Proceedings ITW 2009, June 10-12, 2009, Volos, Greece, 96–100

On minimization of multivariate entropy functionals

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*— The problem of minimizing convex integral functionals subject to moment-like constraints is treated in a general setting when the underlying convex function may be multivariate, perhaps not strictly convex or differentiable. The results are applied to the minimization of f-divergences simultaneously in both variables.**

I. Introduction

Let μ be a nonzero and σ-finite measure on a measurable space $(Z,\mathcal{Z})$, and γ a proper closed convex function on $\mathbb{R}^e$. The function is not necessarily strictly convex or differentiable.

This work studies minimization of the *integral functional*

$$J_\gamma(g) \triangleq \int_Z \gamma(g(z))\,\mu(dz) \tag{1}$$

over the measurable functions $g\colon Z\to\mathbb{R}^e$ satisfying the linear moment-like constraints

$$\int_Z \varphi g\,d\mu = a\,. \tag{2}$$

Here, $a\in\mathbb{R}^d$, φ is a measurable mapping on Z with values in $\mathbb{R}^{d\times e}$, regarded as $d\times e$ matrices, and φg denotes the $\mathbb{R}^d$-valued function obtained by matrix multiplication, regarding the values of g as column vectors.

A. Previous work

The univariate case, $e=1$, has been extensively studied, see [2], [6], [7], [12], [13], etc. The instance $\gamma(t)=t\log t$ is of main interest for information theory. In the multivariate case, the minimization of γ itself has been much studied as well, intertwined with the theory of Bregman [3] distances on $\mathbb{R}^e$. The similar problem for the functionals (1), or even more general functionals based on normal integrands, seems hardly been tackled within information theory or statistics, references in [15, p. 681] point elsewhere.

For $e=2$, a class of functionals (1) of major statistical relevance arises with γ defined by

$$\gamma(t,s)=\begin{cases} sf\left(\frac{t}{s}\right), & t\geqslant 0\,,\ s>0\,,\\ tf'(+\infty)\,, & t\geqslant 0\,,\ s=0\,,\\ +\infty\,, & \text{otherwise.}\end{cases} \tag{3}$$

Here, f is a closed convex function on $\mathbb{R}$ that is finite on the positive axis and $+\infty$ on the negative one. The corresponding integral functional

$$D_f(g_1,g_2)\triangleq J_\gamma(g)=\int_Z \gamma(g_1,g_2)\,d\mu$$

applies to pairs $g=(g_1,g_2)$ of real functions on Z and is known under the name *f-divergence* [5]. Kullback-Leibler I-divergence results from the choice $f(t)=t\log t-t+1$. Its minimization simultaneously in both variables is the substance of the EM-algorithm, see [9].

Analogues of the f-divergences were defined also for $e>2$, as dissimilarity measures [10]. The instance of the functional J_γ with $\gamma(x)=-\prod_{i=1}^e x_i^{\alpha_i}$, $x=(x_1,\dots,x_e)\in\mathbb{R}^e$, where the exponents are positive and sum to 1, plays an important role in the theory of statistical experiments [11].

B. Outline of the results

In the recent work [8] the authors addressed the minimization of the functional (1) in the univariate case, assuming differentiability and strict convexity of γ. The goal was to dispense with constraint qualifications usually imposed in the literature. This approach is extended in Section III to the multivariate case, but here the focus is on dropping the regularity conditions on γ. In particular, for non-differentiable γ a new definition of Bregman distances is suggested in Section II. Also, two errors in previous papers of the first author are corrected in Remark 7 of Section III.

Section IV is devoted to minimization of the functionals (1) with $e=2$ and γ defined by eq. (3), thus to minimization of the f-divergences simultaneously in both variables. Note that this function γ is not strictly convex, and thus disposal of this assumption on γ is necessary.

II. Preliminaries

The *subdifferential* $\partial\gamma(x)$ of γ at $x\in\mathbb{R}^e$ is a closed convex set of the subgradients of γ at x [14, Section 23]. The set of all x's with $\partial\gamma(x)$ nonempty is denoted by $\mathit{dom}(\partial\gamma)$. It is contained in the set $\mathit{dom}(\gamma)$ of all x's with $\gamma(x)$ finite. The two sets have the same relative interior [14, Theorem 23.4].

