Proceedings IEEE ISIT 2012, Cambridge, MA, USA, 150–154

Minimization of Entropy Functionals Revisited

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*—Integral functionals based on convex normal integrands are minimized subject to finitely many moment constraints. The integrands are assumed to be strictly convex but not autonomous or differentiable. The effective domain of the value function is described by a modification of the concept of convex core. The minimization is viewed as a primal problem and studied together with a dual one in the framework of convex duality. Main results assume a dual constraint qualification but dispense with the primal constraint qualification. Minimizers and generalized minimizers are explicitly described whenever the primal value is finite. Existence of a generalized dual solution is established whenever the dual value is finite. A generalized Pythagorean identity is presented using Bregman distance and a correction term. Results are applied to minimization of Bregman distances.**

I. INTRODUCTION

This work addresses minimization of *integral functionals*

(1) $$H_\beta(g) \triangleq \int_Z \beta(z, g(z))\, \mu(dz)$$

of real functions g on a σ-finite measure space $(Z, \mathcal{Z}, \mu)$, subject to the constraint that the *moment vector* $\int_Z \varphi g\, d\mu$ of g is prescribed. Here, φ is a given $\mathbb{R}^d$-valued $\mathcal{Z}$-measurable *moment mapping*.

In (1), β is any mapping $Z \times \mathbb{R} \to (-\infty, +\infty]$ such that $\beta(\cdot, t)$ is $\mathcal{Z}$-measurable for $t \in \mathbb{R}$, and $\beta(z, \cdot)$, $z \in Z$, is in the class Γ of functions γ on $\mathbb{R}$ that are finite and strictly convex for $t > 0$, equal to $+\infty$ for $t < 0$, and satisfy $\gamma(0) = \lim_{t \downarrow 0} \gamma(t)$. In particular, β is a *normal integrand* whence $z \mapsto \beta(z, g(z))$ is $\mathcal{Z}$-measurable if g is [18, Chapter 14]. If neither the positive nor the negative part of $\beta(z, g(z))$ is μ-integrable, the integral in (1) is $+\infty$ by convention. The integrand is *autonomous* if $\beta(z, \cdot) = \gamma$, $z \in Z$, for some $\gamma \in \Gamma$.

Given $a \in \mathbb{R}^d$, let $\mathcal{G}_a$ denote the class of those nonnegative $\mathcal{Z}$-measurable functions g whose moment vector exists and equals a. By the assumptions on β, the minimization of H_β restricts to $\mathcal{G}_a$. Following the key papers [3], [4], *convex duality* has become a standard tool in the mathematically oriented literature on the subject. The *value function*

(2) $$J_\beta(a) \triangleq \inf_{g \in \mathcal{G}_a} H_\beta(g)\,, \qquad a \in \mathbb{R}^d\,,$$

turns out to be closely related to the convex conjugate

(3) $$K_\beta^*(a) = \sup_{\vartheta \in \mathbb{R}^d} \left[\langle \vartheta, a \rangle - K_\beta(\vartheta) \right], \qquad a \in \mathbb{R}^d\,,$$

of the function K_β given by

(4) $$K_\beta(\vartheta) \triangleq \int_Z \beta^*\big(z, \langle \vartheta, \varphi(z) \rangle\big)\, \mu(dz)\,, \qquad \vartheta \in \mathbb{R}^d\,.$$

Here, $\langle \cdot, \cdot \rangle$ denotes the scalar product on $\mathbb{R}^d$ and $\beta^*(z, r)$ is equal to $\sup_{t \in \mathbb{R}} [tr - \beta(z, t)]$ for $r \in \mathbb{R}$. In (2)/(3), the minimization/maximization is called *primal/dual problem*, the infimum/supremum *primal/dual value*, and a minimizer/maximizer, if exists, is a *primal/dual solution*. Since β is strictly convex, the primal solution, denoted by g_a, is unique in the sense that any two minimizers are equal μ-a.e.

Minimization problems as in (2) emerge across various scientific disciplines, notably in *inference*. When g is an unknown probability density, or any nonnegative function, whose moment vector is determined by measurements and a specific choice of β is justified, often the primal solution as above is adopted as the 'best guess' of g. Among autonomous integrands, typical choices of β are $t \ln t$ or $-\ln t$ or t^2 giving $H_\beta(g)$ equal to the negative Shannon or Burg entropy or the squared L^2-norm of $g \geqslant 0$. Note also that the maximum likelihood estimation in exponential families is a special case of the dual problem (3), with β equal to $t \ln t$. When a 'prior guess' h for g is available, that would be adopted before the measurement, it is common to use a non-autonomous integrand β depending on h for which $H_\beta(g)$ represents a non-metric distance of g from h. Two cases are prominent: γ-*divergence* $\int_Z h\, \gamma(g/h) d\mu$ with $\gamma \in \Gamma$ [6], [1], [20] and *Bregman distance* [5], [13], [16], see (6). Then the corresponding primal solution is often referred to as a *projection* of h to $\mathcal{G}_a$. In particular, the most familiar *I-projections* correspond to the information (I-) divergence that belongs to both families of distances. Note that while the minimization of a γ-divergence can be reduced to that of an integral functional with autonomous integrand, this is in general not possible for Bregman distances whence the nonautonomous integrands become unavoidable.

