Proceedings Prague Stochastics 2006, August 21-25, 2006, Prague, Czech Republic, 288-297

Generalized maximum likelihood estimates for infinite dimensional exponential families

Imre Csiszár, František Matúš

Abstract: The notion of generalized maximum likelihood estimate for finite dimensional canonically convex exponential families, studied in detail in previous works of the authors, is extended to an infinite dimensional setting. Existence of the estimate when a generalized log-likelihood function is bounded above, and a continuity property are established. Related literature and examples are discussed.

MSC 2000: 60A10, 94A17, 62B10, 62H12

Key words: Exponential family, likelihood function, maximum likelihood estimate, Kullback-Leibler divergence minimization, generalized information projection, moment constraints, marginal constraints

1 Introduction

For finite dimensional exponential families (EF's), to remedy the nonexistence of a maximum likelihood estimate (MLE) when the log-likelihood function (LLF) is bounded above but maximizers do not exist, the concept of generalized maximum likelihood estimate (GMLE) has been introduced in [8, 9]. Previous approaches considered instead MLE in an extension of the given family, for example, for EF's supported by a finite set [1, Theorem 9.16] or a countable set [3, Chapter 6]. In the former case, the MLE in the closure of a full family in Euclidean topology always exists and has reasonable properties. For finite dimensional EF's in general, the GMLE's are studied in detail in [11], also addressing their relationship to MLE's in the variation distance closure of the family. As shown there, the GMLE does not always correspond to an MLE in that closure, although it does when the latter exists.

In this contribution, MLE and GMLE are considered in an extended framework that admits also infinite dimensional EF's. This framework involves a concept of LLF depending on a linear or, more generally, concave functional φ. In the usual LLF from a sample $x_1, \dots, x_n$ this functional assigns to a function ϑ of individual observations the sample mean $\frac{1}{n}\sum_{i=1}^{n} \vartheta(x_i)$, thus is linear. In the literature, instances of maximization of LLF's in this extended sense (but not called LLF's) have arisen as dual problems to I-divergence minimization, see [6, 2, 4, 7, 14].

This work was supported by the Hungarian National Foundation for Scientific Research under Grant T46376 and by Grant Agency of Academy of Sciences of the Czech Republic under Grant IIA 100750603.

The main result of this paper, Theorem 2.1, extends [9, Theorem 6] on existence of GMLE to the general setting indicated above. As its implication, a continuity property of GMLE is also established as Theorem 2.4.

For infinite dimensional EF's, the usual LLF is often unbounded hence does not admit a GMLE. Still, a bounded LLF may be obtainable replacing φ equal to integral w.r.t. an empirical distribution, yielding a usual LLF, by φ expressible as an integral w.r.t. a probability measure (pm) that behaves better than empirical distributions. In this case a GMLE exists and perhaps even does an MLE. If that pm is a weak limit of a sequence of empirical distributions, this GMLE or MLE admits interpretation via infinite samples. A relationship of LLF maximization to I-divergence minimization, treated in Section 3, provides an independent interpretation of GMLE in infinite dimensional settings, including cases when φ is not linear.

2 Main result

Let $(X, \mathcal{X}, \mu)$ be a σ-finite measure space with positive measure μ. For an $\mathcal{X}$-measurable function $\vartheta\colon X \to \mathbb{R}$ denote

$$\mathit{\Lambda}(\vartheta) = \ln \int_X e^{\vartheta(x)}\,\mu(dx) \tag{1}$$

and observe that $\mathit{\Lambda}$ is a convex functional, due to Hölder inequality. If $\mathit{\Lambda}(\vartheta)$ is finite let Q_ϑ denote the pm with the μ-density q_ϑ given by

$$q_\vartheta(x) = e^{\vartheta(x) - \mathit{\Lambda}(\vartheta)}\,, \qquad x \in X\,. \tag{2}$$

Note that each pm which is measure-theoretically equivalent to μ equals Q_ϑ for a suitable ϑ, and $Q_\vartheta = Q_\tau$ if and only if $\vartheta - \tau$ is constant μ-a.e.