The convex conjugate γ^* of γ is given by

$$\gamma^*(\vartheta)=\sup\nolimits_{x\in\mathbb{R}^e}\left[\langle\vartheta,x\rangle-\gamma(x)\right], \qquad \vartheta\in\mathbb{R}^e\,.$$

A. Bregman distances

The Bregman distance of vectors $x,y\in\mathbb{R}^e$, induced by a convex function γ, is typically defined when γ is differentiable in the interior of $\mathit{dom}(\gamma)$, and y is in that interior. More generally, for $x,y\in\mathit{dom}(\gamma)$ let

$$\Delta_\gamma(x,y)=\gamma(x)-\gamma(y)-\sup\nolimits_{\vartheta\in\partial\gamma(y)}\langle\vartheta,x-y\rangle$$

when $y \in \mathit{dom}(\partial\gamma)$; otherwise let $\varDelta_\gamma(x,y)$ equal 0 or $+\infty$ according to $x = y$ or not. When x or y is out of $\mathit{dom}(\gamma)$ let $\varDelta_\gamma(x,y) = +\infty$. The convexity of γ implies $\varDelta_\gamma(x,y) \geqslant 0$. This inequality is strict if γ is strictly convex on the segment with different endpoints x and y in $\mathit{dom}(\gamma)$. Thus, $\varDelta_\gamma$ is a pseudodistance on $\mathit{dom}(\gamma)$, which is a distance (but generally not a metric) if γ is strictly convex on its domain.

For x, y, θ in $\mathbb{R}^e$ let

$$\varUpsilon_\gamma(x,y,\theta) = \sup_{\vartheta \in \partial\gamma(y)} \langle \vartheta - \theta, x - y \rangle$$

when θ belongs to the subdifferential; otherwise let $\varUpsilon_\gamma(x,y,\theta)$ equal $+\infty$. This nonnegative quantity measures lack of differentiability of the function γ at the point y in the direction $x-y$ w.r.t. the subgradient θ.

Lemma 1. *For $x \in \mathbb{R}^e$, $y \in \mathit{dom}(\partial\gamma)$ and $\theta \in \partial\gamma(y)$*

$$\gamma(x) + \gamma^*(\theta) = \langle \theta, x \rangle + \varDelta_\gamma(x,y) + \varUpsilon_\gamma(x,y,\theta) \qquad (4)$$

Proof. If $x \notin \mathit{dom}(\gamma)$ then both sides of eq. (4) equal $+\infty$. Otherwise, since γ is proper and convex, $\theta \in \partial\gamma(y)$ implies that $\gamma(y) + \gamma^*(\theta) = \langle \theta, y \rangle$ [14, Theorem 23.5]. This combines with the definitions of $\varDelta_\gamma$ and $\varUpsilon_\gamma$ to arrive at (4). ☐

The *Bregman pseudodistance* of measurable functions g, h on Z with values in $\mathbb{R}^e$ is defined by

$$B_\gamma(g,h) \triangleq \int_Z \varDelta_\gamma(g(z), h(z))\, \mu(dz)\,. \qquad (5)$$

The inequality $B_\gamma(g,h) \geqslant 0$ always holds, and for γ strictly convex the equality does only if $g, h \in \mathit{dom}(\gamma)$ and $g = h$, μ-a.e. If additionally $\tau \colon Z \to \mathbb{R}^e$ is measurable let

$$L_\gamma(g,h,\tau) \triangleq \int_Z \varUpsilon_\gamma(g(z), h(z), \tau(z))\, \mu(dz)\,.$$

The above integrands are measurable, omitting details.

B. The value function and its conjugate

A value $J_\gamma(g)$ of the functional J_γ is well defined by eq. (1) if the integral exists, finite or not; otherwise it set equal to $+\infty$. In the sequel, the existence of a measurable function g such that $J_\gamma(g) < +\infty$ and φg is μ-integrable is referred to as the *global assumption*.

For $a \in \mathbb{R}^d$ let the affine set of functions g satisfying eq. (2) be denoted by $\mathcal{G}_a$, and the convex set of functions $g \in \mathcal{G}_a$ with values in $\mathit{dom}(\gamma)$, μ-a.e., by $\mathcal{L}_a$.

Minimization of the functional J_γ under the constraints (2) gives rise to the *value function* H_γ given by

$$H_\gamma(a) \triangleq \inf_{g \in \mathcal{G}_a} J_\gamma(g) = \inf_{g \in \mathcal{L}_a} J_\gamma(g)\,, \quad a \in \mathbb{R}^d\,. \qquad (6)$$

This function is convex but not necessarily proper or closed. The global assumption means that H_γ is not identically $+\infty$. In particular, it may take the value $-\infty$.

