$. Theorem 1 below and eq. (5) imply that such $a \in \mathbb{R}^{1+d}$ exists if and only if $\int_X \gamma^*(\langle\vartheta,\varphi\rangle)d\mu = +\infty$ for all $\vartheta \in \mathbb{R}^{1+d}$.

Example 2. For $X = \mathbb{R}^2$, let μ be the sum of the Lebesgue measure on the horizontal axis and the unit mass at $(0,1)$. Let $\varphi(x) = (1, x_1, x_2)$, $x = (x_1, x_2) \in X$, and $\gamma(t) = t\ln t$. Then $H_\gamma(a) = -\infty$ if $a = (a_0, a_1, a_2)$ satisfies $a_0 > a_2 \geqslant 0$, and $H_\gamma(a) = t\ln t$ for $a = (t, 0, t)$, $t \geqslant 0$. Otherwise, by Lemma 5, $H_\gamma(a) = +\infty$.

Theorem 1. *If μ is finite and $\gamma(0) < +\infty$ or μ is infinite and $\gamma(0) = 0$ then $dom(H_\gamma) = cn_\varphi(\mu)$. If μ is finite and $\gamma(0) = +\infty$ then either $dom(H_\gamma)$ is empty or both $dom(H_\gamma)$ and $dom(H_\gamma^{**})$ are equal to $ri(cn_\varphi(\mu))$.*

Proof. Lemma 5 implies $dom(H_\gamma) \subseteq cn_\varphi(\mu)$. By Lemma 6, if $a \in cn_\varphi(\mu)$ then $\mathcal{L}_a$ contains a bounded function g which is positive only on a set of finite μ-measure. Under either hypothesis for the first assertion of the theorem, $\gamma(g)$ is bounded and nonzero only on a set of finite μ-measure, thus μ-integrable. Hence, $cn_\varphi(\mu) \subseteq dom(H_\gamma)$.

Next, suppose that μ is finite, $\gamma(0) = +\infty$, $dom(H_\gamma) \neq \emptyset$. Fix $h \geqslant 0$ such that $\varphi\,h$ and $\gamma(h)$ are integrable. If $a \in cn_\varphi(\mu)$ then a bounded function $g \in \mathcal{L}_a$ exists by Lemma 6. For $\varepsilon > 0$ let h_ε equal h or $\varepsilon + g$ according as $h \leqslant \varepsilon$ or $h > \varepsilon$. It is not difficult to see that $\varphi\,h_\varepsilon$ and $\gamma(h_\varepsilon)$ are integrable and $\int_X \varphi\,h_\varepsilon\,d\mu$ converges to a as $\varepsilon \downarrow 0$. This proves that $cn_\varphi(\mu)$ is contained in the closure of $dom(H_\gamma)$. As $dom(H_\gamma)$ is a convex subset of $cn_\varphi(\mu)$, the inclusion $ri(cn_\varphi(\mu)) \subseteq dom(H_\gamma)$ follows.

Since $H_\gamma^{**} \leqslant H_\gamma$ the proof can be completed by showing that $H_\gamma^{**}(a) = +\infty$ if $a \notin ri(cn_\varphi(\mu))$. For any such a there exists a hyperplane V and its normal vector $\tau \neq 0$ such that V contains a and $ri(cn_\varphi(\mu))$ is contained in the open halfspace $V_< = \{z \in \mathbb{R}^{1+d} \colon \langle\tau, z\rangle < 0\}$. On account of the definition of the φ-cone of μ, the μ-measure of $\varphi^{-1}(V_<)$ is positive, and that of the complement of $\varphi^{-1}(V \cup V_<)$ is 0.

