essentially the same convergence properties as the doubly-infinite Godard equalizer, with the advantage that those properties are not destroyed by finite truncations.

Comparing the anchored blind equalizers with energy cost functions and the Godard-type blind equalizers, we conclude the following.

a) For doubly-infinite nonrecursive equalization Godard-type cost functions are superior as they achieve global convergence, whereas the anchored equalizers may have inadmissible global minima.
b) For semi-infinite nonrecursive equalization, both equalizers achieve convergence if the channel is minimum-phase, otherwise, both equalizers have spurious local minima. The reason why the Godard equalizer suffers from this problem is the noninvertibility of the semi-infinite channel convolution matrix [5] in the nonminimum-phase case.
c) For finitely parametrized equalizers, the truncation of the Godard-type equalizers leads to ill-convergence even for the simplest channels [4], whereas the anchored equalizers based on convex cost functions inherit the convergence properties of their infinite-dimensional counterparts which can be approximated as accurately as desired. Furthermore, in contrast to the Godard-type equalizers, exact finite-dimensional implementable blind equalization is achievable by anchored blind equalizers.

REFERENCES

[1] K.J. Astrom and T. Soderstrom, "Uniqueness of the maximum likelihood estimates of the parameters of an ARMA model," *IEEE Trans. Automat. Contr.*, vol. AC-19, pp. 769–773, Dec. 1974.
[2] S. Benedetto, E. Biglieri, and V. Castellani, *Digital Transmission Theory.* Englewood Cliffs, NJ: Prentice-Hall, 1987.
[3] A. Benveniste, M. Goursat, and G. Ruget, "Robust identification of a nonminimum phase system: Blind adjustment of a linear equalizer in data communications," *IEEE Trans. Automat. Contr.*, vol. AC-25, pp. 385–399, June 1980.
[4] Z. Ding, R.A. Kennedy, B.D.O. Anderson, and C.R. Johnson, "Existence and avoidance of ill-convergence of Godard blind equalizers in data communication systems," *IEEE Trans. Commun.*, vol. 39, pp. 1313–1327, Sept. 1991.
[5] Z. Ding, C.R. Johnson, and R.A. Kennedy, "On the (non)existence of undesirable equilibria of Godard blind equalizers," *IEEE Trans. Signal Processing*, 1992, to appear.
[6] G.J. Foschini, "Equalizing without altering or detecting data," *AT&T Tech. J.*, vol. 64, pp. 1885–1911, Oct. 1985.
[7] G.B. Giannakis, "On the identifiability of non-Gaussian ARMA models using cumulants," *IEEE Trans. Automat. Contr.*, vol. 35, pp. 18–26, Jan. 1990.
[8] D.N. Godard, "Self-recovering equalization and carrier tracking in two-dimensional data communication systems," *IEEE Trans. Commun.*, vol. COM-28, pp. 1867–1875, Nov. 1980.
[9] C.R. Johnson, "Adaptive IIR filtering: Current results and open issues," *IEEE Trans. Inform. Theory,* vol. IT-30, pp. 237–250, Mar. 1984.
[10] C. R. Johnson, "Admissibility in blind adaptive equalization," *IEEE Control Syst. Mag.*, vol. 11, pp. 3–15, Jan. 1991.
[11] L. Ljung and T. Soderstrom, *Theory and Practice of Recursive Identification.* Cambridge, MA: MIT Press, 1987.
[12] R.W. Lucky, "Automatic Equalization for Digital Communications," *BSTJ*, vol. 44, pp. 547–588, Apr. 1965.
[13] M. Marden, *Geometry of Polynomials.* Providence, RI: American Mathematical Society, 1966.
[14] N. Nayeri and W.K. Jenkins, "Alternate realizations of adaptive IIR filters and properties of their performance surfaces," *IEEE Trans. Circuits Syst.*, vol. CAS-36, pp. 485–496, Apr. 1989.
[15] O. Shalvi and E. Weinstein, "New criteria for blind deconvolution of non-minimum phase systems (channels)," *IEEE Trans. Inform. Theory*, vol. 36, pp. 312–321, Mar. 1990.
[16] P. Stoica and T. Soderstrom, "Uniqueness of the maximum likelihood estimates of ARMA model parameters—An elementary proof," *IEEE Trans. Automat. Contr.*, vol. AC-27, pp. 736–738, June 1982.
[17] J.K. Tugnait, "Identification of nonminimum phase linear stochastic systems," *Automatica*, vol. 22, pp. 457–464, July 1986.
[18] S. Vembu, S. Verdú, R.A. Kennedy, and W. Sethares, "Convex cost functions in blind equalization," *Proc. 24th Conf. Inform. Sci. Syst.*, Baltimore, MD, Mar. 1991, pp. 792–797.
[19] S. Verdú, "On the selection of memoryless adaptive laws for blind equalization in binary communications," *Proc. 6th. Int. Conf. Analysis Optimization Syst.* Nice, France, June 1984, pp. 239–249.