Lemma 2. *Under the global assumption,*

$$H^*_\gamma(\vartheta) = \int_Z \gamma^*(\vartheta\varphi)\, d\mu\,, \qquad \vartheta \in \mathbb{R}^d\,,$$

where the integral is finite or $+\infty$.

Here, $\vartheta\varphi$ denotes the $\mathbb{R}^e$-valued function obtained by matrix multiplication, regarding ϑ as a row vector.

Proof. By definition,

$$H^*_\gamma(\vartheta) = -\inf_{a \in \mathbb{R}^d} \big[-\langle \vartheta, a \rangle + \inf_{g \in \mathcal{G}_a} J_\gamma(g) \big]\,.$$

The infimum rewrites to

$$\inf_{a \in \mathbb{R}^d} \inf_{g \in \mathcal{G}_a} \int_Z \big[-\langle \vartheta, \varphi g(z) \rangle + \gamma(g(z)) \big] \mu(dz)\,,$$

and thus to

$$\inf_{g \in \mathcal{G}} \int_Z \big[-\langle \vartheta\varphi(z), g(z) \rangle + \gamma(g(z)) \big] \mu(dz)$$

where $\mathcal{G}$ denotes the space of functions g with μ-integrable φg. If the above integrals do not exist they are set equal to $+\infty$ as in the definition of J_γ. It suffices to prove that the above infimum equals

$$\int_Z \inf_{x \in \mathbb{R}^e} \big[-\langle \vartheta\varphi(z), x \rangle + \gamma(x) \big] \mu(dz)\,.$$

If φ is μ-integrable then the linear space $\mathcal{G}$ is decomposable in the sense of [15, Definition 14.59] whence the infimum and integral can be exchanged by [15, Theorem 14.60]. Otherwise, $\mathcal{G}$ need not be decomposable, but it is possible to modify [15, Theorem 14.60] to cover also this case, omitting details. ☐

III. Main results

For a measurable function τ on Z with values in $\mathit{dom}(\partial\gamma^*)$, μ-a.e., the set-valued mapping $z \mapsto \partial\gamma^*(\tau(z))$, $z \in Z$, is measurable in the sense of [15, Definition 14.1], using [15, Theorem 14.56]. Hence, it admits a Castaing representation [15, Theorem 14.5], and consequently a *measurable selection*. Thus, there exists a $\mathcal{Z}$-measurable function h such that $h(z)$ belongs to $\partial\gamma^*(\tau(z))$, μ-a.e. In the following theorem this conclusion is applied to τ of the form $\vartheta\varphi$.

Theorem 1. *Under the global assumption, for $\vartheta \in \mathit{dom}(H^*_\gamma)$ with $\vartheta\varphi$ taking values in $\mathit{dom}(\partial\gamma^*)$, μ-a.e., and a measurable selection h from $z \mapsto \partial\gamma^*(\vartheta\varphi(z))$ the equality*

$$J_\gamma(g) + H^*_\gamma(\vartheta) = \langle \vartheta, a \rangle + B_\gamma(g,h) + L_\gamma(g,h,\vartheta\varphi)$$

holds for all $a \in \mathbb{R}^d$ and functions $g \in \mathcal{G}_a$.

Proof. For $z \in Z$ let $x = g(z)$, $\theta = \vartheta\varphi(z)$, and $y = h(z)$. Since γ is closed $y \in \partial\gamma^*(\theta)$ is equivalent to $\theta \in \partial\gamma(y)$ by [14, Theorem 23.5]. Then, the assumptions on ϑ, g and h imply that Lemma 1 applies and eq. (4) holds with the above substitutions μ-a.e. The assertion follows by integration w.r.t. μ using the following arguments: *(i)* the integral of $\gamma^*(\vartheta\varphi)$ is finite and equals $H^*_\gamma(\vartheta)$ due to Lemma 2 and the assumption $\vartheta \in \mathit{dom}(H^*_\gamma)$, *(ii)* the integral of $\langle \vartheta\varphi, g \rangle$ equals $\langle \vartheta, a \rangle$ because $g \in \mathcal{G}_a$, and *(iii)* the quantities $\varDelta_\gamma$ and $\varUpsilon_\gamma$ are nonnegative. ☐

Remark 1. If in Theorem 1 a selection h belongs to $\mathcal{L}_a$ then the infima in (6) are finite and attained by $g = h$.