A *canonically convex* EF is a family of pm's $\mathcal{E}_\Theta = \{Q_\vartheta\colon \vartheta \in \Theta\}$ where Θ is a nonempty convex set of $\mathcal{X}$-measurable functions $\vartheta : X \to \mathbb{R}$ with $\mathit{\Lambda}(\vartheta)$ finite. In this paper, Θ always denotes a set of functions with these properties.

Let $\boldsymbol{\Phi}(\Theta)$ denote the set of all concave functionals $\varphi\colon \Theta \to [-\infty, +\infty)$, not identically $-\infty$. For $\varphi \in \boldsymbol{\Phi}(\Theta)$, the function $\vartheta \mapsto \varphi(\vartheta) - \mathit{\Lambda}(\vartheta)$ of $\vartheta \in \Theta$ is called the LLF based on φ for the family $\mathcal{E}_\Theta$. If this LLF attains its maximum at some $\vartheta^* \in \Theta$ then ϑ^* is called an MLE based on φ.

When Θ is finite dimensional it can be represented via a 'canonical statistic' $g\colon X \mapsto \mathbb{R}^d$, an $\mathcal{X}$-measurable vector-valued function whose real-valued component functions are linearly independent, as the family of mappings

$$\vartheta_{\xi,g}\colon x \mapsto \langle \xi, g(x)\rangle \tag{3}$$

indexed by ξ in a convex set $\mathit{\Xi} \subseteq \mathbb{R}^d$. Here, $\langle\cdot,\cdot\rangle$ is the scalar product on $\mathbb{R}^d$. Hence, $\mathcal{E}_\Theta$ is a canonically convex EF in the usual sense [1]. The usual (normalized) LLF from an i.i.d. sample $x_1, \ldots, x_n$ is the function $\xi \mapsto \langle \xi, a\rangle - \mathit{\Lambda}(\xi)$ of $\xi \in \mathit{\Xi}$ where $a = \frac{1}{n}\sum_{i=1}^n g(x_i)$ is the sample mean of the canonical statistic, and $\mathit{\Lambda}(\xi) = \mathit{\Lambda}(\vartheta_{\xi,g})$. An MLE from the sample is a maximizer $\xi^* \in \mathit{\Xi}$ of this LLF. Clearly, these standard concepts are covered by the above general definitions, taking the linear functional φ that assigns to each function $\vartheta \in \Theta$ its sample mean $\frac{1}{n}\sum_{i=1}^{n} \vartheta(x_i)$, and identifying the functions $\vartheta_{\xi,g}$ with the vectors ξ.

For $\varphi \in \boldsymbol{\Phi}(\Theta)$ denote

$$\Psi^*(\varphi) = \sup_{\vartheta\in\Theta} \left[\, \varphi(\vartheta) - \Lambda(\vartheta) \,\right]. \tag{4}$$

The notation is motivated by the fact that when φ is a linear functional on a linear space containing Θ then $\Psi^*(\varphi)$ is the value at φ of the convex conjugate of the convex function Ψ equal to Λ on Θ and to $+\infty$ elsewhere.

The main result of this contribution is the following theorem that provides existence and uniqueness of a pm that is a universal limit of Q_{ϑ_n} along any sequence $\vartheta_n \in \Theta$ maximizing the LLF based on φ. This pm is called the GMLE based on φ.

Theorem 2.1. *Given $\varphi \in \boldsymbol{\Phi}(\Theta)$ with $\Psi^*(\varphi)$ finite, there exists a unique pm $R^*(\varphi)$ satisfying*

$$\Psi^*(\varphi) - \left[\, \varphi(\vartheta) - \Lambda(\vartheta) \,\right] \geqslant D(R^*(\varphi)\|Q_\vartheta)\,, \qquad \vartheta \in \Theta\,. \tag{5}$$

Here, D denotes I-divergence (Kullback-Leibler information divergence or relative entropy), defined for any pm's P and Q by

$$D(P\|Q) \triangleq \begin{cases} \displaystyle\int \ln \frac{dP}{dQ}\; dP\;, & \text{if } P \ll Q, \\ +\infty\,, & \text{otherwise.} \end{cases} \tag{6}$$

An immediate consequence of Theorem 2.1 is that if an MLE $\vartheta^* \in \Theta$ based on φ exists then the GMLE $R^*(\varphi)$ is equal to Q_{ϑ^*}.