Since $dom(H_\gamma)$ is nonempty, eqs. (5) and (6) hold and imply together with Lemma 8 that $dom(H_\gamma^*) \neq \emptyset$. In addition, for any $\vartheta \in dom(H_\gamma^*)$ and $t \geqslant 0$

$$H^{**}(a) \geqslant \langle\vartheta + t\tau, a\rangle - \int_X \gamma^*(\langle\vartheta + t\tau, \varphi\rangle)\,d\mu.$$

Here $\langle\tau, a\rangle = 0$ since $a \in V$. On the set $\varphi^{-1}(V)$ the integrand is equal to $\gamma^*(\langle\vartheta,\varphi\rangle)$, while on $\varphi^{-1}(V_<)$, where $\langle\tau,\varphi\rangle < 0$, it decreases to $\gamma^*(-\infty) = -\gamma(0) = -\infty$ when $t \uparrow +\infty$, because γ^* is nondecreasing. By monotone convergence, the integral tends to $-\infty$, proving that $H_\gamma^{**}(a) = +\infty$. □

The following example shows that $dom(H_\gamma)$ can be empty when μ is finite and $\gamma(0) = +\infty$. A necessary and sufficient condition for emptiness remains elusive.

Example 3. Let $X = \mathbb{R}$, $\mu = \sum_{n\geqslant 1} \frac{1}{n^2}\,\delta_n$ where δ_n is the unit mass at n, $\varphi(x) = (1, x)$, $x \in \mathbb{R}$, and $\gamma(t) = e^{2/t}$, $t > 0$. If $J_\gamma(g)$ is finite for some function $g \geqslant 0$ then $\frac{1}{n^2}\,e^{2/g(n)}$ converges to 0, and thus $g(n) \geqslant \frac{1}{\ln n}$ eventually. Therefore $\int \varphi_1\,g\,d\mu$, equal to $\Sigma_{n\geqslant 1}\ n\,g(n)\,\frac{1}{n^2}$, diverges. This implies that g is not in the union of the families $\mathcal{L}_a$, thus H_γ is identically $+\infty$.

IV. Generalized minimizer of the primal problem

The primal and dual problems have been well understood under the hypothesis of the following lemma, equivalent to the primal constraint qualification. A proof invokes general results on convex functions [19, Theorems 23.4, 23.5] and combines them with (6) and Lemma 4.

Lemma 10. *If $a \in ri(dom(H_\gamma))$ and $H_\gamma(a) > -\infty$ then the duality gap $H_\gamma(a) - H_\gamma^{**}(a)$ equals zero and the supremum in eq. (6) is attained by some $\vartheta \in dom(H_\gamma^*)$. Even if the maximizer ϑ is not unique, the function $u(\langle\vartheta,\varphi\rangle)$ does not depend on its choice, equals the generalized primal minimizer g_a and its integral is at most a_0. For all $g \in \mathcal{L}_a$*

$$J_\gamma(g) = H_\gamma(a) + B_\gamma(g, g_a) + \int_X g\,|\gamma'(0) - \langle\vartheta,\varphi\rangle|_+\,d\mu. \quad (7)$$

Here, the function g_a satisfying ineq. (4) is obtained directly, not relying on the prior existence result, as a specific member of the family $\mathcal{F}_\gamma = \mathcal{F}_{\mu,\varphi,\gamma}$ of all functions $f_\vartheta = u(\langle\vartheta,\varphi\rangle)$ with $\vartheta \in dom(H_\gamma^*)$ and $\langle\vartheta,\varphi\rangle < \gamma'(+\infty)$, μ-a.e. The second condition, equivalent to the finiteness of f_ϑ, may or may not hold for all $\vartheta \in dom(H_\gamma^*)$, depending on γ. In the case $\gamma(t) = t\ln t$ the family $\mathcal{F}_\gamma$ consists of the positive multiples of the probability densities in the full exponential family $\mathcal{E}_{\mu,\varphi}$ based on μ with the sufficient statistic φ.