The primal problem is well understood when the following *primal constraint qualification* (PCQ) holds

(PCQ) $$a \in ri(dom(J_\beta)) \text{ and } J_\beta(a) > -\infty\,.$$

Here, ri stands for the relative interior and dom for the effective domain. Theorem 1 below implies by standard convex duality results [17] that under the PCQ the primal and dual values coincide, a dual solution exists and explicitly specifies the primal solution when the latter exists. This covers for example the classical maximization of Shannon differential entropy over probability densities on $\mathbb{R}$ with given mean and variance. Even under the PCQ the primal solution need not exist, as in Example 1. Note that while the PCQ often holds for each $a \in dom(J_\beta)$, perhaps with exception of the origin, in many inference problems of practical interest this is not the case. Hence, dispensing with the PCQ is desirable.

This work is a follow-up of the authors' contribution [11] at ISIT'08 motivated by their previous work [9] about I-projections. As there, the PCQ is dispensed with, and in the case when no primal/dual solutions exist, generalized solutions in the sense of [19], [7] are studied. In [11], as in most of the previous literature, it is assumed that the integrand is autonomous, differentiable, and that the moment mapping has one coordinate function identically equal to 1. In this contribution, these assumptions are avoided, saving as many conclusions as possible. For previous works not making these assumptions see [14], [15], using advanced tools of functional analysis. No such tools are used here, and neither is differential geometry, see [2], which is powerful but requires strong regularity assumptions.

Non-autonomous integrands do not entail conceptual difficulties since problems with measurability can be handled via the machinery of normal integrands [18]. Non-differentiability of β causes few results to fail. Absence of the special coordinate of φ is cured by adopting a dual constraint qualification. Some results here are new even in the framework of [11].

Space limitations do not admit a detailed presentation of results, let alone their proofs. For these, and more references, see the full paper [12] available on arXiv.

II. Main ingredients

The key for the close relationship of the primal and dual problems is the following fact which, in this generality, does not seem to directly follow from known results. It can be proved via modifications of arguments in [18].

Theorem 1. *If $J_\beta \not\equiv +\infty$ then $J_\beta^* = K_\beta$.*

The hypothesis holds if K_β is finite on an open set. The weaker assumption that K_β is proper (not identically $+\infty$ and not attaining $-\infty$) is not sufficient for $J_\beta \not\equiv +\infty$, see [12].

Theorem 2. *If K_β is finite in a neighborhood of $\vartheta \in \mathbb{R}^d$ then it is differentiable there, and $J_\beta(\nabla K_\beta(\vartheta))$ is finite.*

A special role is played by the set Θ_β of those $\vartheta \in dom(K_\beta)$ for which the function $r \mapsto \beta^*(z, r)$ is finite in a neighborhood of $r = \langle \vartheta, \varphi(z) \rangle$ for μ-a.a. $z \in Z$. Equivalently,

$$\Theta_\beta = \big\{ \vartheta \in dom(K_\beta) : \langle \vartheta, \varphi(z) \rangle < \beta'(z, +\infty) \;\; \mu\text{-a.e.} \big\}$$

where $\beta'(z, +\infty)$ denotes the limit of right derivatives $\beta'_+(z, t)$ when $t \uparrow +\infty$. Since $\beta(z, \cdot) \in \Gamma$ implies that $\beta^*(z, \cdot)$ is differentiable on $(-\infty, \beta'(z, +\infty))$, if $\vartheta \in \Theta_\beta$ then

$$z \mapsto (\beta^*)'(z, \langle \vartheta, \varphi(z) \rangle), \quad z \in Z,$$

defines a function f_ϑ on Z, up to a μ-negligible set. Similarly to [11], the family $\mathcal{F}_\beta \triangleq \{f_\vartheta : \vartheta \in \Theta_\beta\}$ plays the role of generalized exponential families, provided that the following *dual constraint qualification* (DCQ) holds

$$\text{(DCQ)} \qquad \Theta_\beta \text{ is nonempty.}$$

The DCQ follows from $dom(K_\beta) \neq \emptyset$ when one component of φ is 1. When K_β is finite on an open set, the DCQ holds if and only if

$$\text{(5)} \quad \mu\big\{ z \in Z : \varphi(z) = \mathbf{0} \text{ and } \lim_{t \uparrow +\infty} \beta(z, t) \neq \pm\infty \big\} = 0 .$$

If the DCQ holds then the maximization in the dual problem can be restricted to Θ_β without changing the dual value and loosing a dual solution. Under the DCQ, if dual solutions ϑ exist then each one induces the same function f_ϑ, called here the *effective dual solution* g_a^*.