Remark 2.2. By Pinsker inequality, (5) implies that whenever

$$\vartheta_n \in \Theta \;\text{ and }\; \varphi(\vartheta_n) - \Lambda(\vartheta_n) \to \Psi^*(\varphi) < +\infty, \tag{7}$$

the pm's Q_{ϑ_n} converge to $R^*(\varphi)$ in variation distance, thus $R^*(\varphi)$ is the universal limit alluded to above. In particular, while the GMLE need not belong to the family $\mathcal{E}_\Theta$ it does belong to $\mathrm{cl}_v(\mathcal{E}_\Theta)$, the closure of $\mathcal{E}_\Theta$ in variation distance.

Theorem 2.1 is an extension of the similar result proved for finite dimensional EF's and the usual concept of LLF in [9, Theorem 6]. That result was later completed by an explicit construction of the GMLE in [11, Theorem 4.1]. The proof of Theorem 2.1 presented below is a streamlined generalization of the former proof. A simple lemma is sent forward.

Lemma 2.3. *For arbitrary $\mathcal{X}$-measurable functions ϑ and τ on X with $\Lambda(\vartheta)$ and $\Lambda(\tau)$ finite and $0 < t < 1$*

$$\begin{aligned} t\Lambda(\vartheta)+&(1-t)\Lambda(\tau) - \Lambda(t\vartheta + (1-t)\tau) \\ &= t\,D(Q_{t\vartheta+(1-t)\tau}\|Q_\vartheta) + (1-t)\,D(Q_{t\vartheta+(1-t)\tau}\|Q_\tau)\,. \end{aligned}$$

Proof. The identities follow immediately from the definitions (1), (2) and (6). □

Proof of Theorem 2.1. Consider any sequence ϑ_n satisfying (7). The convexity of Λ, the definition (4) and the concavity of φ imply

$$\begin{aligned}0 &\geqslant \Lambda\big(\tfrac{\vartheta_m+\vartheta_n}{2}\big) - \tfrac{\Lambda(\vartheta_m)+\Lambda(\vartheta_n)}{2} \geqslant \varphi\big(\tfrac{\vartheta_m+\vartheta_n}{2}\big) - \Psi^*(\varphi) - \tfrac{\Lambda(\vartheta_m)+\Lambda(\vartheta_n)}{2}\\ &\geqslant \frac{1}{2}\big[\varphi(\vartheta_m) - \Lambda(\vartheta_m) + \varphi(\vartheta_n) - \Lambda(\vartheta_n)\big] - \Psi^*(\varphi)\,.\end{aligned}$$

Thus, $\Lambda\big(\frac{\vartheta_m+\vartheta_n}{2}\big) - \frac{\Lambda(\vartheta_m)+\Lambda(\vartheta_n)}{2}$ is arbitrarily close to 0 when m and n are sufficiently large. By Lemma 2.3 and Pinsker inequality, the sequence Q_{ϑ_n} is Cauchy in variation distance, and thus converges to a pm Q^* in variation distance. This Q^* does not depend on the particular choice of the sequence ϑ_n. Indeed, it τ_n is another sequence with the property (7) then, by the result just proven, the pm's of $\mathcal{E}_\Theta$ indexed by $\vartheta_1, \tau_1, \vartheta_2, \tau_2, \ldots$ have a limit in variation distance, and hence Q_{ϑ_n} and Q_{τ_n} converge to the same limit Q^*.