Remark 2. If γ is strictly convex then γ^* is essentially smooth [14, Theorem 26.3], thus $\mathit{dom}(\gamma^*)$ has nonempty interior where γ^* is differentiable and $\partial\gamma^*$ is empty otherwise. Hence, in this case $h = \nabla\gamma^*(\vartheta\varphi)$ is the unique selection in Theorem 1. On the other hand, if γ is essentially smooth then the term $L_\gamma(g,h,\vartheta\varphi)$ in Theorem 1 vanishes. In fact, by the above proof, $\vartheta\varphi \in \partial\gamma(h)$ whence $\varUpsilon_\gamma(g,h,\vartheta\varphi) = 0$, μ-a.e.

Remark 3. The integration in the above proof reveals also that in Theorem 1 the integral defining $J_\gamma(g)$ exists, finite or $+\infty$.

The minimization of $J_\gamma(g)$ over $g \in \mathcal{G}_a$ will be investigated via the second conjugate of the value function. Due to Lemma 2, under the global assumption

$$H_\gamma^{**}(a) = \sup_{\vartheta \in \mathbb{R}^d} \left[\langle \vartheta, a \rangle - \int_Z \gamma^*(\vartheta\varphi)\, d\mu \right], \quad a \in \mathbb{R}^d. \quad (7)$$

Theorem 2. *If a belongs to $ri(dom(H_\gamma))$ and $H_\gamma(a)$ is finite then the supremum in* (7) *is attained and coincides with $H_\gamma(a)$. If, additionally, a maximizer ϑ in* (7) *exists such that the values of $\vartheta\varphi$ belong μ-a.e. to $dom(\partial\gamma^*)$ then*

$$J_\gamma(g) = H_\gamma(a) + B_\gamma(g,h) + L_\gamma(g,h,\vartheta\varphi), \qquad g \in \mathcal{G}_a, \quad (8)$$

for any measurable selection h from $z \mapsto \partial\gamma^*(\vartheta\varphi(z))$.

Proof. By the assumptions on a, the value function is proper [14, Theorems 7.2], and the first implication follows from [14, Theorems 23.4, 23.5]. For any maximizer ϑ, the integral in (7) is finite whence $\vartheta\varphi$ belongs to $dom(\gamma^*)$, μ-a.e. Under the stronger assumption on $\vartheta\varphi$, the second assertion follows from Theorem 1 and Lemma 2, noting that the assumptions on a imply the global one. $\square$

When Theorem 2 applies the infima in eq. (6) coincide with the supremum in (7), which is an attained maximization problem without constraints in a finite dimension.

Remark 4. In the second part of Theorem 2, if $J_\gamma(g)$ equals the finite infimum $H_\gamma(a)$ for some $g \in \mathcal{G}_a$ then $g \in \mathcal{L}_a$ and $B_\gamma(g,h)$ vanishes for any of the selections h. In particular, if γ is strictly convex then g is equal to the unique selection $\nabla\gamma^*(\vartheta\varphi)$ by Remark 2. Even if no such minimizer g exists, $B_\gamma(g_n,h) \to 0$ for every minimizing sequence $g_n \in \mathcal{G}_a$ in (6), thus when $J_\gamma(g_n) \to H_\gamma(a)$. Such a function h is referred to as a *generalized minimizer* of the problem (6). When γ is not strictly convex, it is not necessarily unique.

Remark 5. The first hypothesis of Theorem 2 is of a very common kind, known as the (primal) constraint qualification. For the univariate case and a certain class of functions γ, an explicit characterization of $dom(H_\gamma)$ and its relative interior is available in [8, Theorem 1]. For that case, the constraint qualification can be dispensed with [8, Theorem 2]. Extensions of these results to the present generality remain unclear.

Remark 6. The hypothesis on ϑ attaining the maximum in (7) is trivially satisfied when $dom(\gamma^*) = dom(\partial\gamma^*)$, in particular when $dom(\gamma^*)$ is relatively open. In the framework of [8], a maximizing ϑ always satisfies this hypothesis whenever one of the coordinates of φ is the constant function 1, which is an assumption commonly adopted also elsewhere. If that assumption were dropped, Theorem 2 could fail even in that framework, as demonstrated below.