Proposition 1. *For any $a \in \mathbb{R}^d$, if $f_\vartheta \in \mathcal{G}_a$ for some $\vartheta \in \Theta_\beta$ with $K_\beta(\vartheta)$ finite then f_ϑ equals the effective dual solution g_a^*, as well as the primal solution g_a. Under the* PCQ, *the primal solution g_a exists if and only if $\mathcal{G}_a \cap \mathcal{F}_\beta \neq \emptyset$.*

As a consequence, if K_β is essentially smooth and (5) holds then the primal solution g_a exists for each $a \in ri(dom(J_\beta))$.

The Bregman distance based on β is given by

$$\text{(6)} \qquad B_\beta(g, h) \triangleq \int_Z \Delta_\beta(z, g(z), h(z))\, \mu(dz)$$

where g, h are nonnegative $\mathcal{Z}$-measurable functions and Δ_β is a nonnegative integrand such that $\Delta_\beta(z, s, t)$ for $z \in Z$ and $s, t \geqslant 0$ is equal to

$$\gamma(s) - \gamma(t) - \gamma'_{sgn(s-t)}(t)[s - t] \quad \text{if } \gamma'_+(t) \text{ is finite,}$$

and $s \cdot (+\infty)$ otherwise. Here, $\gamma \in \Gamma$ abbreviates $\beta(z, \cdot)$ and $sgn(r)$ denotes $+$ if $r \geqslant 0$ and $-$ if $r < 0$.

Bregman distances enter into the primal problem via the following identity.

Proposition 2. *Assuming the* PCQ *for $a \in \mathbb{R}^d$ and the* DCQ,

$$\text{(7)} \qquad H_\beta(g) = J_\beta(a) + B_\beta(g, g_a^*) + C_\beta(g), \qquad g \in \mathcal{G}_a .$$

In (7), C_β is a sophisticated nonnegative correction functional defined explicitly in [12].

These results have been known in special settings with β autonomous and (7) in a weaker form as the inequality obtained by neglecting the correction. If β is essentially smooth then the correction vanishes anyhow and (7) is known as a *Pythagorean identity*. For autonomous β that is differentiable but not essentially smooth, the correction is determined in [11].

If the primal value $J_\beta(a)$ is finite and all sequences g_n in $\mathcal{G}_a$ with $H_\beta(g_n) \to J_\beta(a)$ converge to a common limit function $\hat{g}_a$ then $\hat{g}_a$ is called the *generalized primal solution*. Here, the convergence is locally in measure, thus in measure on each set $A \in \mathcal{Z}$ of finite μ-measure. Proposition 2 implies that, subject to the PCQ and DCQ, the generalized primal solution $\hat{g}_a$ exists and equals the effective dual solution g_a^*. Under the PCQ, the generalized primal solution exists if and only if the DCQ holds. This is a new result here, included in Theorem 6*(iii)*.

The following example shows that the generalized primal solution $\hat{g}_a$ may be independent of a, and its moment vector need not exist. For other examples illustrating possible irregularities see [12].

Example 1. Let μ be the Cauchy distribution on $Z = \mathbb{R}$ with $d\mu = [\pi(1 + z^2)]^{-1}dz$, let β be the autonomous integrand given by $\beta(z,t) = t\ln t$, $t > 0$, and $\varphi(z) = (1, z)$, $z \in Z$. For $a = (a_1, a_2)$ with $a_1 \geqslant 0$ and $g \in \mathcal{G}_a$, if ν denotes the measure with $d\nu = g\,d\mu$ then $H_\beta(g)$ equals the I-divergence $D(\nu\|\mu)$. In the particular case $a_1 = 1$, the primal problem is equivalent to minimization of I-divergence from μ over the probability measures ν with mean a_2. In dual problems, $\beta^*(z,r) = e^{r-1}$, $\vartheta = (\vartheta_1, \vartheta_2)$ and $K_\beta(\vartheta_1, \vartheta_2) = \int_{\mathbb{R}} e^{\vartheta_1+\vartheta_2 z-1}\,\mu(dz)$ is equal to e^{ϑ_1-1} if $\vartheta_2 = 0$, and $+\infty$, otherwise. Hence, Θ_β equals $\mathit{dom}(K_\beta) = \mathbb{R} \times \{0\}$. For a in $\mathit{ri}(\mathit{dom}(J_\beta)) = (0, +\infty) \times \mathbb{R}$

$$J_\beta(a) = K^*_\beta(a) = \max_{\vartheta_1\in\mathbb{R}} \left[a_1\vartheta_1 - e^{\vartheta_1-1}\right] = a_1 \ln a_1\,,$$

$(1+\ln a_1, 0)$ is the dual solution, and the effective dual solution g^*_a identically equals a_1. Since φg^*_a is not μ-integrable the primal solution does not exist, by Proposition 1. Nevertheless, g^*_a equals the generalized primal solution $\hat{g}_a$, by Proposition 2.