It will be shown that (5) holds with Q^* playing the role of $R^*(\varphi)$. For any ϑ and τ in Θ, and $0 < t < 1$, Lemma 2.3 implies

$$\Lambda(\vartheta) + \tfrac{1-t}{t}\,\Lambda(\tau) - \tfrac{1}{t}\,\Lambda(t\vartheta + (1-t)\tau) \geqslant D(Q_{t\vartheta+(1-t)\tau}\|Q_\vartheta)\,.$$

By the definition (4) of $\Psi^*(\varphi)$, the left-hand side can be estimated above by

$$\tfrac{1}{t}\,\Psi^*(\varphi) - \tfrac{1}{t}\,\varphi(t\vartheta + (1-t)\tau) + \Lambda(\vartheta) + \tfrac{1-t}{t}\,\Lambda(\tau)$$

which is at most

$$\Psi^*(\varphi) - \varphi(\vartheta) + \Lambda(\vartheta) + \tfrac{1-t}{t}\Big[\Psi^*(\varphi) - \varphi(\tau) + \Lambda(\tau)\Big]$$

by the concavity of φ. Apply this estimate to $\vartheta \in \Theta$ fixed, $\tau = \vartheta_n$ satisfying (7) and $t = t_n$ such that $t_n \to 0$ and $\frac{1}{t_n}[\Psi^*(\varphi) - \varphi(\vartheta_n) + \Lambda(\vartheta_n)] \to 0$. It follows that

$$\Psi^*(\varphi) - \varphi(\vartheta) + \Lambda(\vartheta) \geqslant \liminf_{n\to\infty}\, D(Q_{\theta_n}\|Q_\vartheta)\,, \qquad \vartheta \in \Theta\,, \tag{8}$$

where $\theta_n = t_n\vartheta + (1-t_n)\vartheta_n$. The sequence θ_n is in Θ and if $\varphi(\vartheta)$ is finite then $\varphi(\theta_n) - \Lambda(\theta_n) \to \Psi^*(\varphi)$ which is a consequence of

$$\Psi^*(\varphi) \geqslant \varphi(\theta_n) - \Lambda(\theta_n) \geqslant t_n\big[\varphi(\vartheta) - \Lambda(\vartheta)\big] + (1-t_n)\big[\varphi(\vartheta_n) - \Lambda(\vartheta_n)\big]$$

obtained by the concavity of $\varphi - \Lambda$. Thus, θ_n satisfies (7), and hence $Q_{\theta_n} \to Q^*$ in variation distance. By the lower-semicontinuity of I-divergence, (8) implies (5) with $R^*(\varphi)$ replaced by Q^*. This proves the existence while the uniqueness is obvious from Remark 2.2. □

The explicit characterization in [11, Theorem 4.1] of GMLE's for finite dimensional EF's admitted to establish various properties of the GMLE mapping, including characterizations of domain and range. It is not apparent to what extent these results generalize to the infinite dimensional setting. A continuity result in [11, Theorem 5.5], however, depends only on the defining inequality (5) of GMLE, and its extension is straightforward, applying the previous proof verbatim.

Theorem 2.4. *If a sequence of functionals $\varphi_n \in \boldsymbol{\Phi}(\Theta)$ converges weakly to some $\varphi \in \boldsymbol{\Phi}(\Theta)$, that is, $\varphi_n(\vartheta) \to \varphi(\vartheta)$ for all $\vartheta \in \Theta$, and $\Psi^*(\varphi_n) \to \Psi^*(\varphi) < +\infty$, then $R^*(\varphi_n) \to R^*(\varphi)$ in variation distance.*

Remark 2.5. The assertion holds also for convergence of nets, by the analogous arguments. Note that in [11, Theorem 5.5] also a stronger conclusion is made, namely the rI-convergence $D(R^*(\varphi)\|R^*(\varphi_n)) \to 0$. This follows because in the finite dimensional case for pm's in the range of the GMLE mapping, convergence in variation distance implies the rI-convergence [11, Lemma 5.6]. It remains open whether this stronger conclusion holds in general.

3 GMLE and minimization of I-divergence

Instances of the problem of maximizing LLF's in the general sense have arisen as dual problems of I-divergence minimization subject to linear or somewhat more general constraints. Such minimization problems have been studied often in a more general context of minimizing an f-divergence in the sense of [5, 17], covering the I-divergence minimization with the choice $f(t) = t\ln t - t + 1$, see for example [2, 7, 14, 16]. A brief discussion of this topic follows without any attempt at completeness.