Convergence of Best ϕ-Entropy Estimates

Marc Teboulle and Igor Vajda, *Senior Member, IEEE*

***Abstract*—Minimization problems involving ϕ-entropy functionals (a generalization of Boltzmann–Shannon entropy) are studied over a given set A and a sequence of sets A_n and the properties of their optimal solutions x_ϕ, x_n. Under certain conditions on the objective functional and the sets A and A_n, it is proven that as n increases to infinity, the optimal solution x_n converges in L_1 norm to the best ϕ-entropy estimate x_ϕ.**

***Index Terms*—Entropy functionals, norm convergence, maximum entropy methods, convex optimization, set-convergence.**

I. INTRODUCTION

Let us consider a convex, continuous function $\phi : \mathbb{R}_+ \to \mathbb{R}$ and put

$$\phi(u) = \begin{cases} \lim_{v \downarrow 0} \phi(v), & \text{if } u = 0 \\ +\infty, & \text{if } u < 0. \end{cases}$$

Denote by $\phi'_+(u)$ the (finite) right-hand derivative of ϕ at $u \in \mathbb{R}_+$. It is well known (see, Rockafellar [25]), that for any $v \in \mathbb{R}_+$ it holds

$$\phi(u) \geq \phi(v) + \phi'_+(v)(u - v), \qquad u \in \mathbb{R}. \tag{1}$$

It follows from here

$$\phi(u) \in (-\infty, \infty], \qquad u \in \mathbb{R}.$$

Consider further a finite measure space (S, μ) and the corresponding Banach spaces $L_\alpha(S, \mu), 1 \leq \alpha \leq \infty$, with norms $\|\cdot\|_\alpha$. It follows from (1) that the formula

$$I_\phi(x) := \int_S \phi(x(s)) d\mu(s) \tag{2}$$

defines a mapping $I_\phi : L_1(S, \mu) \to (-\infty, \infty]$. This mapping is called the ϕ-entropy.[1]

Manuscript received November 5, 1991; revised May 3, 1992. M. Teboulle was supported in part by AFOSR Grant 91008 and NSF Grant DMS-9201297. I. Vajda was supported by CSAS Grant 17503.

M. Teboulle is with the Department of Mathematics and Statistics, University of Maryland, Baltimore County Campus, Baltimore, MD 21228.

I. Vajda is with the Institute for Information Theory and Automation, 18208 Prague, Czechoslovakia.

IEEE Log Number 9203638.

[1] Throughout this correspondence, we use the extended real line arithmetic rules common to integration theory. These rules include e.g., $0 \cdot \infty = 0$, see e.g., Rudin [27, p. 19].

By a best ϕ-entropy estimate from a given set $A \subset L_1(S, \mu)$ we mean a solution x_ϕ of the optimization problem

$$(P_\phi) \qquad \text{minimize } I_\phi(x) \text{ subject to } x \in A.$$

Special versions of this problem play an important role in mathematical statistics, information theory and many other areas of engineering and have been extensively studied, see e.g., Csiszár [12]; Liese, Vajda [19, pp. 121–133], Vajda [29, pp. 274–297], and Kay and Marple [18]. Recently, *dual methods* for solving (P_ϕ) have also been studied under various choices and conditions for ϕ and the set A, (see e.g., Ben-Tal, Borwein, and Teboulle [4], [5], Borwein and Lewis [8], Decarreau *et al.* [15]).

A typical example of the set A is

$$A_n = \left\{ x \in L_1(S, \mu) : \int_S x a_i d_\mu = b_i \;\; i = 1, \cdots, n, x \geq 0 \right\},$$

where the a_i's are given functions in $L_\infty(S, \mu)$ and $b_1, \cdots, b_n$ are observed moments.

An important question arising with the corresponding sequence of optimization problems

$$(P_n) \qquad \text{minimize } I_\phi(x) \text{ subject to } x \in A_n$$

is then to know under what conditions on the problem's data and how, the best ϕ-entropy estimate x_n solution of (P_n) will converge as the number of given moments A_n increases. This question was recently studied in Borwein and Lewis [7]. It was proved there that if the objective functional in (P_ϕ) is chosen to be the (minus) Boltzmann–Shannon entropy and if A_n is a nested sequence of closed subsets $L_1 \supset A_1 \supset A_2 \supset \cdots$ satisfying $A = \bigcap_{n=1}^\infty A_n$, then x_n converges in L_1 norm to the best entropy estimates minimizing the Boltzmann–Shannon entropy.