Example 1. Let $Z = \{-1,0,1\}$ be endowed with the counting measure, $e = 1$ and $\gamma(x) = 1/2x$ for $x > 0$ while $\gamma(x) = +\infty$ otherwise. Thus, $\gamma^*(\vartheta) = -\sqrt{-\vartheta}$ for $\vartheta \leqslant 0$ and $\gamma^*(\vartheta) = +\infty$ otherwise. Let $d = 1$ and $\varphi(z) = z$. Then, representing a function g on Z by a triple $(r,s,t) \in \mathbb{R}^3$,

$$H_\gamma(a) = \inf\{\tfrac{1}{2r} + \tfrac{1}{2s} + \tfrac{1}{2t} : r,s,t > 0,\ t - r = a\} = 0, \ a \in \mathbb{R}.$$

It follows that $\vartheta = 0$ is the unique element of $dom(H_\gamma^*)$. Then, $\vartheta\varphi$ identically equals 0, the endpoint of $dom(\gamma^*) = (-\infty, 0]$ where $\partial\gamma^*(0) = \emptyset$. Hence, no $\vartheta \in dom(H_\gamma^*)$ satisfies the additional hypothesis of Theorem 2. Moreover, each minimizing sequence g_n has the form $(r_n, s_n, r_n + a)$ with $r_n, s_n \to +\infty$ arbitrarily. Therefore, for g represented by $r,s,t > 0$

$$B_\gamma(g_n, g) = \frac{(r_n - r)^2}{2r_n r} + \frac{(s_n - s)^2}{2s_n s} + \frac{(r_n + a - t)^2}{2(r_n + a)t} \to +\infty$$

which implies that no generalized minimizer exists. In particular, no function h renders the Pythagorean-like inequality $J_\gamma(g) \geqslant H_\gamma(a) + B_\gamma(g,h)$ true for all $g \in \mathcal{L}_a$.

Remark 7. This example reveals two errors in previous papers of the first author. Nonexistence of the generalized minimizer in Example 1 contradicts [6, Theorem 1(c)], stating the Pythagorean-like inequality. Failure of a maximizer in (7) to satisfy the hypothesis in Theorem 2 contradicts the second statement of [7, Theorem II.2]. Regretting these errors, the first author notes that the proof of the existence of a generalized minimizer in [6] is valid under the erroneously omitted assumption that members of minimizing sequences have bounded Bregman distances from some fixed function.

IV. Minimization of f-divergence

In this section, $e = 2$ and γ is defined by eq. (3). For simplicity we assume that f is strictly convex, differentiable on the positive axis and $f'(+\infty) = +\infty$. By the last assumption,

$$J_\gamma(g) = D_f(g_1, g_2) = \int_{\{g_2 > 0\}} g_2 f\left(\tfrac{g_1}{g_2}\right) d\mu$$

if $g = (g_1, g_2)$ is a pair of nonnegative functions on Z such that $g_2 = 0$ implies $g_1 = 0$, μ-a.e. Otherwise, $D_f(g_1, g_2) = +\infty$. The minimization problem of this section takes the form

$$H_\gamma(a) = \inf_{(g_1,g_2) \in \mathcal{L}_a} D_f(g_1, g_2), \quad a \in \mathbb{R}^d. \quad (9)$$

A. Preliminaries and technicalities

The assumptions on f imply that the convex conjugate f^* is everywhere finite, differentiable, constant on $(-\infty, f'(0)]$ and increasing on $(f'(0), +\infty)$.

If $f(0) = +\infty$ then the domain of γ is the union of the open positive quadrant and the origin $(0,0)$. Otherwise, the domain contains additionally the halfline $\ell = \{(0,s) : s > 0\}$.

Lemma 3. *At $y = (u,v)$ with u,v positive*

$$\nabla\gamma(y) = \left(f'\left(\tfrac{u}{v}\right), f\left(\tfrac{u}{v}\right) - \tfrac{u}{v} f'\left(\tfrac{u}{v}\right) \right). \quad (10)$$

For $y = (0,0)$

$$\partial\gamma(y) = \{(\alpha,\beta) \in \mathbb{R}^2 : f^*(\alpha) + \beta \leqslant 0\}. \quad (11)$$

If $f(0)$ is finite then for all $y \in \ell$

$$\partial\gamma(y) = \{(\alpha, f(0)) : \alpha \leqslant f'(0)\}. \quad (12)$$

A proof relies on [14, Theorem 23.2].

Lemma 4. *For $\theta = (\alpha, \beta) \in \mathbb{R}^2$*

$$\gamma^*(\theta) = \begin{cases} 0\,, & f^*(\alpha) + \beta \leqslant 0\,, \\ +\infty\,, & \textit{otherwise,} \end{cases} \tag{13}$$

and $\partial\gamma^*(\theta) = \{v(f^{*\prime}(\alpha), 1) \colon v \geqslant 0\}$ if $f^*(\alpha) + \beta = 0$.

Eq. (13) follows from [14, Corollary 13.5.1].