In this section it is assumed that μ is a pm. Let $\mathcal{U}$ be a linear space of real $\mathcal{X}$-measurable functions on X, $\boldsymbol{C}$ a convex set of linear functionals on $\mathcal{U}$, and let $\mathcal{S}$ denote the set of all pm's P on $(X, \mathcal{X})$ such that each $\vartheta \in \mathcal{U}$ is P-integrable and

$$\int \vartheta \, dP = \gamma(\vartheta)\,, \qquad \vartheta \in \mathcal{U}\,, \tag{9}$$

holds for some $\gamma \in \boldsymbol{C}$ where γ may depend on P. Of special interest is the case when $\boldsymbol{C}$ consists of a single linear functional. Since $\mathcal{S}$ is a convex set of pm's, results of [19, 6] apply to the minimization of $D(P\|\mu)$ subject to $P \in \mathcal{S}$: when the infimum of $D(P\|\mu)$ over $P \in \mathcal{S}$, denoted by $D(\mathcal{S}\|\mu)$, is finite, there exists a unique pm $\mathit{\Pi}$ called the generalized I-projection of μ to $\mathcal{S}$ such that

$$D(P\|\mu) \geqslant D(P\|\mathit{\Pi}) + D(\mathcal{S}\|\mu)\,, \qquad P \in \mathcal{S}\,. \tag{10}$$

In addition, $D(\mathcal{S}\|\mu)$ is attained if and only if $\mathit{\Pi}$ belongs to $\mathcal{S}$, which always takes place if $\mathcal{S}$ is variation closed. In this case, $\mathit{\Pi}$ is the unique minimizer, thus the I-projection of μ to $\mathcal{S}$.

Considering the minimization of $D(P\|\mu)$ subject to $P \in \mathcal{S}$ as a primal problem, its Lagrangian dual is to maximize the function $\vartheta \mapsto \tilde{\varphi}(\vartheta) - \mathit{\Lambda}(\vartheta)$ over $\vartheta \in \mathcal{U}$ where $\tilde{\varphi}$ is the concave functional given by

$$\tilde{\varphi}(\vartheta) = \inf_{\gamma \in \boldsymbol{C}} \gamma(\vartheta)\,, \qquad \vartheta \in \mathcal{U}\,,$$

in particular if $\boldsymbol{C}$ consists of a single linear functional γ then $\tilde{\varphi} = \gamma$. The maximization obviously restricts to the convex set $\Theta = \{\vartheta \in \mathcal{U} \colon \mathit{\Lambda}(\vartheta) < +\infty\}$, and assuming that the restriction φ of $\tilde{\varphi}$ to Θ is not identically $-\infty$, it is equivalent to the maximization of the extended LLF based on the concave functional $\varphi \in \boldsymbol{\Phi}(\Theta)$, for the EF $\mathcal{E} = \mathcal{E}_\Theta$. In general, the primal infimum is greater than or equal to the dual supremum and their difference is referred to as the *duality gap.* The following proposition estimates the gap in terms of the infimum of $D(P\|Q)$ over $P \in \mathcal{S}$ and $Q \in \mathcal{E}$, denoted by $D(\mathcal{S}\|\mathcal{E})$.

Proposition 3.1. *If $D(\mathcal{S}\|\mu)$ is finite then $D(\mathcal{S}\|\mathcal{E})$ is a lower bound of the duality gap $D(\mathcal{S}\|\mu) - \Psi^*(\varphi)$ which is tight if $\boldsymbol{C}$ consists of a single functional. In addition, if the duality gap vanishes then the generalized I-projection Π of μ to $\mathcal{S}$ equals the* GMLE $R^*(\varphi)$.