In this correspondence, we extend and strengthen this result. Under an appropriate assumption (see Section IV), we prove that if A_n are nested around A in the sense that for each n either $A_n \subset A$ or $A \subset A_n$ then $\|x_n - x_\phi\|_1 \to 0$ as $n \to \infty$ and that this statement holds not only for the Boltzmann–Shannon entropy, but for the more general class of ϕ-entropy under consideration. At this juncture, we would like to thank a referee for pointing out to us the work of Visitin [30], and the more recent one of Borwein and Lewis [10], where results similar to ours have been obtained using very different techniques. After giving in Section II some examples of ϕ-entropy, existence and characterization of best ϕ-entropy estimates are studied in Section III, and in Section IV we prove our convergence result. Our proofs are self-contained and rely on basic convex and functional analysis arguments.

II. Examples of ϕ-Entropy

If $\phi(u) = u \log u, u > 0$, then the (minus) ϕ-entropy is the well-known Boltzmann–Shannon entropy of the signed measure

$$\nu(E) = \int_E x(s) d\mu(s), \qquad E \subset S \text{ measurable}.$$

To avoid the trivial infinity value one has to restrict it to $x \geq 0$. The Boltzmann–Shannon entropy of probability measures is one of the well-known fundamental concepts of information theory (see Blahut [6]). If $\phi(u) = u - u^a, u > 0$, for a fixed $a \in (0, 1)$ then

$$I_\phi(x) = \|x\|_1 - \int_S x^a(s) d\mu(s), \qquad \text{for } x \geq 0.$$

In the particular case, where $a = 1/2, S = (-\pi, \pi)$ and μ is uniform on S, the Ornstein's distance of a Gaussian random signal $X_1, X_2, \cdots$ with power 1 and spectral density $x \in L_1(S, \mu)$ from the white noise signal $Y_1, Y_2, \cdots$ with power 1 happens to be

$$\frac{I_\phi(x)}{\pi} = 2\left(1 - \frac{1}{\sqrt{2\pi}} \int_{-\pi}^{\pi} \sqrt{x(s)}\, ds\right) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left(\sqrt{x(s)} - 1\right)^2 ds$$

(cf. Gray, Neuhoff, and Shields [16]). Other examples of ϕ-entropies together with their various applications in information theory, can be found e.g., in Arimoto [1], Poor [24], Clarke and Barron [13], Longo *et al.* [22], Jones and Byrne [17], and Lin [20].

However, best known are the ϕ-entropies in the case where S is finite, $S = \{1, \cdots, m\}$, and μ is the uniform counting measure on S. Then, the ϕ-entropies

$$I_\phi(x) = \sum_{s=1}^m \phi(x(s)), \qquad x \equiv (x(1), \cdots, x(m)) \in (0, \infty)^m$$

are well-known measures of inequality of coordinates of x and $-I_\phi(x)$ measures of equality (see Marshall and Olkin [23]). The ϕ-entropies of probability vectors x, i.e., vectors with $x(i) \geq 0$,

$$\|x\|_1 = \sum_{s=1}^m x(s) = 1,$$

have been introduced independently by several authors (see e.g., Vosatka [31], Ben Bassat [2], Burbea and Rao [11], Ben-Tal and Teboulle [3]).

If we take

$$\phi_a(u) = \begin{cases} \frac{u - u^a}{1 - a}, & \text{for } a \in (0, 1), \\ u \ln u, & \text{for } a = 1, \end{cases}$$

then the following continuity takes place for every $x \equiv (x(1), \cdots, x(m)) \in (0, \infty)^m$:

$$\lim_{a \uparrow 1} I_{\phi_a}(x) = \lim_{a \uparrow 1} \frac{1}{1 - a} \left(1 - \sum_{s=1}^m x^a(s)\right) = \sum_{s=1}^m x(s) \ln x(s) = I_{\phi_a}(x).$$

Various other examples can be found in the recent work Dacunha-Castelle and Gamboa [14].

It follows from Proposition 2.14 in [19], that this approximation of Boltzmann–Shannon entropy holds for every μ and $x \in L_1(S, \mu)$ such that $\mu(S) = 1, x \geq 0, \|x\|_1 = 1$. Using an evident modification of the proof presented there one can extend this limit relation to arbitrary μ and $x \in L_1(S, \mu)$ under consideration.

Let us return to ϕ-entropies of probability vectors $x \in (0, \infty)^m$. Using the convexity of ϕ-entropy (see Lemma 1 below), one easily obtains the inequalities

$$I_\phi(x_s) \geq I_\phi(x) \geq I_\phi(x_u), \qquad s = 1, \cdots, m,$$

where $x_s \in (0, \infty)^m$ has all coordinates zero except the sth which is 1 and $x_u = (m^{-1}, m^{-1}, \cdots, m^{-1}) \in (0, \infty)^m$ is the uniform probability vector. This relation explains why $-I_\phi(x)$ is a good measure of uncertainty of a random variable distributed by x. It is often preferred to put the uncertainty of a constant to be equal zero, i.e.,

$$I_\phi(x_s) = 0, \qquad s = 1, \cdots, m.$$

This is satisfied with the normalization $\phi(0) = \phi(1) = 0$, which is verified for the functions $\phi_a, a \in (0, 1]$ previously considered.