Additionally to (7), Bregman distances emerge also in the dual problem (3), via the following existence result.

Theorem 3. *Assuming the* DCQ, *for every $a \in \mathbb{R}^d$ with $K^*_\beta(a)$ finite there exists a unique $\mathcal{Z}$-measurable function h_a such that*

$$\text{(8)}\quad K^*_\beta(a) - \left[\langle\vartheta, a\rangle - K_\beta(\vartheta)\right] \geqslant B_\beta(h_a, f_\vartheta)\,, \qquad \vartheta \in \Theta_\beta\,.$$

As a consequence, if the dual problem has a solution $\vartheta \in \Theta_\beta$ then h_a equals $g^*_a = f_\vartheta$. In general, (8) implies that whenever ϑ_n is a maximizing sequence for $\langle\vartheta, a\rangle - K_\beta(\vartheta)$, the Bregman distances $B_\beta(h_a, f_{\vartheta_n})$ tend to zero, and thus f_{ϑ_n} converges to h_a locally in measure. The function h_a in (8) is regarded as *generalized dual solution* for a, extending the concept of *generalized maximum likelihood estimate* introduced in [9] and explicitly constructed in [10]. Our current proof of Theorem 3 is non-constructive, except for the case of equal primal and dual values, when h_a is equal to the generalized primal solution $\hat{g}_a$.

III. The effective domain of the value function

The set of the moment vectors $\int_Z \varphi g\,d\mu$ of all nonnegative functions g with φg integrable is a convex cone, referred to as the *φ-cone $\mathit{cn}_\varphi(\mu)$* of μ. It contains the effective domain of J_β. In this section, the domain is described via faces of $\mathit{cn}_\varphi(\mu)$.

Recall that a subset C of $\mathbb{R}^d$ is a cone if $\mathbf{0} \in C$, and $tx \in C$ whenever $t > 0$ and $x \in C$. A *face* of a convex set C is a nonempty convex subset F of C such that every closed line segment in C with a relative interior point in F is contained in F. The face is proper if $F \neq C$. A face of a convex cone is a convex cone.

The φ-cones will be studied via a new notion defined next, for Borel measures Q on $\mathbb{R}^d$ that are σ-finite on $\mathbb{R}^d \setminus \{\mathbf{0}\}$.

Definition 1. The *conic core $\mathit{cnc}(Q)$* of Q is the intersection of all convex Borel cones with Q-negligible complements.

A predecessor of this construction appeared first in [8]: the convex core $\mathit{cc}(Q)$ of Q is the intersection of all convex Borel sets with Q-negligible complements. By [10, Theorem 3], $\mathit{cc}(Q)$ consists of the integrals $\int_{\mathbb{R}^d} x\,P(dx)$ where P is a probability measure dominated by Q.

Lemma 1. *The closure of $\mathit{cnc}(Q)$ coincides with the smallest closed convex cone with Q-negligible complement.*

Let μ_φ denote the φ-image of μ. In general, it is not σ-finite.

Lemma 2. *If ν is a measure equivalent to μ and the image ν_φ is σ-finite on $\mathbb{R}^d \setminus \{\mathbf{0}\}$ then $\mathit{cn}_\varphi(\mu) = \mathit{cnc}(\nu_\varphi)$.*

It follows that $\varphi(z) \in \mathit{cl}(\mathit{cn}_\varphi(\mu))$ for μ-a.a. $z \in Z$.

Lemma 3. *If F is a face of $\mathit{cnc}(Q)$ then $\mathit{cnc}(Q^{\mathit{cl}(F)}) = F$.*

Here, $Q^{\mathit{cl}(F)}$ denotes the restriction of Q to $\mathit{cl}(F)$. Hence, the φ-cone of the restriction of μ to $\varphi^{-1}(\mathit{cl}(F))$ equals F, and $\mu(\varphi^{-1}(\mathit{cl}(F))) > 0$ except perhaps for $F = \{\mathbf{0}\}$.