Proof. For $P \in \mathcal{S}$ with $D(P\|\mu)$ finite and $\vartheta \in \Theta$, (6) and (2) imply

$$D(P\|Q_\vartheta) = D(P\|\mu) - \int \vartheta \, dP + \Lambda(\vartheta)\,. \tag{11}$$

Since the integral equals $\gamma(\vartheta)$ for some $\gamma \in \boldsymbol{C}$ and is lower bounded by $\varphi(\vartheta)$,

$$D(P\|Q_\vartheta) \leqslant D(P\|\mu) - \big[\varphi(\vartheta) - \Lambda(\vartheta)\big]\,. \tag{12}$$

It follows that if $D(\mathcal{S}\|\mu)$ is finite then it dominates $\Psi^*(\varphi)$ and the lower bound on the duality gap obtains by limiting along $P_n \in \mathcal{S}$ with $D(P_n\|\mu) \to D(\mathcal{S}\|\mu)$ in parallel to $\vartheta_n \in \Theta$ satifying $\varphi(\vartheta_n) - \Lambda(\vartheta_n) \to \Psi^*(\varphi)$. If $\boldsymbol{C}$ consists of a single functional then (12) holds with equality, implies the equality of $D(\mathcal{S}\|Q_\vartheta)$ to $D(\mathcal{S}\|\mu) - \big[\varphi(\vartheta) - \Lambda(\vartheta)\big]$ for $\vartheta \in \Theta$, and hence the gap equals $D(\mathcal{S}\|\mathcal{E})$. If the gap vanishes, the limiting in (12) as above gives $D(P_n\|Q_{\vartheta_n}) \to 0$. The inequality (10) implies $D(P_n\|\Pi) \to 0$, and $D(R^*(\varphi)\|Q_{\vartheta_n}) \to 0$ holds by Theorem 2.1. Hence, $\Pi = R^*(\varphi)$ follows via Pinsker inequality. ☐

In the literature, conditions are available for the duality gap to vanish, and for the existence of a maximizer, resp. minimizer. Results of this kind are reviewed in the best understood finite dimensional setting for the special case when $\boldsymbol{C}$ consists of a single functional.

Formally, let $\mathcal{U}$ be the linear space consisting of all functions $\vartheta_{\xi,g}$, $\xi \in \mathbb{R}^d$, see (3), and let $\Theta = \{\vartheta_{\xi,g}\colon \Lambda(\vartheta_{\xi,g}) < +\infty\}$. Then $\mathcal{E} = \mathcal{E}_\Theta$ is a full exponential family in the usual sense with the canonical statistic $g\colon X \to \mathbb{R}^d$. The linear functionals on $\mathcal{U}$ are of form $\gamma = \gamma_a$, $a \in \mathbb{R}^d$, where $\gamma_a(\vartheta_{\xi,g}) = \langle \xi, a\rangle$. When $\boldsymbol{C} = \{\gamma_a\}$, the convex set $\mathcal{S}$ of pm's satisfying (9) consists of those pm's P that satisfy the familiar moment constraint

$$\int g \, dP = a\,. \tag{13}$$

The set $\boldsymbol{\Phi}(\Theta)$ of functionals on Θ contains the restriction of γ_a to Θ. Let $\Psi^*(a)$ and $R^*(a)$ denote $\Psi^*(\varphi)$ and $R^*(\varphi)$ for φ equal to this restriction of γ_a.

The two propositions below follow from the main results of [9], giving the role of μ to the image μ_g. They involve the concept of convex core of a measure on $\mathbb{R}^d$, which is a convex subset of the convex support of the measure, and of the extension

$\text{ext}(\mathcal{E})$ of a full exponential family $\mathcal{E}$. Details are not given here since references are made only to the following assertions.

Fact A. The convex core of μ_g contains $a \in \mathbb{R}^d$ if and only if some pm P dominated by μ satisfies (13).

Fact B. The relative interior of the convex core of μ_g contains $a \in \mathbb{R}^d$ if and only if some pm P measure-theoretically equivalent to μ satisfies (13).

Moreover, $\text{ext}(\mathcal{E})$ is a union of full exponential families over the faces of the convex core of μ_g, each family with the same canonical statistic g as $\mathcal{E}$. This extension includes $\mathcal{E}$ as the family corresponding to the core itself, and it is closed in the variation distance [10, Theorem 4] thus contains the variation closure $\text{cl}_v(\mathcal{E})$ of $\mathcal{E}$, sometimes properly.