III. Existence of Best ϕ-Entropy Estimates

We first summarize some basic functional properties of ϕ-entropy. The next Lemma could alternatively be proved using results from Rockafellar [26]. We give a self-contained proof enlightening the importance of (3). The uniform integrability on subsets of $L_1(S,\mu)$ with upperbounded ϕ-entropy, established in this proof, is directly employed in the proof of Lemma 2.

Lemma 1: The ϕ-entropy defined by (2) is a lower semicontinuous proper convex function on $L_1(S,\mu)$. If ϕ is strictly convex on $\mathbb{R}_+$ then the ϕ-entropy is strictly convex on its domain

$$\operatorname{dom} I_\phi := \{x \in L_1(S,\mu) | I_\phi(x) < \infty\}.$$

If

$$\lim_{u\to\infty} \frac{\phi(u)}{u} = \infty, \tag{3}$$

then the level sets

$$B_\alpha = \{x \in L_1(S,\mu) | I_\phi(x) \le \alpha\}, \qquad \alpha \in \mathbb{R},$$

are weakly compact.

Proof: It is clear from (2) that I_ϕ is proper convex. The lower semicontinuity will be proved if we prove that each B_α is closed. Consider a sequence $x_n \in B_\alpha, x_n \to \overline{x} \in L_1(S,\mu)$. Since $x_n \ge 0$, it follows that $x \ge 0$. Moreover, there exists a subsequence $x_{n_k} \to \overline{x}$ μ-a.s. Since ϕ is assumed to be continuous on $(0,\infty)$, this implies $\phi(x_{n_k}) \to \phi(\overline{x})$ μ-a.s. Hence, by Fatou's lemma

$$\alpha \ge \liminf_{k\to\infty} I_\phi(x_{n_k}) \ge I_\phi(\overline{x}),$$

which proves that B_α is closed.

Since B_α is convex, Mazur's theorem implies that B_α is also weakly closed (see e.g., V.1 in Yoshida [32]). This implies that I_ϕ is weakly lower semicontinuous too. Therefore, to prove the weak compactness of B_α it suffices to show the sequential compactness of B_α. To this end it suffices to show that all elements of B_α are uniformly μ-integrable. If $x \in B_\alpha$ and $\|x\|_1 > 0$, then it follows from (3) that for all sufficiently large u and for $v = \|x\|_1$,

$$\begin{aligned}\int_{\{x>u\}} x\,d\mu &= \int_{\{x>u\}} x\,\frac{\phi(x)-\phi(v)-\phi'_+(v)(x-v)}{\phi(x)-\phi(v)-\phi'_+(v)(x-v)}\,d\mu \\ &\le \int_S \left[\phi(x)-\phi(v)-\phi'_+(v)(x-v)\right] d\mu \\ &\quad\cdot \frac{u}{\phi(u)-\phi(v)-\phi'_+(v)(u-v)},\end{aligned}$$

where the inequality follows from the fact that the expression in brackets is by (1) nonnegative and also increasing in x for sufficiently large x. Since the last integral is bounded above by

$$I_\phi(x) - \phi(v) \le \alpha - \phi(v),$$

it holds

$$\begin{aligned}\lim_{u\to\infty}\int_{\{x>u\}} x\,d\mu &\le (\alpha - \phi(v)) \\ &\quad\cdot \lim_{u\to\infty} \frac{u}{\phi(u)-\phi(v)-\phi'_+(v)(u-v)} = 0.\end{aligned}$$

The convergence on the right-hand side is uniform for $v = \|x\|_1$, $x \in B_\alpha$, since the Jensen inequality implies

$$\alpha \ge \int_S \phi(x)\,d\mu \ge \mu(S)\phi\left(\frac{\|x\|_1}{\mu(S)}\right)$$

so that $\|x\|_1$ is bounded on B_α.

It remains to prove the strict convexity of I_ϕ. Since ϕ is given strictly convex on $\mathbb{R}_+$, then it holds for all x, $y \in \operatorname{dom} I_\phi$

$$\frac{\phi(x)+\phi(y)}{2} - \phi\left(\frac{x+y}{2}\right) > 0, \qquad \mu\text{ -a.s.}$$

Hence, it follows from (2),

$$\frac{I_\phi(x)+I_\phi(y)}{2} - I_\phi\left(\frac{x+y}{2}\right) > 0. \qquad \square$$

For examples of kernels ϕ satisfying (3), we refer the reader to Ben-Tal, Borwein, and Teboulle [5] and Borwein and Lewis [8].

It follows from the lower semicontinuity of I_ϕ established in Lemma 1 that for convex, compact $A \subset L_1(S,\mu)$ the best estimate exists. The following assertion goes deeper. This assertion is not new (cf. Proposition 8.5 in Liese and Vajda [19]), but its proof is short and we present it here for the sake of completeness.