Theorem 4. *The conic core $\mathit{cnc}(Q)$ consists of the integrals $\int_{\mathbb{R}^d} xP(dx)$ for all finite (or σ-finite) measures P dominated by Q such that x is P-integrable.*

It follows that $\mathit{cnc}(Q)$ equals the conic hull of $\mathit{cc}(Q)$. The faces of the two sets are not related to each other in general. However, if Q is concentrated on a hyperplane that does not contain the origin then there is a bijection between the families of faces of $\mathit{cc}(Q)$ and $\mathit{cnc}(Q)$, up to the face $\{\mathbf{0}\}$ of the latter: the faces of $\mathit{cnc}(Q)$ are the conic hulls of the faces of $\mathit{cc}(Q)$. In particular, this takes place for $Q = \mu_\varphi$ whenever one coordinate of φ is 1, a fact that has been used in [11].

Lemma 4. *The moment vector $\int_Z \varphi g\,d\mu$ of a function $g \geqslant 0$ with the φg integrable belongs to a face F of $\mathit{cn}_\varphi(\mu)$ if and only if $g(z) = 0$ for μ-a.a. $z \in Z$ with $\varphi(z) \notin \mathit{cl}(F)$.*

Let $\boldsymbol{F}_\beta$ denote the family of faces F of $\mathit{cn}_\varphi(\mu)$ such that the positive part of the integral $\int_{\{\varphi\notin \mathit{cl}(F)\}} \beta(\cdot, 0)\,d\mu$ is finite. The improper face $\mathit{cn}_\varphi(\mu)$ belongs to $\boldsymbol{F}_\beta$, and if $F \subseteq G$ are faces of $\mathit{cn}_\varphi(\mu)$ and F belongs to $\boldsymbol{F}_\beta$ then so does also G.

Theorem 5. $\mathit{dom}(J_\beta) = \bigcup_{F\in\boldsymbol{F}_\beta} \mathit{ri}(F)$ *whenever $J_\beta \not\equiv +\infty$.*

This shows that $\mathit{dom}(J_\beta)$ is closed to positive multiples.

Another consequence is that $\int_Z \beta(\cdot, 0)\,d\mu < +\infty$ is a sufficient condition for $\mathit{dom}(J_\beta) = \mathit{cn}_\varphi(\mu)$. This condition is also necessary when $\{\mathbf{0}\}$ is a face of $\mathit{cn}_\varphi(\mu)$ and $\varphi \neq \mathbf{0}$, μ-a.e., in particular, when one coordinate of φ identically equals 1.

Example 2. Let $Z = \mathbb{R} \times \{0, 1\}$ and $\mu = \mu_1 \times \mu_2$ where μ_1 is a finite measure on $\mathbb{R}$ equivalent to the Lebesgue measure and μ_2 is the counting measure. Let the moment mapping be given by $\varphi(z) = (1, z_1, z_2)$, $z = (z_1, z_2) \in Z$. Then the φ-cone of μ equals the sum of the sets

$$F = \{(r, s, 0)\colon r > 0\,,\ s \in \mathbb{R}\} \cup \{\mathbf{0}\}$$
$$G = \{(r, s, r)\colon r > 0\,,\ s \in \mathbb{R}\} \cup \{\mathbf{0}\}\,.$$

Its proper faces are F, G and $\{\mathbf{0}\}$. Let the integrand β be given by $\beta(z,t) = (t + z_2)^{-1}$, $z \in Z$, $t > 0$. Then $\beta(\cdot, 0)$ equals $+\infty$ on $\varphi^{-1}(F) = \mathbb{R} \times \{0\}$, and 1 on $\varphi^{-1}(G) = \mathbb{R} \times \{1\}$. Theorem 5 implies that $\mathit{dom}(J_\beta)$ equals

$$\mathit{ri}(F) \cup \mathit{ri}(\mathit{cn}_\varphi(\mu)) = \{(r, s, q)\colon r > q \geqslant 0\,,\ s \in \mathbb{R}\}\,.$$

IV. Dispensing with the PCQ in the primal problem

In this section, the primal problem is studied when the value $J_\beta(a)$ is finite but the PCQ is not assumed, $a \notin ri(dom(J_\beta))$. The main idea is to replace the measure μ by its restriction to the set $\{\varphi \in cl(F)\} \triangleq \varphi^{-1}(cl(F))$ where F is an appropriate face of the φ-cone $cn_\varphi(\mu)$. To indicate this, the letter F is added to indices, e.g. $\mathcal{G}_{F,a}$ denotes the class of $\mathcal{Z}$-measurable functions $g\colon Z \to [0,+\infty)$ such that the integral $\int_{\{\varphi\in cl(F)\}} \varphi g\, d\mu$ exists and equals a.

