Proc. Geometric Science of Information 2013, Springer LNCS 8085, 302–307.

Generalized minimizers of convex integral functionals and Pythagorean identities

Imre Csiszár[1] and František Matúš[2]
csiszar@renyi.hu, matus@utia.cas.cz

[1] A. Rényi Institute of Mathematics, Hungarian Academy of Sciences, H-1364 Budapest, P.O.Box 127, Hungary
[2] Institute of Information Theory and Automation, Academy of Sciences of the Czech Republic, 18208 Prague, P.O.Box 18, Czech Republic

Abstract. Integral functionals based on convex normal integrands are minimized subject to finitely many moment constraints. 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. The minimizers and generalized minimizers are explicitly described whenever the primal value is finite, assuming a dual constraint qualification but not the primal constraint qualification. A generalized Pythagorean identity is presented using Bregman distance and a correction term.

1 The problem

This contribution addresses minimization of *integral functionals*

$$\mathsf{H}_\beta(g) \triangleq \int_Z \beta(z, g(z))\, \mu(dz) \tag{1}$$

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*.

It is assumed throughout that 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. 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 $\gamma \in \Gamma$. More details on integrands can be found in [18, Chapter 14].

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 over g with the moment vector equal to a gives rise to the *value function*

$$\mathsf{J}_\beta(a) \triangleq \inf_{g \in \mathcal{G}_a} \mathsf{H}_\beta(g)\,, \qquad a \in \mathbb{R}^d\,. \tag{2}$$

Following [2, 3], that restrict however to autonomous integrands, let

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

where $\langle\cdot,\cdot\rangle$ denotes the scalar product and $\beta^*(z,r) = \sup_{t\in\mathbb{R}}[tr - \beta(z,t)]$ for $r \in \mathbb{R}$. The convex conjugate of the function K_β is given by

$$\mathsf{K}^*_\beta(a) = \sup_{\vartheta\in\mathbb{R}^d} \big[\langle \vartheta, a\rangle - \mathsf{K}_\beta(\vartheta))\big]\,, \qquad a \in \mathbb{R}^d\,. \tag{3}$$

The minimization/maximization in (2)/(3) is *primal/dual problem*, the infimum/supremum is *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 μ-a.e. unique. When all minimizing sequences in (2) with the finite primal value $\mathsf{J}_\beta(a)$ converge to a common limit locally in measure then the limit function is called *generalized primal solution* and is denoted by $\hat{g}_a$.

2 Main ingredients

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* $\mathsf{cn}_\varphi(\mu)$ of μ. It contains the effective domain $\mathsf{dom}(\mathsf{J}_\beta)$ of J_β. A new insight into the structure of this cone and its faces from [11, 12], building on [7], helps to recognize when the primal value is finite.

Theorem 1. *If $\mathsf{J}_\beta \not\equiv +\infty$ then $\mathsf{dom}(\mathsf{J}_\beta)$ equals union of relative interiors $\mathsf{ri}(F)$ over the faces F of $\mathsf{cn}_\varphi(\mu)$ with $\int_{\{\varphi\notin\mathsf{cl}(F)\}} \beta(\cdot,0)\, d\mu < +\infty$.*

The primal/dual problems are studied together because of their close relationship stemming from the following equality.

Theorem 2. *If $\mathsf{J}_\beta \not\equiv +\infty$ then $\mathsf{J}^*_\beta = \mathsf{K}_\beta$.*

This implication is based on [18, Theorem 14.60] but modifications in the proof are necessary. As a consequence, $\mathsf{J}^{**}_\beta = \mathsf{K}^*_\beta$ where the biconjugate J^{**}_β coincides with the value function on the relative interior of $\mathsf{dom}(\mathsf{J}_\beta)$ [17]. Correspondingly, when $a \in \mathbb{R}^d$ enjoys the *primal constraint qualification*

$$a \in \mathsf{ri}(\mathsf{dom}(\mathsf{J}_\beta)) \text{ and } \mathsf{J}_\beta(a) > -\infty \tag{PCQ}$$

interplay between the primal/dual problems have been well understood. Under PCQ, the primal and dual values coincide and a dual solution exists, reducing the primal problem to the dual one which is finite dimensional and unconstrained.