The *extension* $\mathit{exn}(\mathcal{F}_\gamma)$ of $\mathcal{F}_\gamma$ is defined as the union of families $\mathcal{F}_\gamma^F$ with F a face of $\mathit{cc}_\varphi(\mu)$ or the empty set. This family consists of the functions $f_{F,\vartheta} = u(\langle\vartheta,\varphi\rangle)\mathbf{1}_{\hat{F}}$ with ϑ in $\mathit{dom}(H^*_{\nu,\varphi,\gamma})$ and $\langle\vartheta,\varphi\rangle < \gamma'(+\infty)$, ν-a.e., where ν is equal to the restriction of μ to $\hat{F} = \varphi^{-1}(\mathit{cl}(F))$, and $\mathbf{1}_A$ denotes the indicator function of $A \subseteq X$. Note that $\mathcal{F}_\gamma^\emptyset$ contains only the function equal identically to zero.

For $\gamma(t) = t\ln t$, this extension consists of the nonnegative multiples of the densities in $\mathit{ext}(\mathcal{E}_{\mu,\varphi})$, the extension in the sense of [8].

The main result of this section, Theorem 2, does not resort to the constraint qualification. It identifies the generalized primal minimizer g_a as a member of $\mathit{exn}(\mathcal{F}_\gamma)$, for any $a \in \mathbb{R}^{1+d}$ with $H_\gamma(a)$ finite. In particular, H_γ is allowed to take the value $-\infty$ at some points different from a, as in Example 2 with $a = (t, 0, t)$, which is a contingency excluded by the hypothesis of Lemma 10. Note, however, that on account of Theorem 1 no enhancements of this lemma are needed when $\mathit{cc}_\varphi(\mu)$ is relatively open, or when $\gamma(0) = +\infty$.

By Theorem 1, if $a \in \mathit{dom}(H_\gamma)$ with $a_0 = t > 0$ then $\frac{1}{t}\,a$ belongs to $\mathit{cc}_\varphi(\mu)$. Denote by $F(a)$ the unique face F of this convex set with $\frac{1}{t}\,a \in \mathit{ri}(F)$, and extend this notation by $F(0) = \emptyset$. Further, let $\hat{F}(a) = \varphi^{-1}(\mathit{cl}(F(a)))$ and

$$C_a = \gamma(0) \cdot \mu(X \setminus \hat{F}(a)).$$

Theorem 2. *For every $a \in \mathbb{R}^{1+d}$ with $H_\gamma(a)$ finite,*

$$H_\gamma(a) = C_a + \max\nolimits_{\vartheta\in\mathbb{R}^{1+d}} \left[\langle\vartheta, a\rangle - \int_{\hat{F}(a)} \gamma^*(\langle\vartheta,\varphi\rangle)\, d\mu\right].$$

For any maximizer ϑ the function $f_{F(a),\vartheta}$ belongs to $\mathcal{F}_\gamma^{F(a)}$, equals g_a, and eq. (7) *holds for $g \in \mathcal{L}_a$.*

Proof. If a belongs to $\mathit{ri}(\mathit{dom}(H_\gamma))$ then $F(a) = \mathit{cc}_\varphi(\mu)$, the complement of $\hat{F}(a)$ has μ-measure zero, and the assertions follow from Lemma 10.

Otherwise, excluding the trivial case $a = 0$, let ν denote the restriction of μ to $\hat{F}(a)$. Then $\mathit{cc}_\varphi(\nu) = F(a)$, and as $\frac{1}{t}\,a$ belongs to the relative interior of this set, Theorem 1 implies that $a \in \mathit{ri}(\mathit{dom}(H_{\nu,\gamma}))$.