For a face F of $cn_\varphi(\mu)$ and $a \in \mathbb{R}^d$, the minimization in

$$J_{F,\beta}(a) \triangleq \inf_{g\in\mathcal{G}_{F,a}} H_{F,\beta}(g)\,,$$

where $H_{F,\beta}(g) \triangleq \int_{\{\varphi\in cl(F)\}} \beta(z,g(z))\,\mu(dz)$, is the *$F$-primal problem* and the maximization in

$$K^*_{F,\beta}(a) \triangleq \sup_{\vartheta\in\mathbb{R}^d} \left[\langle\vartheta, a\rangle - K_{F,\beta}(\vartheta)\right],$$

where $K_{F,\beta}(\vartheta) \triangleq \int_{\{\varphi\in cl(F)\}} \beta^*\big(z,\langle\vartheta,\varphi(z)\rangle\big)\,\mu(dz)$, is the *$F$-dual problem* for a. If the F-primal value $J_{F,\beta}(a)$ is finite and attained then the minimizers vanishing outside $\{\varphi \in cl(F)\}$ define the μ-unique *F-primal solution $g_{F,a}$* for a. The *generalized F-primal solution $\hat g_{F,a}$* is defined likewise.

Let the set $\Theta_{F,\beta}$ consist of those $\vartheta \in dom(K_{F,\beta})$ for which the function $r \mapsto \beta^*(z,r)$ is finite around $r = \langle\vartheta,\varphi(z)\rangle$ for μ-a.a. $z \in Z$ with $\varphi(z) \in cl(F)$. For $\vartheta \in \Theta_{F,\beta}$ let $f_{F,\vartheta}(z)$ equal $(\beta^*)'(z,\langle\vartheta,\varphi(z)\rangle)$ if $\varphi(z) \in cl(F)$ and the derivative exists, and zero otherwise.

The assumption $\Theta_{F,\beta} \neq \emptyset$ plays the role of DCQ in the F-dual problem and is weaker than the DCQ for the original problem (3). The role of PCQ in the F-primal problem for a with $J_\beta(a)$ finite is played by the assumption that $a \in ri(F)$. Under these assumptions, the standard results under PCQ and DCQ imply attainment in the F-dual problem for $a \in \mathbb{R}^d$, where each F-dual solution ϑ belongs to $\Theta_{F,\beta}$ and gives rise to the same function $f_{F,\vartheta}$. This function is referred to as the *effective F-dual solution $g^*_{F,a}$* for a.

For $a \in cn_\varphi(\mu)$ let $F(a)$ denote the unique face of $cn_\varphi(\mu)$ whose relative interior contains a.

Theorem 6. *For $a \in \mathbb{R}^d$ such that $J_\beta(a)$ is finite*

(i) the $F(a)$-dual value $K^*_{F(a),\beta}(a)$ is attained and the primal value $J_\beta(a)$ equals $\int_{\{\varphi\notin cl(F(a))\}} \beta(\cdot,0)\,d\mu + K^*_{F(a),\beta}(a)$,

(ii) the primal solution g_a exists if and only if $\Theta_{F(a),\beta}$ is not empty and the moment vector of $g^*_{F(a),a}$ exists and equals a, in which case $g_a = g^*_{F(a),a}$,

(iii) the generalized primal solution $\hat g_a$ exists if and only if $\Theta_{F(a),\beta}$ is not empty, in which case $\hat g_a = g^*_{F(a),a}$.

If J_β equals $-\infty$ at some point then J^*_β is identically $+\infty$ and the dual problems (3) bear no information on the primal ones. However, Theorem 6 makes sense since J_β can still be finite at some point a and the $F(a)$-dual problem provides complete understanding of the primal problem for this a. This situation is illustrated in [12, Example 10.7].

Definition 2. The *extension* $exn(\mathcal{F}_\beta)$ of the family $\mathcal{F}_\beta$ is defined as union of the families $\mathcal{F}_{F,\beta} = \{f_{F,\vartheta}\colon \vartheta \in \Theta_{F,\beta}\}$ over the faces F of $cn_\varphi(\mu)$.

Corollary 1. *For $a \in \mathbb{R}^d$ with $J_\beta(a)$ finite the primal solution g_a exists if and only if the families $exn(\mathcal{F}_\beta)$ and $\mathcal{G}_a$ intersect, in which case the intersection equals $\{g_a\}$.*

This corollary practically amounts to solving the equation

$$\int_Z \varphi f_{F,\vartheta}\, d\mu = a$$

over the faces F of $cn_\varphi(\mu)$ and $\vartheta \in \Theta_{F,\beta}$.

The following result seems to be the most general version of Pythagorean identity.

Theorem 7. *If $a \in \mathbb{R}^d$, $J_\beta(a)$ is finite and $\Theta_{F(a),\beta} \neq \emptyset$ then*

$$H_\beta(g) = J_\beta(a) + B_\beta(g,\hat g_a) + C_\beta(g)\,, \qquad g \in \mathcal{G}_a\,. \tag{9}$$

The generalized primal solution $\hat g_a$ in (9) is described explicitly by Theorem 6*(iii)*.