Further conclusions depend on the *dual constraint qualification*

$$\begin{aligned} &\text{there exists } \vartheta \in \mathsf{dom}(\mathsf{K}_\beta) \text{ s.t. for } \mu\text{-a.a. } z \in Z \\ &\quad \text{the function } r \mapsto \beta^*(z,r) \text{ is finite around } \langle \vartheta, \varphi(z)\rangle. \end{aligned} \tag{DCQ}$$

Under PCQ and DCQ, each dual solution ϑ is a witness for DCQ and the function g^*_a given by $z \mapsto \beta^{*\prime}(z, \langle \vartheta, \varphi(z)\rangle)$ does not depend on its choice. The primal solution exists if and only if $g^*_a \in \mathcal{G}_a$, in which case $g_a = g^*_a$.

The generalized primal solution $\hat{g}_a$ exists and equals g^*_a. In addition, *generalized Pythagorean identity* holds in the form

$$\mathsf{H}_\beta(g) = \mathsf{J}_\beta(a) + \mathsf{B}_\beta(g, g^*_a) + \mathsf{C}_\beta(g)\,, \qquad g \in \mathcal{G}_a\,. \tag{Pyth}$$

This involves the Bregman distance based on β

$$\mathsf{B}_\beta(g,h) \triangleq \int_Z \Delta_\beta(z, g(z), h(z))\, \mu(dz)$$

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 $s, t \geqslant 0$ equals

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

and $s\cdot(+\infty)$ otherwise. In the expression above, $\gamma \in \Gamma$ abbreviates $\beta(z,\cdot)$ and $\mathsf{sgn}(r)$ denotes $+$ if $r \geqslant 0$ and $-$ if $r < 0$. The identity (Pyth) contains also a new nonnegative correction functional C_β that admits explicit formula. It vanishes when β is essentially smooth in second coordinate.

If PCQ holds but DCQ fails then no sequence $g_n \in \mathcal{G}_a$ with $\mathsf{H}_\beta(g_n) \to \mathsf{J}_\beta(a)$ converges locally in measure, thus the generalized primal solution cannot exist [12, Theorem 4.17].

3 Main results

The primal problem can be attacked by the techniques presented in Section 2 also when the value $\mathsf{J}_\beta(a)$ is finite but the PCQ is not assumed, $a \notin \mathsf{ri}(\mathsf{dom}(\mathsf{J}_\beta))$. In such cases, the primal and dual values can differ and the dual problem bears often no information on the primal one.

To dispense with PCQ, the dual problem is modified, depending on the geometric position of a in the φ-cone of μ. The idea is to replace the measure μ by its restriction to the set $\{\varphi \in \mathsf{cl}(F)\}$ where F is the unique face of the φ-cone $\mathsf{cn}_\varphi(\mu)$ that contains a in its relative interior. To indicate this change of measure μ, the letter F is added to indices and labels, like in $\mathsf{K}_{F,\beta}$, $g^*_{F,a}$, F-dual, DCQ_F, etc. The modified DCQ_F is not stronger than the original one.

Theorem 3. *Let $a \in \mathbb{R}^d$ such that $\mathsf{J}_\beta(a)$ is finite, and F be the face of $\mathsf{cn}_\varphi(\mu)$ such that $a \in \mathsf{ri}(F)$.*

(i) The F-dual value $\mathsf{K}^*_{F,\beta}(a) = \sup_{\vartheta\in\mathbb{R}^d}[\langle\vartheta, a\rangle - \mathsf{K}_{F,\beta}(\vartheta)]$ is attained and the primal value $\mathsf{J}_\beta(a)$ equals $\int_{\{\varphi\notin\mathsf{cl}(F)\}} \beta(\cdot,0)\,d\mu + \mathsf{K}^*_{F,\beta}(a)$.

(ii) The primal solution g_a exists if and only if DCQ_F holds and the moment vector of $g^*_{F,a}$ exists and equals a, in which case $g_a = g^*_{F,a}$.