By Lemma 7, for $g \in \mathcal{L}_a$

$$J_{\mu,\gamma}(g) = C_a + \int_{\hat{F}(a)} \gamma(g)\, d\mu = C_a + J_{\nu,\gamma}(g)$$
$$H_{\mu,\gamma}(a) = C_a + H_{\nu,\gamma}(a)\,.$$

In particular, $H_{\nu,\gamma}(a)$ is finite. Thus, Lemma 10 applies to ν in the role of μ, and the assertions of Theorem 2 follow on account of the above two equations. □

Remark 2. The integral in Theorem 2, equal to $H^*_{\nu,\gamma}(\vartheta)$, is differentiable in the interior of $\mathit{dom}(H^*_{\nu,\gamma})$ and its gradient is equal to $\int_{\hat{F}(a)} \varphi\, u(\langle\vartheta,\varphi\rangle)\, d\mu$. When some ϑ in this interior attains the maximum in Theorem 2, the gradient at ϑ equals a. This means that g_a satisfies the moment constraints (1) thus it is a proper primal minimizer. A reasonably general and easily verifiable sufficient condition for primal attainment for all a in $\mathit{dom}(H_\gamma)$ is, however, not apparent: while $\mathit{dom}(H^*_\gamma) = \mathbb{R}^{1+d}$ suffices, the essential smoothness of H^*_γ does not, in general.

V. Generalized maximizer of the dual problem

In this section, H_γ is assumed not to take the value $-\infty$.

When in eq. (6) the maximum is attained for some ϑ, it is convenient to regard not this ϑ but the function $f_\vartheta = u(\langle\vartheta,\varphi\rangle)$ as dual maximizer. Then for $a \in \mathit{ri}(\mathit{dom}(H_\gamma))$ a unique dual maximizer exists, by Lemma 10. The next theorem shows that a generalized dual maximizer exists for all $a \in \mathit{dom}(H^{**}_\gamma)$.

Theorem 3. *For every $a \in \mathit{dom}(H^{**}_\gamma)$ there exists a unique nonnegative function h_a such that for $\vartheta \in \mathit{dom}(H^*_\gamma)$ satisfying $\langle\vartheta,\varphi\rangle < \gamma'(+\infty)$, μ-a.e.,*

$$\begin{aligned} H^{**}_\gamma(a) - \left[\langle\vartheta, a\rangle - \int_X \gamma^*(\langle\vartheta,\varphi\rangle)\, d\mu\right] \geqslant \\ B_\gamma(h_a, f_\vartheta) + \int_X h_a\, |\gamma'(0) - \langle\vartheta,\varphi\rangle|_+\, d\mu\,. \end{aligned} \tag{8}$$

The function h_a belongs to $\mathit{exn}(\mathcal{F}_\gamma)$. If the duality gap equals zero then $h_a = g_a$.

Remark 3. The supremum in (6) can be equivalently obtained over $\vartheta \in \mathit{dom}(H^*_\gamma)$ with μ-integrable f_ϑ. As μ-integrability implies $\langle\vartheta,\varphi\rangle < \gamma'(+\infty)$, μ-a.e., the latter condition does not restrict the quantification of ineq. (8) substantially.

Proof. For $a \in \mathit{dom}(H_\gamma)$, Lemma 4 is applied to a sequence g_n in $\mathcal{L}_a$ with $J_\gamma(g_n) \to H_\gamma(a)$ and $\vartheta \in \mathit{dom}(H^*_\gamma)$ with f_ϑ finite μ-a.e., and implies

$$\begin{aligned} H_\gamma(a) - \left[\langle\vartheta, a\rangle - \int_X \gamma^*(\langle\vartheta,\varphi\rangle)\, d\mu\right] \geqslant \\ \liminf_{n\to\infty}\ B_\gamma(g_n, f_\vartheta) + \int_X g_n\, |\gamma'(0) - \langle\vartheta,\varphi\rangle|_+\, d\mu\,. \end{aligned}$$

As $B_\gamma(g_n, g_a)$ goes to 0 by ineq. (4), some subsequence of g_n converges to g_a μ-a.e. Then, Fatou lemma implies that this $\liminf$ is at least the right-hand side of ineq. (8) with h_a replaced by g_a. This proves that in the case $H_\gamma(a) = H^{**}_\gamma(a)$, ineq. (8) holds with $h_a = g_a$. By Theorem 2, $h_a \in \mathit{exn}(\mathcal{F}_\gamma)$.