Theorem 1: If $\emptyset \ne A \subset L_1(S,\mu)$ is weakly closed (e.g., convex and closed) and (3) of Lemma 1 holds then the solution x_ϕ of (P_ϕ) exists. If moreover ϕ is strictly convex on $\mathbb{R}_+$, and $I_\phi(x_\phi) < \infty$ then this solution is unique.

Proof: Put

$$\alpha = \inf_{x\in A} I_\phi(x).$$

If $\alpha = \infty$, then there is nothing to prove. Suppose $\alpha < \infty$, consider the nonempty set

$$C_\alpha = A \cap \{x \in L_1(S,\mu) | I_\phi(x) \le 2\alpha\}.$$

By Lemma 1, this set is weakly compact and $I_\phi(x)$ is weakly lower semicontinuous. Hence, the best ϕ-entropy estimate from C_α, and consequently from A, exists. Let us now assume that ϕ is strictly convex on $\mathbb{R}_+$ and consider two best ϕ-entropy estimates $x_\phi, y_\phi \in A$. If $I_\phi(x_\phi) = I_\phi(y_\phi) < \infty$, then $x_\phi, y_\phi \ge 0$ and

$$I_\phi\left(\frac{x_\phi+y_\phi}{2}\right) - \frac{I_\phi(x_\phi)+I_\phi(y_\phi)}{2} = 0.$$

Using (2), this means

$$\phi\left(\frac{x_\phi+y_\phi}{2}\right) - \frac{\phi(x_\phi)+\phi(y_\phi)}{2} = 0 \qquad \mu\text{-a.s.}$$

and the strict convexity of ϕ leads to $x_\phi = y_\phi$ μ-a.s. Therefore, the elements $x_\phi, y_\phi \in L_1(S,\mu)$ coincide. $\square$

Our next result gives a simple characterization of best ϕ-entropy estimate. It is an extension of Theorem 5a in Rüschendorf [28].

Theorem 2: Let ϕ be differentiable on $\mathbb{R}_+$ with derivative $\phi'(u), u \in \mathbb{R}_+$, and put

$$\phi'(0) = \lim_{v\downarrow 0} \frac{\phi(v)-\phi(0)}{v}, \qquad \text{where } -\infty < \phi(0) \le \infty.$$

If $A \subset \operatorname{dom} I_\phi$ is convex then $\overline{x} \in A$ solves (P_ϕ) iff

$$\int_S \phi'(\overline{x})(x-\overline{x})\,d\mu \ge 0, \qquad x \in A. \tag{4}$$

Proof: Let $\overline{x} \in A$. It follows from the assumed properties of ϕ in the domain $(0,\infty)$ that for every $x \in A$ and $\varepsilon \in (0,1]$ it holds everywhere on S

$$\phi'(\overline{x})(x-\overline{x}) \le \frac{\phi((1-\varepsilon)\overline{x}+\varepsilon x)-\phi(\overline{x})}{\varepsilon} \le \phi(x)-\phi(\overline{x}).$$

Hence, it follows from the monotone convergence theorem for integrals that

$$\{\phi'(\overline{x})(x-\overline{x}) | x \in A\} \subset L_1(S,\mu)$$

and

$$\int_S \phi'(\overline{x})(x-\overline{x})\,d\mu = \lim_{\varepsilon\downarrow 0} \frac{I_\phi((1-\varepsilon)\overline{x}+\varepsilon x) - I_\phi(\overline{x})}{\varepsilon} \tag{5}$$

$$\leq I_\phi(x) - I_\phi(\overline{x}), \qquad x \in A. \tag{6}$$

From (4) and (6), we have $\overline{x} = x_\phi$.
Conversely, let $\overline{x} = x_\phi$. Then,

$$I_\phi((1-\varepsilon)\overline{x}+\varepsilon x) - I_\phi(\overline{x}) \geq 0, \qquad x \in A,$$

and (5) implies (4). □

Example 1: Let us consider the set A to be the convex set A_n defined in the Introduction, and a function ϕ satisfying the assumptions of Theorem 2. Denote the restriction of $\phi'(u)$ on $\mathbb{R}_+$ by $\psi(u)$, assume that ψ is increasing on $\mathbb{R}_+$ and denote its inverse by ψ^{-1}. For example, if $\phi(u) = u\ln u$, then $\psi(u) = \ln u + 1$. Let us suppose that there exist real $c_1, \cdots, c_n$ such that

$$\overline{x} = \psi^{-1}\left(\sum_{i=1}^n c_i a_i\right) \tag{7}$$

belongs to A. Since a_i's are bounded, this takes place iff there exists a solution $c_1, \cdots, c_n$ of the system of equations

$$\int_S a_i \psi^{-1}\left(\sum_{j=1}^n c_j a_j\right) d\mu = b_i, \qquad i = 1, \cdots, n.$$