(iii) The generalized primal solution $\hat{g}_a$ exists if and only if DCQ_F holds, in which case $\hat{g}_a = g^*_{F,a}$.

(iv) If DCQ_F holds then

$$\mathsf{H}_\beta(g) = \mathsf{J}_\beta(a) + \mathsf{B}_\beta(g, \hat{g}_a) + \mathsf{C}_\beta(g)\,, \qquad g \in \mathcal{G}_a\,.$$

This result can be applied to study Bregman projections, thus to the minimization problems $\inf_{g\in\mathcal{G}_a} \mathsf{B}_\beta(g,h)$, $a\in\mathbb{R}^d$, where h is a given function. This works well when h is nonnegative, $\mathcal{Z}$-measurable and $h(z)>0$ if the right derivative $\beta'_+(z,\cdot)$ at 0 is $-\infty$. Pythagorean identities in this case are worked out in [12, Section 8]. When $\vartheta\in\mathbb{R}^d$ satisfies the implication in DCQ the mapping

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

defines a function f_ϑ on Z, up to a μ-negligible set. Similarly to [10], the family $\mathcal{F}_\beta$ of such functions plays the role of generalized exponential families. The Bregman projections of $h=f_\vartheta$ can be related to the original primal and dual problems.

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

Theorem 4. *Assuming the* DCQ*, for every $a\in\mathbb{R}^d$ with $\mathsf{K}^*_\beta(a)$ finite there exists a unique $\mathcal{Z}$-measurable function h_a such that for all ϑ witnessing* DCQ

$$\mathsf{K}^*_\beta(a)-\big[\langle\vartheta,a\rangle-\mathsf{K}_\beta(\vartheta)\big]\geqslant \mathsf{B}_\beta(h_a,f_\vartheta)\,.$$

The above inequality implies that whenever ϑ_n is a maximizing sequence in the dual problem, the Bregman distances $\mathsf{B}_\beta(h_a,f_{\vartheta_n})$ tend to zero, and thus f_{ϑ_n} converges to h_a locally in measure. The function h_a is regarded as *generalized dual solution* for a, extending the concept of *generalized maximum likelihood estimate* introduced in [8] and explicitly constructed in [9]. Our current proof of Theorem 4 is non-constructive, except for the case of equal primal and dual values, when h_a is equal to the generalized primal solution $\hat{g}_a$.

Space limitations do not admit a more detailed presentation of related results, discussion of assumptions, embedding to the existing literature and examples. For all these the reader is referred to the full paper [12] or an abridged version suitable for first reading [11].

4 Discussion

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 $\mathsf{H}_\beta(g)$ equal to the negative Shannon or Burg entropy or the squared L^2-norm of nonnegative g. 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 $\mathsf{H}_\beta(g)$ represents a non-metric distance of g from h. Two cases are prominent: *γ-divergence* $\int_Z h\,\gamma(g/h)d\mu$ with $\gamma\in\Gamma$ [5, 20] and *Bregman distance* [4, 13, 16]. Then the corresponding primal solution is often referred to as a *projection* of h to $\mathcal{G}_a$. The

most familiar projections correspond to the information (I-) divergence that belongs to both families of distances.

This work was preceded by [8, 10] studying the 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, 6] are studied. In [10], 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. The latter implies DCQ. 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 [1], which is powerful but requires strong regularity assumptions.

Non-autonomous integrands do not entail substantial 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.

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., Duality relationships for entropy-like minimization problems, *SIAM J. Control Optim.* **29** (1991) 325–338.
3. 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.
4. 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.
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. 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 exponential families. *Probab. Th. and Rel. Fields* **141** (2008) 213–246.
10. Csiszár, I. and Matúš, F., On minimization of entropy functionals under moment constraints. *Proc. ISIT 2008*, Toronto, Canada, 2101–2105.
11. I. Csiszár and F. Matúš (2012) Minimization of entropy functionals revisited. *Proceedings IEEE ISIT 2012*, Cambridge, MA, USA, 150–154.

12. Csiszár, I. and Matúš, F., Generalized minimizers of convex integral functionals, Bregman distance, Pythagorean identities. *Kybernetika* **48** (2012) 637–689.
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.