For $a \in \mathit{dom}(H^{**}_\gamma)$ with nonzero duality gap, the existence of h_a satisfying ineq. (8) can be proved by mimicking the proof sketched in [8] and detailed in [9, Theorem 2.1] for the special case of I-divergence. The proof of $h_a \in \mathit{exn}(\mathcal{F}_\gamma)$ relies upon Lemma 1. Details are omitted here. □

For the special case $\gamma(t) = t\ln t$ an explicit construction of the function h_a is available in [10]. It is believed that the techniques of that work can provide a constructive description of h_a, and an explicit formula for the duality gap, in general. The authors intend to return to these generalizations later.

The following example illustrates that, in general, the duality gap can be positive and g_a different from h_a even if μ is finite and φ bounded, which is impossible when $\gamma(t) = t\ln t$.

Example 4. On the unit square $X = [0,1]^2$, let μ be the sum of the Lebesgue measure λ and the unit masses at $(0, 1/3)$ and $(0, 2/3)$. Let $\varphi(x) = (1, x_1, x_2)$, $x \in \mathbb{R}^2$, and $\gamma(t) = -2t^{1/2}$ for $t > 0$ whence $\gamma^*(r) = -r^{-1}$ for $r < 0$.

For $a = (1, 0, 1/3)$ it is easy to see that $H_\gamma(a) = J_\gamma(g_a)$ where g_a is equal to 1 at $(0, 1/3)$ and to 0 otherwise. Thus, $H_\gamma(a) = -2$. By (5), for $\vartheta = (\vartheta_0, \vartheta_1, \vartheta_2) \in \mathbb{R}^3$

$$H^*_\gamma(\vartheta) = \int_X \gamma^*(\vartheta_0 + \vartheta_1 x_1 + \vartheta_2 x_2)\, \mu(dx)$$

is finite only if ϑ_0, $\vartheta_0+\vartheta_1$, $\vartheta_0+\vartheta_2$ and $\vartheta_0+\vartheta_1+\vartheta_2$ are negative, in which case $-H^*_\gamma(\vartheta)$ equals

$$(\vartheta_0+\tfrac{1}{3}\vartheta_2)^{-1}+(\vartheta_0+\tfrac{2}{3}\vartheta_2)^{-1}+\int_X(\vartheta_0+\vartheta_1x_1+\vartheta_2x_2)^{-1}\lambda(dx)\,.$$

By limiting as $\vartheta_1\downarrow-\infty$, $H^{**}_\gamma(a)$ is equal to

$$\begin{aligned}\sup\nolimits_\vartheta\left[\vartheta_0+\tfrac{1}{3}\vartheta_2-H^*_\gamma(\vartheta)\right]&=\sup\nolimits_{\vartheta_0<0,\ \vartheta_0+\vartheta_2<0}\\&\left[\vartheta_0+\tfrac{1}{3}\vartheta_2+(\vartheta_0+\tfrac{1}{3}\vartheta_2)^{-1}+(\vartheta_0+\tfrac{2}{3}\vartheta_2)^{-1}\right]\\=\sup\nolimits_{t,r<0}\left[2t+r+\tfrac{1}{2t+r}+\tfrac{1}{t+r}\right]&=-2\sqrt{3}\,.\end{aligned}$$

It follows that the duality gap $H_\gamma(a)-H^{**}_\gamma(a)$ is $2(\sqrt{3}-1)$. It is also easy to identify a maximizing sequence and to find explicitly the function h_a, different from g_a.

VI. Pythagorean theorem

The refinement eq. (7) of ineq. (4) simplifies for $a\in\mathbb{R}^{1+d}$ with $H_\gamma(a)$ finite and γ steep to

$$J_\gamma(g)=H_\gamma(a)+B_\gamma(g,g_a)\,,\qquad g\in\mathcal{L}_a\,.$$

This and Lemma 4 immediately imply the *Pythagorean identity* for Bregman distances, eq. (9). One point of this section is that this identity, and related results well-known in regular cases, do require γ to be steep, thus $\gamma'(0)=-\infty$.