Lemma 5. *$dom(\gamma^*) = dom(\partial\gamma^*) = \partial\gamma(y)$ where $y = (0, 0)$.*

Lemma 6. *For $x = (t, s) \in dom(\gamma)$, $y = (u, v) \in dom(\partial\gamma)$*

$$\varDelta_\gamma(x, y) = \begin{cases} s\,\varDelta_f\big(\frac{t}{s}, \frac{u}{v}\big)\,, & s, v > 0\,, \\ 0\,, & u = 0\,,\, v = 0\,. \end{cases}$$

Proof. For $s, v > 0$ the assertion follows by simple calculus, using (10) in the case $u > 0$ and (12) when $u = 0$. For $s > 0$ and $u = v = 0$

$$\varDelta_\gamma(x, y) = sf\big(\tfrac{t}{s}\big) - \sup_{\vartheta \in \partial\gamma(y)} \langle \vartheta, x\rangle = 0$$

because by (11) the supremum equals

$$\sup\,\{\alpha t + \beta s \colon \alpha \in \mathbb{R}\,,\, \beta \leqslant -f^*(\alpha)\} = sf^{**}\big(\tfrac{t}{s}\big)$$

and $f^{**} = f$. If instead $s = 0$, the assertion is trivial. □

The previous lemma is partially covered by identities stated in [1, Remark 6.10] for a broader class of functions γ than those given by (3).

Lemma 7. *For x, y as in Lemma 6 and $\theta = (\alpha, \beta) \in \partial\gamma(y)$*

$$\varUpsilon_\gamma(x, y, \theta) = \begin{cases} 0\,, & u > 0\,,\, v > 0\,, \\ t\,[f'(0) - \alpha]\,, & u = 0\,,\, v > 0\,, \\ \gamma(x) - \langle\theta, x\rangle\,, & u = 0\,,\, v = 0\,, \end{cases}$$

where the second case takes place only when $f'(0)$ is finite. In the third case, $\varUpsilon_\gamma(x, y, \theta)$ vanishes if and only if $x = (0, 0)$ or $x = s(f^{*\prime}(\alpha), 1)$ and $f^*(\alpha) + \beta = 0$.

Proof. The first case follows by the differentiability of γ at y. In the second one, Lemma 3 applies to recognize that $f'(0)$ is finite, to conclude that $\beta = f(0)$ and to compute $\varUpsilon_\gamma(x, y, \theta)$ as the maximum of the scalar product of $(w - \alpha, f(0) - \beta)$ and $x - y = (t, s - v)$ over $w \leqslant f'(0)$. The last case follows from Lemma 1, where $\varDelta_\gamma(x, y) = 0$ by Lemma 6, and $\gamma^*(\theta) = 0$, on account of the assumption $\theta \in \partial\gamma(y)$ and Lemmas 5 and 4. Hence, $\varUpsilon_\gamma(x, y, \theta) = 0$ if and only if $\gamma(x) - \langle\theta, x\rangle = -\gamma^*(\theta)$ which is equivalent to $x \in \partial\gamma^*(\theta)$. Then, the last assertion follows from Lemma 4. □

B. Specialization of the main results

The minimization problem (9) is studied by applying the results of Section III and the above technical lemmas. Let φ_1 and φ_2 be the columns of φ so that if $\vartheta \in \mathbb{R}^d$ then $\vartheta\varphi$ has the coordinates $\vartheta\varphi_1$ and $\vartheta\varphi_2$. By Lemma 4, the values of $\vartheta\varphi$ belong to $dom(\partial\gamma^*)$ if and only if

$$f^*(\vartheta\varphi_1) + \vartheta\varphi_2 \leqslant 0\,. \tag{14}$$

For ϑ satisfying (14) μ-a.e. it follows from the second assertion of Lemma 4 that $h = (h_1, h_2)$ is a measurable selection from $\partial\gamma^*(\vartheta\varphi)$ if and only if, μ-a.e.,

$$h_1 = f^{*\prime}(\vartheta\varphi_1)h_2 \tag{15}$$

and $h_2 = 0$ outside the set

$$Z_\vartheta = \{z \in Z : f^*(\vartheta\varphi_1(z)) + \vartheta\varphi_2(z) = 0\}\,.$$

The global assumption is satisfied, because the f-divergence $D_f(g_1, g_2)$ is zero if both g_1 and g_2 vanish identically. Hence, eq. (7) takes the form

$$H^{**}_\gamma(a) = \sup\{\langle\vartheta, a\rangle \colon \vartheta \text{ satisfies (14) } \mu\text{-a.e.}\}\,, \quad a \in \mathbb{R}^d\,. \tag{16}$$

Theorem 2 is specialized for the present case as follows.