V. Bregman projections

This section is devoted to the minimization in

$$\inf_{g\in\mathcal{G}_a} B_\beta(g,h)\,, \qquad a \in \mathbb{R}^d\,, \tag{10}$$

an often emerging special case of (2).

The function h is assumed to be nonnegative, $\mathcal{Z}$-measurable and $h(z) > 0$ if $\beta'_+(z,0) = -\infty$. Then, for $t \geqslant 0$

$$(z,t) \mapsto \beta(z,t) - \beta(z,h(z)) - \beta'_{sgn(t-h(z))}(h(z))[t-h(z)]$$

is an integrand denoted by $[\beta h]$, and $g \mapsto B_\beta(g,h) = H_{[\beta h]}(g)$ is a functional of the form (1). It follows that the infimum in (10) equals $J_{[\beta h]}(a)$ and the minimization is in the framework of the primal problem (2). A (generalized) primal solution is renamed to a *(generalized) Bregman projection* of h to $\mathcal{G}_a$.

The dual problem to (10) features the function $K_{[\beta h]}$, see (4) with β replaced by $[\beta h]$. The crucial set $\Theta_{[\beta h]}$ consists of those ϑ in $dom(K_{[\beta h]})$ that satisfy

$$\langle\vartheta,\varphi(z)\rangle < \beta'(z,+\infty) - \beta'_+(z,h(z)) \quad \text{for } \mu\text{-a.a. } z \in Z\,.$$

The DCQ holds because $\vartheta = \mathbf{0}$ always belongs to $\Theta_{[\beta h]}$. The family $\mathcal{F}_{[\beta h]}$ is parameterized by $\vartheta \in \Theta_{[\beta h]}$ and consists of the functions given μ-a.e. by

$$f_{[\beta h],\vartheta}(z) \triangleq (\beta^*)'\big(z,\langle\vartheta,\varphi(z)\rangle + \beta'_{sgn(\langle\vartheta,\varphi(z)\rangle)}(z,h(z))\big)\,.$$

In particular, $f_{[\beta h],\vartheta} = h$ when $\vartheta = \mathbf{0}$.

Since $[\beta h] \geqslant 0$, the PCQ reduces to $a \in ri(dom(J_{[\beta h]}))$. Assuming $J_{[\beta h]} \not\equiv +\infty$, thus existence of a and $g \in \mathcal{G}_a$ with $B_\beta(g,h)$ finite, the relative interiors of $dom(J_{[\beta h]})$ and $cn_\varphi(\mu)$ coincide, by Theorem 5. Thus, the PCQ is equivalent to the condition $a \in ri(cn_\varphi(\mu))$, not depending on h.

Theorems 6 and 7 can be reformulated as follows. In these reformulations, in addition to restricting μ, the integrand β is replaced by $[\beta h]$, as indicated in indices. Accordingly, $(F,[\beta h])$-problems, $(F,[\beta h])$-solutions, etc., come into play.

Theorem 8. *For every $a \in dom(J_{[\beta h]})$*

(i) the $(F(a),[\beta h])$-dual value is attained and

$$J_{[\beta h]}(a) = \int_{\{\varphi \notin cl(F(a))\}} [\beta h](\cdot,0)\, d\mu + K^*_{F(a),[\beta h]}(a)\,,$$

(ii) the Bregman projection $g_{[\beta h],a}$ of h to $\mathcal{G}_a$ exists if and only if the moment vector of the effective $(F(a),[\beta h])$-dual solution $g^*_{F(a),[\beta h],a}$ exists and equals a, in which case $g_{[\beta h],a} = g^*_{F(a),[\beta h],a}$,

(iii) the generalized Bregman projection $\hat{g}_{[\beta h],a}$ of h to $\mathcal{G}_a$ exists and equals $g^*_{F(a),[\beta h],a}$.

Theorem 9. *For every $a \in dom(J_{[\beta h]})$*

$$B_\beta(g,h) = J_{[\beta h]}(a) + B_{[\beta h]}(g,\hat{g}_{[\beta h],a}) + C_{[\beta h]}(g)\,, \quad g \in \mathcal{G}_a\,.$$

A new feature here is the presence of two kinds of Bregman distances, based on β and $[\beta h]$.