The family of nonnegative functions h admitting the approximation $B_\gamma(g_n,h)\to0$ by sequences g_n in $\mathcal{L}_a$ is denoted by $cl_\gamma(\mathcal{L}_a)$. This is an analogue of the I-closure in [8].

Lemma 11. *If γ is steep, $a\in ri(dom(H_\gamma))$ and $H_\gamma(a)>-\infty$, then g_a is the unique element of the intersection of $cl_\gamma(\mathcal{L}_a)$ and $\mathcal{F}_\gamma$. For all $g\in\mathcal{L}_a$ and $h\in\mathcal{F}_\gamma$*

$$B_\gamma(g,h)=B_\gamma(g,g_a)+B_\gamma(\mathcal{L}_a,h)\qquad(9)$$

where $B_\gamma(\mathcal{L}_a,p)$ denotes the infimum of $B_\gamma(g,h)$ subject to $g\in\mathcal{L}_a$, and g_a is the unique function with the last property.

The following two examples show that the steepness assumption on γ cannot be avoided in Lemma 11, neither for the intersection to be a singleton nor for (9).

Example 5. Let $X=\{1,2,3\}$ with μ the counting measure, and identify functions on X with points in $\mathbb{R}^3$. Let $d=1$, $\varphi_1=(1,-1,0)$, then for $a=(a_0,a_1)\in\mathbb{R}^2$

$$\mathcal{L}_a=\{(\tfrac{t+a_1}{2},\tfrac{t-a_1}{2},a_0-t)\colon a_0\geqslant t\geqslant|a_1|\}.$$

Let $\gamma(t)=\frac{t^2}{2}$, then $u(r)=|r|_+$, and

$$\mathcal{F}_\gamma=\{(|\vartheta_0+\vartheta_1|_+\,,|\vartheta_0-\vartheta_1|_+\,,|\vartheta_0|_+)\,:\,(\vartheta_0,\vartheta_1)\in\mathbb{R}^2\,\}$$

is the union of three two-dimensional cones in $\mathbb{R}^3$. Obviously, eq. (9) does not hold, neither does its weakened version with the equality replaced by an inequality.

Example 6. Let μ have the density $\frac{2}{x^3}\,\mathbf{1}_{(1,+\infty)}$ w.r.t. Lebesgue measure on $\mathbb{R}$, $\varphi(x)=(1,x)$, $x\in\mathbb{R}$, and $\gamma(t)=\frac{t^2}{2}$. Then $\mathcal{F}_\gamma$ consists of the functions $f_\vartheta(x)=|\vartheta_0+\vartheta_1x|_+$ with $\vartheta_1\leqslant0$, but this function vanishes μ-a.e. if $\vartheta_0\leqslant-\vartheta_1$. If $\vartheta_0>-\vartheta_1\geqslant0$ then $\int_X\varphi\,f_\vartheta\,d\mu$ is equal to

$$a=(a_0,a_1)=\vartheta_0\Big(\big[(1+\tfrac{\vartheta_1}{\vartheta_0}\big]^2,\,2\big[1+\tfrac{\vartheta_1}{\vartheta_0}-\tfrac{\vartheta_1}{\vartheta_0}\ln\tfrac{-\vartheta_1}{\vartheta_0}\big]\Big)$$

These points a exhaust the cone given by $2a_0\geqslant a_1>a_0>0$ while $a\neq0$ belongs to $dom(H_\gamma)$ whenever $a_1>a_0>0$. It is possible to prove, but details are omitted here, that if $a\neq0$, $g\in\mathcal{L}_a$ and g^2 is μ-integrable then g belongs to $cl_\gamma(\mathcal{L}_{a+(0,\varepsilon)})$ for any $\varepsilon>0$. It follows that $cl_\gamma(\mathcal{L}_{(1,2)})$ contains a continuum of functions $f_\vartheta\in\mathcal{F}_\gamma$.