Theorem 3. *If a belongs to $ri(dom(H_\gamma))$ and $H_\gamma(a)$ is finite then the supremum in (16) is attained at some ϑ and coincides with $H_\gamma(a)$. For this ϑ and any $g = (g_1, g_2) \in \mathcal{L}_a$*

$$\begin{aligned} D_f(g_1, g_2) = H_\gamma(a) + \textstyle\int_{A\cap B} g_2\,\varDelta_f\big(\frac{g_1}{g_2}, f^{*\prime}(\vartheta\varphi_1)\big)\,d\mu \\ + \textstyle\int_{A\cap C} g_1\big[f'(0) - \vartheta\varphi_1\big]\,d\mu + \int_{Z\setminus A}\big[\gamma(g) - \langle\vartheta\varphi, g\rangle\big]\,d\mu \end{aligned} \tag{17}$$

where A is any measurable subset of Z_ϑ and B, C are the subsets of Z given by $g_2 > 0$ and $\vartheta\varphi_1 \leqslant f'(0)$, respectively.

Proof. The first assertion follows directly from Theorem 2. The finiteness of (16) at a maximizer ϑ implies that $\vartheta\varphi$ belongs to $dom(\gamma^*)$, μ-a.e. The additional hypotheses of Theorem 2 holds because $dom(\gamma^*) = dom(\partial\gamma^*)$ by Lemma 5. If $A \subseteq Z_\vartheta$ is measurable then $h_2 = \mathbf{1}_A$ and $h_1 = f^{*\prime}(\vartheta\varphi_1)\mathbf{1}_A$ give rise to a measurable selection $h = (h_1, h_2)$ from $z \mapsto \partial\gamma^*(\vartheta\varphi(z))$ as discussed above. Then, the second assertion of Theorem 2 applies and (17) results from (8). In fact, by the substitutions

$$\begin{aligned} g = (g_1, g_2) \ &\text{for}\ x = (t, s) \\ h = (h_1, h_2) \ &\text{for}\ y = (u, v) \\ \vartheta\varphi = (\vartheta\varphi_1, \vartheta\varphi_2) \ &\text{for}\ \theta = (\alpha, \beta) \end{aligned}$$

Lemma 6 implies that

$$B_\gamma(g, h) = \textstyle\int_{\{g_2 > 0\,,\, h_2 > 0\}} g_2\,\varDelta_f\big(\frac{g_1}{g_2}, \frac{h_1}{h_2}\big)\,d\mu \tag{18}$$

equals the integral on the first line of eq. (17). By Lemma 7,

$$L_\gamma(g, h, \vartheta\varphi) = \textstyle\int_{\{h_1 = 0\}} \varUpsilon_\gamma(g, h, \vartheta)\,d\mu\,. \tag{19}$$

Splitting the integration further to A and $Z \setminus A$, the two integrals on the second line of eq. (17) emerge. The function h_1 vanishes on A if and only if $f^{*\prime}(\vartheta\varphi_1) = 0$ which is equivalent to $\vartheta\varphi_1 \leqslant f'(0)$. □

Remark 8. A set A in Theorem 3 originated from a selection h in Theorem 2. Though the selections in the above proof are not in the most general form, it is not difficult to see that no generality was lost when rewriting (8) to (17). In fact, for arbitrary selection $h = (h_1, h_2)$, satisfying (15) and $h_2 = 0$ outside Z_ϑ, μ-a.e., the integral in (18) depends on h only through $\{h_2 > 0\}$ and the ratio $\frac{h_1}{h_2} = f^{*\prime}(\vartheta\varphi_1)$ on this set, and (19) depends on h only through $\{h_1 = 0\}$, expressible also via that set and ratio. It follows that all these selections h are generalized minimizers in the problem (9). This may be of limited value in general, but if either of them happens to satisfy the given constraints, $h \in \mathcal{L}_a$, then it has to be a true minimizer, by Remark 1.

Remark 9. Under the assumptions of Theorem 3, if g is a minimizer in the problem (9) then g_2 vanishes outside Z_ϑ and $g_1 = f^{*\prime}(\vartheta\varphi_1)g_2$, μ-a.e., thus g is among the generalized minimizers discussed in Remark 8. In fact, in eq. (17) with $A = Z_\vartheta$ the last integral vanishes, and thus $g = 0$, μ-a.e. outside Z_ϑ, by the last assertion of Lemma 7. Then, in eq. (17) $B = \{g_2 > 0\} \cap Z_\vartheta$ can play the role of A, the first integral vanishes, and therefore $g_1 = f^{*\prime}(\vartheta\varphi_1)g_2$ on B, μ-a.e., by the properties of f. The finiteness of $D_f(g_1, g_2)$ implies that $g_1 = 0$ whenever $g_2 = 0$ so that $g_1 = f^{*\prime}(\vartheta\varphi_1)g_2$, μ-a.e.