Lemma 5. *For any nonnegative $\mathcal{Z}$-measurable functions $g, \tilde{g}$ on Z, $B_\beta(g,\tilde{g}) \geqslant B_{[\beta h]}(g,\tilde{g})$, with equality if $\beta(z,\cdot)$ is differentiable at $t = h(z)$ for μ-a.a. $z \in Z$ with $h(z) > 0$.*

In particular, for $\beta(z,\cdot)$ differentiable at each $t > 0$, $z \in Z$, the two Bregman distances coincide. In that case, Theorem 9 implies that if the Bregman projection $g_{[\beta h],a}$ exists, then

$$(11) \qquad B_\beta(g,h) \geqslant B_\beta(g, g_{[\beta h],a}) + B_\beta(g_{[\beta h],a}, h)\,,\ g \in \mathcal{G}_a\,.$$

If β is even essentially smooth then $h > 0$, the correction term in Theorem 9 vanishes, and (11) becomes an equality, which is the usual Pythagorean identity for Bregman distances.

If β is not differentiable, the Pythagorean inequality (11) may fail.

Example 3. Let β be autonomous, given by γ differentiable except at $t = 1$, let μ be a probability measure on $(Z,\mathcal{Z})$ and $\varphi \equiv 1$. Then, $\mathcal{G}_a$ consists of the nonnegative $\mathcal{Z}$-measurable functions whose μ-integral equals a. For the function $h \equiv 1$ and $a > 0$, the Bregman distance $B_\beta(g,h)$ equals

$$\int_Z \left[\gamma(g(z)) - \gamma(1) - \gamma'_{sgn(g(z)-1)}(1)[g(z)-1]\right] d\mu(z)\,.$$

This is minimized subject to $g \in \mathcal{G}_a$ when $g \equiv a$, by Jensen inequality. In other words, the Bregman projection $g_{[\beta h],a}$ of h to $\mathcal{G}_a$ exists and equals identically a. By a simple calculation,

$$B_\beta(g,h) < B_\beta(g, g_{[\beta h],a}) + B_\beta(g_{[\beta h],a}, h)\,, \quad g \in \mathcal{G}_a\,,$$

when $0 < a < 1$, $\int_Z \gamma(g)\, d\mu$ is finite and the set $\{g > 1\}$ is not μ-negligible.

The choice $h = f_\vartheta$, for some $\vartheta \in \Theta_\beta$, meets the running assumption and is of special interest.

Theorem 10. *If β is differentiable, $J_\beta \not\equiv +\infty$ and $\vartheta \in \Theta_\beta$ satisfies $\langle \vartheta, \varphi\rangle \geqslant \beta'_+(\cdot,0)$, μ-a.e., then $dom(J_\beta) = dom(J_{[\beta f_\vartheta]})$ and for $a \in dom(J_\beta)$*

$$B_\beta(g, f_\vartheta) = J_{[\beta f_\vartheta]}(a) + B_\beta(g,\hat{g}_a) + C_\beta(g)\,, \ \ g \in \mathcal{G}_a\,.$$

This identity implies that the generalized primal solution $\hat{g}_a$ from Theorem 6*(iii)* coincides with the generalized Bregman projection of f_ϑ to $\mathcal{G}_a$, and that the existence of a sequence g_n in $\mathcal{G}_a$ with $B_\beta(g_n, f_\vartheta) \to 0$ is sufficient for $\hat{g}_a = f_\vartheta$, extending the first assertion of Proposition 1. If β is essentially smooth then each $\vartheta \in \Theta_\beta$ meets the hypothesis of Theorem 10, but this need not be so if β is merely differentiable. For $\vartheta \in \Theta_\beta$ not meeting that hypothesis, the above assertions may fail already for the autonomous integrand β given by $\gamma(t) = t^2, t \geqslant 0$, see [12, Example 10.11].

ACKNOWLEDGEMENT

The work on this contribution was supported by the Hungarian National Foundation for Scientific Research under Grant K76088, and by Grant Agency of the Czech Republic under Grants P202/10/0618 and 201/08/0539.

REFERENCES

[1] Ali, S.M. and Silvey, S.D., A general class of coefficients of divergence of one distribution from another. *J. Roy. Statist. Soc. Ser. B* **28** (1966) 131–142.

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

[3] Borwein, J.M. and Lewis, A.S., Duality relationships for entropy-like minimization problems, *SIAM J. Control Optim.* **29** (1991) 325–338.

[4] 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.

[5] 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.

[6] 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.

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

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

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

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

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

[12] Csiszár, I. and Matúš, F., Generalized minimizers of convex integral functionals, Bregman distance, Pythagorean identities. (accepted to *Kybernetika* (2012), also on arXiv)

[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., Minimization of entropy functionals. *J. Math. Anal. Appl.* **346** (2008) 183–204.

[15] Léonard, C., Entropic projections and dominating points. *ESAIM: Probability and Statistics* **14** (2010) 343–381.

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

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

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

[19] Topsoe, F., Information-theoretical optimization techniques. *Kybernetika* **15** (1979) 8–27.

[20] Vajda, I., *Theory of Statistical Inference and Information.* Kluwer Academic Publishers, Dordrecht 1989.