Finally, a strengthening of Lemma 11 is stated without proof.

Theorem 4. *If γ is steep and H_γ is finite at $a\in\mathbb{R}^{1+d}$ then g_a is the unique element of $cl_\gamma(\mathcal{L}_a)\cap exn(\mathcal{F}_\gamma)$, and* (9) *holds for all $g\in\mathcal{L}_a$ and $h\in exn(\mathcal{F}_\gamma)$. Moreover, g_a is the unique function with the last property.*

Acknowledgment

This work was supported by Hungarian National Foundation for Scientific Research, Grant T046376, by Grant Agency of Academy of Sciences of the Czech Republic, Grant IAA 100750603, and by Grant Agency of the Czech Republic, Grant 201/08/0539.

References

[1] Amari S. and Nagaoka H., *Methods of Information Geometry.* Translations of Mathematical Monographs, Vol. 191, Oxford Univ. Press, 2000.

[2] Borwein, J.M. and Lewis, A.S., Partially-finite programming in L_1 and the existence of maximum entropy estimates. *SIAM J. Optimization* **3** (1993) 248–267.

[3] Bregman, L.M., The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathrematical Physics* **7** (1967) 200-217.

[4] Censor, Y. and Zenios, S.A., *Parallel Optimization.* Oxford University Press, New York 1997.

[5] Csiszár, I., Generalized projections for non-negative functions. *Acta Math. Hungar.* **68 (1–2)** (1995) 161–185.

[6] Csiszár, I., Gamboa F. and Gassiat E., MEM pixel correlated solutions for generalized moment and interpolation problems. *IEEE Trans. Inform. Theory* **45** (1999) 2253–2270.

[7] Csiszár, I. and Matúš, F., Convex cores of measures on $\mathbb{R}^d$. *Studia Sci. Math. Hungar.* **38** (2001) 177–190.

[8] Csiszár, I. and Matúš, F., Information projections revisited. *IEEE Trans. Inform. Theory* **49** (2003) 1474–1490.

[9] Csiszár, I. and Matúš, F., Generalized maximum likelihood estimates for infinite dimensional exponential families. *Proceedings Prague Stochastics'06*, Prague, Czech Republic, 2006, 288–297.

[10] Csiszár, I. and Matúš, F., Generalized maximum likelihood estimates for exponential families. *Probability Theory and Related Fields* **141** (2008) 213–246.

[11] Dawid, A.P. and Grünwald P.D., Game theory, maximum entropy, minimum discrepancy, and robust Bayesian decision theory. *Ann. Statist.* **32** (2004), 1367-1433.

[12] Eguchi S., Information geometry and statistical pattern recognition. *Sugaku Expositions, Amer. Math. Soc.*, **19** (2006) 197–216.

[13] Jones, L. and Byrne, C., General entropy criteria for inverse problems with application to data compression, pattern classification and cluster analysis. *IEEE Trans. Inform. Theory* **36** (1990) 23–30.

[14] Léonard, C., Minimizers of energy functionals. *Acta Math. Hungar.* **93** (2001) 281-325.

[15] Léonard, C., Minimizers of energy functionals under not very integrable constraints. *J. Convex Anal.* **10** (2003) 63-68.

[16] Léonard, C., Minimization of entropy functionals. (arXiv:0710.1462)

[17] Murata N., Takenouchi T., Kanamori T. and Eguchi S., Information geometry of U-Boost and Bregman divergence. *Neural Computation* **16** (2004) 1437–1481.

[18] Rockafellar, R.T., Integrals which are convex functionals. *Pacific Journal of Mathematics* **24** (1968) 525–539.

[19] Rockafellar, R.T., *Convex Analysis.* Princeton University Press, Princeton 1970.