Example 2. Let $Z = (0, 1]$ be endowed with the Lebesgue measure, $d = 2$ and $f(t) = \frac{1}{2}t^2$, $t \geqslant 0$. Then $f^*(s) = |s|_+^2/2$. Further, $\gamma(x) = \frac{t^2}{2s}$ for $x = (t, s)$ with $t \geqslant 0$, $s > 0$. Let φ_1 consist of the functions $z, 0$ and φ_2 of $0, z^2$. Then $\mathcal{L}_a$ with $a = (a_1, a_2)$ contains pairs $g = (g_1, g_2)$ of nonnegative functions satisfying

$$\int_0^1 z g_1(z)\, dz = a_1 \quad \text{and} \quad \int_0^1 z^2 g_2(z)\, dz = a_2\,.$$

Let a_1, a_2 be positive. By eq. (16),

$$\begin{aligned} H_\gamma^{**}(a) &= \sup\{\vartheta_1 a_1 + \vartheta_2 a_2 \colon |\vartheta_1|_+^2 + 2\vartheta_2 \leqslant 0\} \\ &= \max\{\vartheta_1 a_1 - \tfrac{1}{2}\vartheta_1^2 a_2 \colon \vartheta_1 \geqslant 0\} = \frac{a_1^2}{2a_2} \end{aligned}$$

where the supremum is attained at $\vartheta_1 = \frac{a_1}{a_2}$ and $\vartheta_2 = -\frac{a_1^2}{2a_2^2}$, uniquely. Then, $H_\gamma(a) = H_\gamma^{**}(a)$.

The mapping φ was chosen on purpose to make ineq. (14) with the maximizing $\vartheta = (\vartheta_1, \vartheta_2)$ tight, and thus to achieve $Z_\vartheta = Z$. Hence, $h = (h_1, h_2)$ is a measurable selection from $z \mapsto \partial\gamma^*(\vartheta\varphi(z))$ if and only if h_2 is a nonnegative measurable function on Z and

$$h_1(z) = |\vartheta_1 z|_+ h_2(z) = \tfrac{a_1}{a_2} z h_2(z)\,,$$

a.e. In particular, the selection $h = 3(a_1 z, a_2)$ satisfies the constraints, $h \in \mathcal{L}_a$, and thus is a minimizer of the problem (9) by Remark 8. Any minimizer $g = (g_1, g_2)$ in (9) must have the form $(\frac{a_1}{a_2} z g_2, g_2)$ with $g_2 \geqslant 0$ by Remark 9. A pair of this form is a minimizer if and only if $z^2 g_2(z)$ integrates to a_2. Let us remark that this example can be analyzed also in a simple direct way, without any convex analysis.

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] Bauschke, H.H. and Borwein, J.M., Legendre functions and the method of random Bregman projections. *J. Convex Anal.* **4** (1997) 27-67.

[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 Mathematical Physics* **7** (1967) 200–217.

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

[5] Csiszár, I., Eine informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizität von Markoffschen Ketten. *Publ. Math. Inst. Hungar. Acad. Sci.* **8** (1963) 85–108.

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

[7] 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.

[8] Csiszár, I. and Matúš, F., On minimization of entropy functionals under moment constraints. *Proc. ISIT 2008*, Toronto, Canada, 2101–2105.

[9] Csiszár, I. and Tusnády, G., Information geometry and alternating minimization procedures, *Statistics and Decisions* **1** (1984) 205–237.

[10] Györfi, L. and Nemetz, T., f-dissimilarity: a generalization of the affinity of several distribution. *Annals of Inst. Statist. Math.* **30** (1978) 105–113.

[11] Le Cam, L., *Asymptotic Methods in Statistical Decision Theory*. Springer Verlag, Berlin, Heidelberg, New York, 1986.

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

[13] Léonard, C., Minimization of entropy functionals. *J. Math. Anal. Appl.* **346** (2008) 183-204.

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

[15] Rockafellar, R.T. and Wets, R.J-B., *Variational Analysis.* Springer Verlag, Berlin, Heidelberg, New York 2004.