The positive case $\overline{x}$ given by (7) is the best ϕ-entropy estimate x_ϕ. This follows from Theorem 2 since (7) satisfies for every $x \in A$ the condition

$$\int_S x\psi(\overline{x})\,d\mu = \int_S x\sum_{i=1}^n c_i a_i d\mu = \sum_{i=1}^n c_i b_i$$
$$= \int_S \overline{x}\sum_{i=1}^n c_i a_i\,d\mu = \int_S \overline{x}\psi(\overline{x})\,d\mu.$$

If it holds

$$\int_S \phi\left(\psi^{-1}\left(\sum_{i=1}^n c_i a_i\right)\right) d\mu < \infty, \tag{8}$$

then any other solution $c_1^*, \cdots, c_n^*$ of the previously considered system of equations satisfies the relation

$$\sum_{i=1}^n (c_i - c_i^*) a_i = 0, \qquad \mu\text{-a.s.}$$

This assertion follows from the uniqueness in Theorem 1 since A_n is closed in $L_1(S,\mu)$ and that ϕ with a derivative strictly increasing on $\mathbb{R}_+$ is strictly convex on $\mathbb{R}_+$. It follows from here in particular that for a_i's linearly independent on a set of positive μ-measure there is at most one solution of the mentioned system of equations.

IV. Convergence for Approximate Optimization Constraints

Throughout this section, we consider ϕ-entropy $I_\phi(x), x \in L_1(S,\mu)$, for ϕ strictly convex on $\mathbb{R}_+$ and satisfying (3). In addition to (P_ϕ), we consider a sequence of sets $A_n \subset L_1(S,\mu)$ and a sequence of optimization problems

$$(P_n) \qquad \text{minimize } I_\phi(x) \text{ subject to } x \in A_n.$$

Assumption 1: There exist solutions x_ϕ of (P_ϕ) and x_n of (P_n) and it holds

$$\lim_{n\to\infty} I_\phi(x_n) = I_\phi(x_\phi) < \infty. \tag{9}$$

The relation (9) will be interpreted as a condition of *asymptotic approximation* of the constraint A by the constraints A_n. Of course, (9) does not necessarily mean the setwise convergence $A_n \to A$, (in the sense defined below) but the converse is sometimes true, namely: if $A_n \to A$ setwise, then (9) holds.

Example 2: Let ϕ satisfy (3) of Section III, e.g., let $\phi(u) = u\log u$ or $u^a - u, u \in \mathbb{R}_+$, where $a > 1$ (these examples satisfies the normalization $\phi(0) = \phi(1) = 0$ of Section II). If $A, A_1, A_2, \cdots$ are weakly closed (e.g., convex and closed), then the existence part of Assumption 1 holds. If moreover, $A_n \to A$ setwise, in the sense that $A_1 \supset A_2 \supset \cdots$ and

$$A = \bigcap_{n=1}^\infty A_n,$$

then it follows from the properties of ϕ-entropy established in Lemma 1 and from Theorem 3.1 in Borwein and Lewis [7] (for a complete version of this theorem, see Borwein and Lewis [9]) that (9) holds too. See also the recent work of Borwein and Lewis [10] which addresses the question of checking (9).

One of the main result of Borwein and Lewis [7] essentially is that if the constraints A, A_n satisfy the assumptions of Example 2 then the solutions x_ϕ, x_n of $(P_\phi), (P_n)$, respectively, for the Boltzmann–Shannon entropy (i.e., corresponding to the special choice $\phi(u) = u\log u$) satisfy the limit relation $\|x_n - x_\phi\|_1 \to 0$ as $n \to \infty$. We extend this result for any ϕ-entropy under consideration and when the A_n are nested around A.

Lemma 2: If Assumption 1 holds and $A_n \supset A$ for all n then $\|x_n - x_\phi\|_1 \to 0$ as $n \to \infty$.

Proof: Consider for $u, v > 0$, the difference

$$\triangle(u,v) = \frac{\phi(u)+\phi(v)}{2} - \phi\left(\frac{u+v}{2}\right) \geq 0.$$

The finiteness in (9) implies $x_n, x_\phi \geq 0$ for all sufficiently large n. Hence, for these n, it holds

$$\int_S \triangle(x_n, x_\phi)\,d\mu = \frac{I_\phi(x_n)+I_\phi(x_\phi)}{2} - I_\phi\left(\frac{x_n+x_\phi}{2}\right)$$
$$\leq \frac{I_\phi(x_\phi)-I_\phi(x_n)}{2}$$

since $A_n \supset A$ implies

$$I_\phi(x_n) \leq I_\phi\left(\frac{x_n+x_\phi}{2}\right). \tag{10}$$

Consequently,

$$\lim_{n\to\infty}\int_S \triangle(x_n, x)\,d\mu = 0.$$

On the other hand, for each fixed $c > 0$ and $0 < v \leq c$ the function $u \to \triangle(u,v)$ is nondecreasing in the domain $u > c$. To see this, consider

$$u_i = v + 2t_i, \qquad i = 1,2, t_2 > t_1 > 0,$$

and restrict ourselves to $t_2 \geq 2t_1$ (the case $t_1 < t_2 < 2t_1$ can be treated analogically). Let $\psi(u)$ be the straight line passing through the planar points $(u_i, \phi(u_i)), i = 1,2$. It holds

$$\psi(v+2t_i) = \phi(u_i), \qquad \phi(v+t_i) = \phi\left(\frac{u_i+v}{2}\right), \qquad i = 1,2.$$

The convexity of $\phi(u)$, together with $t_1 > 2t_1$, implies

$$\psi(v+t_1) \leq \phi(v+t_1), \qquad \psi(v+t_2) \geq \phi(v+t_2).$$

Since the linear function of variable $t \in \mathbb{R}$ defined by

$$L(t) = \frac{\phi(v)+\psi(v+2t)}{2} - \psi(v+t)$$

is constant on $\mathbb{R}$, it holds $L(t_1) = L(t_2)$. The desired relation $\Delta(u_1, v) \leq \Delta(u_2, v)$ follows from this equality and from the previous relations. Therefore, for each $\varepsilon, c > 0$

$$\delta(\varepsilon, c) := \inf_{\substack{0 \leq u < \infty, 0 \leq v \leq c \\ |u-v| > \varepsilon}} \Delta(u, v) = \inf_{\substack{0 \leq u, v \leq c \\ |u-v| > \varepsilon}} \Delta(u, v).$$

Moreover, by the assumed continuity and strict convexity of ϕ, we then have the right-hand infimum that is positive. Therefore, it holds

$$\begin{aligned} \mu(\{|x_n - x_\phi| > \varepsilon\}) &\leq \mu(\{|x_n - x_\phi| > \varepsilon, x_\phi \leq c\}) \\ &\quad + \mu(\{x_\phi > c\}) \leq \frac{1}{\delta(\varepsilon, c)} \int_{\{|x_n - x_\phi| > \varepsilon, x_\phi \leq c\}} \\ &\quad \Delta(x_n, x_\phi)\, d\mu + \frac{\|x_\phi\|_1}{c} \\ &\leq \frac{1}{\delta(\varepsilon, c)} \int_S \Delta(x_n, x_\phi)\, d\mu + \frac{\|x_\phi\|_1}{c}. \end{aligned}$$

Sending first $n \to \infty$ and then $c \to \infty$, we obtain

$$\lim_{n \to \infty} \mu(\{|x_n - x_\phi| > \varepsilon\}) = 0.$$

So x_n tends in μ-measure to x_ϕ. But it follows from the uniform μ-integrability of elements of B_α (cf. the proof of Lemma 1) that all but finitely many x_n are uniformly μ-integrable. Hence, by the Theorem of L_r-convergence in [21, Ch. 3, Section 9.4], which obviously holds not only for probability, but also for finite measures, the convergence in μ-measure implies the convergence in $L_1(\mu)$. □

Lemma 3: If Assumption 1 holds and $A_n \subset A$ for all n then $\|x_n - x_\phi\|_1 \to 0$ as $n \to \infty$.

Proof: We proceed as in the previous proof with the only change that, instead of (10), in this case $A_n \subset A$ implies

$$I_\phi(x_\phi) \leq I_\phi\Big(\frac{x_n + x_\phi}{2}\Big).$$

Hence, the basic inequality has in this case the form

$$\int_S \Delta(x_n, x_\phi)\, d\mu \leq \frac{I_\phi(x_n) - I_\phi(x_\phi)}{2}.$$

The rest is the same as in the previous proof. □

Theorem 3: If Assumption 1 holds and for every n either $A_n \supset A$ or $A_n \subset A$, then $\|x_n - x_\phi\|_1 \to 0$ as $n \to \infty$.

Proof: Decompose the sequence x_n into two subsequences depending on whether $A_n \supset A$ or $A_n \subset A$, respectively. Applying on one of these subsequences Lemma 2 and on the other Lemma 3 we obtain that x_ϕ is the common limit of both subsequences. □

REFERENCES

[1] S. Arimoto, "Information-theoretical considerations on estimation problems," *Inform. Contr.*, vol. 19, pp. 181–194, 1971.

[2] M. Ben Bassat, "*f*-entropies, probability of error, and feature selection," *Inform. Contr.*, vol. 39, pp. 277–292, 1978.

[3] A. Ben-Tal and M. Teboulle, "Rate distortion theory with generalized information measures via convex programming duality," *IEEE Trans. Inform. Theory*, vol. IT-32, pp. 630–641, Sept. 1986.

[4] A. Ben-Tal, J. M. Borwein, and M. Teboulle, "A dual approach to multidimensional L_p spectral estimation problems," *SIAM J. Contr. Optimizat.*, vol. 26, pp. 985–996, 1988.

[5] ——, "Spectral estimation via convex programming," in *Conf. in Honor of A. Charnes 70th Birthday*, Univ. of Texas, Austin, TX, Oct. 1987. Also in *Systems and Management Science by Extremal Methods,* F. Y. Phillips and J. J. Rousseau, Eds. Boston, MA: Kluwer Academic Publishers, 1992, ch. 18, pp. 275–289.

[6] R. E. Blahut, *Principles and Practice of Information Theory.* Reading, MA: Addison-Wesley, 1987.

[7] J. M. Borwein and A. S. Lewis, "Convergence of best entropy estimates," *SIAM J. Optimizat.*, vol. 1, pp. 191–205, 1991.

[8] ——, "Duality relationships for entropy-like minimization problems," *SIAM J. Contr. Optimizat.*, vol. 29, pp. 325–338, 1991.

[9] ——, "On the convergence of moment problems," *Trans. Amer. Math. Soc.*, vol. 325, pp. 249–271, 1991.

[10] ——, "Strong convexity and optimization," Tech. Report CORR 91–16, Univ. of Waterloo, Aug. 1991.

[11] J. Burbea and R. Rao, "On the convexity of some divergence measures based on entropy functions," *IEEE Trans. Inform. Theory*, vol. IT-28, pp. 489–495, May 1982.

[12] I. Csiszár, "*I*-divergence geometry of probability distributions and minimization problems," *Ann. Probab.*, vol. 3, pp. 146–158, 1975.

[13] B. S. Clarke and A. R. Barron, "Information–theoretic asymptotics of Bayes methods," *IEEE Trans. Inform. Theory*, vol. 36, pp. 453–471, May 1990.

[14] D. Dacunha-Castelle and F. Gamboa, "Maximum d'entropie et probleme des moments," *Annales de l'Institut Henri Poincare*, vol. 26, pp. 567–596, 1990.

[15] A. Decarreau, D. Hilhorst, C. Lemarechal, and J. Navaza, "Dual methods in entropy maximization: Application to some problems in crystallography," *SIAM J. Optimizat.*, to appear.

[16] R. M. Gray, D. L. Neuhoff, and P. C. Shields, "A generalization of Ornstein's $\bar{d}$-distance with applications to information theory," *Ann. Probab.*, vol. 3, pp. 315–328, 1975.

[17] L. K. Jones and C. L. Byrne, "Generalized entropy criteria for inverse problems, with applications to data compression, pattern classification and cluster analysis," *IEEE Trans. Inform. Theory*, vol. 36, pp. 23–30, Jan. 1990.

[18] S. M. Kay and S. L. Marple, "Spectrum Analysis: A modern perspective," *Proc. IEEE*, vol. 69, pp. 1380–1419, Nov. 1981.

[19] F. Liese and I. Vajda, *Convex Statistical Distances.* Leipzig: B. G. Teubner, 1987.

[20] J. Lin, "Divergence measures based on the Shannon entropy," *IEEE Trans. Inform. Theory* vol. 37, pp. 145–151, Jan. 1991.

[21] M. Loeve, *Probability Theory.* Princeton, NJ: Van Nostrand, 1963.

[22] M. Long, T. D. Lookabaugh, and R. M. Gray, "Quantization for decentralized hypothesis testing under communication constraints," *IEEE Trans. Inform Theory,* vol. 36, pp. 241–255, Mar. 1990.

[23] A. W. Marshall and I. Olkin, *Inequalities: Theory of Majorization and Its Applications.* New York: Academic Press, 1979.

[24] H. V. Poor, "Robust decision design using a distance criterion," *IEEE Trans. Inform. Theory,* vol. IT-26, pp. 578–587, May 1980.

[25] R. T. Rockafellar, *Convex Analysis.* Princeton, NJ: Princeton Univ. Press, 1970.

[26] ——, "Integral functionals, normal integrands and measure selections," in *Lecture Notes in Mathematics.* New York: Springer Verlag, 1976, vol. 543, pp. 157–207.

[27] W. Rudin, *Real and Complex Analysis.* New York: McGraw-Hill, 1974.

[28] L. Rüschendorf, "On the minimum discrimination information theorem," *Statist. Decisions*, Supplementary, vol. 1, 1984.

[29] I. Vajda, *Theory of Statistical Inference and Information.* Boston, MA: Kluwer, 1989.

[30] A. Visintin, "Strong convergence results related to strict convexity," *Commun. Partial Differential Equations*, vol. 9, pp. 439–466, 1984.

[31] P. Vosatka, "On some properties of conditional entropies," thesis, Fac. of Math. and Phys., Charles Univ., Prague, Czechoslovakia, 1969.

[32] R. Yoshida, *Functional Analysis.* Berlin: Springer, 1965.