Proposition 3.2. *For $\mathcal{S}$ equal to the set of all pm's P satisfying the moment constraint* (13)*, $D(\mathcal{S}\|\mu)$ is finite if and only if a satisfies the equivalent conditions in Fact A. Then the generalized I-projection Π of μ to $\mathcal{S}$ belongs to* $\text{ext}(\mathcal{E})$*, moreover, it belongs to $\mathcal{E}$ if and only if the equivalent conditions in Fact B hold. The I-projection of μ to $\mathcal{S}$ exists if and only if $\mathcal{S}$ intersects* $\text{ext}(\mathcal{E})$ *in which case the I-projection is the single element Π of this intersection. The maximum in the dual problem is attained (an* MLE *exists) if and only if the equivalent conditions in Fact B are satisfied.*

Proposition 3.3. *When $D(\mathcal{S}\|\mu)$ is finite, the following assertions are equivalent.*

- *(i)* $D(\mathcal{S}\|\mu) = \Psi^*(a)$ *(vanishing duality gap)*
- *(ii)* $D(\mathcal{S}\|\mathcal{E}) = 0$
- *(iii)* $\Pi = R^*(a)$
- *(iv)* Π *belongs to* $\text{cl}_v(\mathcal{E})$*.*

Implications of these propositions include the fact that if an MLE exists, thus the given $a \in \mathbb{R}^d$ satisfies the equivalent conditions in Fact B, then the generalized I-projection Π of μ to $\mathcal{S}$ equals the pm in $\mathcal{E}$ corresponding to this MLE [13]. Also, the identity *(iii)* holds for all a satisfying the equivalent conditions in Fact A if and only if $\text{ext}(\mathcal{E})$ coincides with $\text{cl}_v(\mathcal{E})$. Simple sufficient conditions for the latter are the finiteness of $\Lambda(\vartheta_{\xi,g})$ for all $\xi \in \mathbb{R}^d$, or $\text{ext}(\mathcal{E}) = \mathcal{E}$ since the extension is variation closed. The last equality holds if and only if the convex core of μ_g has no proper faces, for which a sufficient condition is that each hyperplane in $\mathbb{R}^d$ has μ_g measure zero; with the terminology of [2], that condition means that the system consisting of the d components of the vector valued function g and the constant function is pseudo-Haar. The explicit characterization of $\text{cl}_v(\mathcal{E})$ obtained in [10] permits to formulate also a necessary and sufficient condition for $\text{ext}(\mathcal{E}) = \text{cl}_v(\mathcal{E})$, not presented here because of space-demanding technicalities.

In infinite dimensional situations, less complete information is available and only two examples are presented.

Example 3.4. Let X be a locally convex topological vector space and $\mathcal{U}$ the topological dual X' of X. Recall that a Borel pm P on X has a mean in the weak sense if $a \in X$ exists such that all $\vartheta \in X'$ are P-integrable and $\int \vartheta \, dP = \vartheta(a)$. The mean is unique since X is locally convex. If $\boldsymbol{C}$ is a nonempty convex subset of X, with its elements a considered also as the linear functionals $\gamma_a \colon \vartheta \mapsto \vartheta(a)$ on $\mathcal{U} = X'$, then the set $\mathcal{S}$ given by (9) consists of all Borel pm's on X whose mean exists and belongs to $\boldsymbol{C}$. Determination of the finite infimum $D(\mathcal{S}\|\mu)$ and the generalized I-projection $\mathit{\Pi}$ of μ to $\mathcal{S}$ is an instance of the primal problem considered above. The dual problem is to maximize the LLF based on a restriction of the concave functional $\tilde{\varphi}$ whose value at $\vartheta \in \mathcal{U}$ is equal to $\inf_{a \in \boldsymbol{C}} \vartheta(a)$. In other words, the LLF is defined by

$$\vartheta \;\mapsto\; \inf_{a \in \boldsymbol{C}} \vartheta(a) - \ln \int_X e^{\vartheta(x)} \, \mu(dx)$$

for $\vartheta \in X'$ with $\mathit{\Lambda}(\vartheta)$ finite. In [6, Theorem 3] sufficient conditions for the duality gap to vanish, existence of an MLE ϑ^* and $\mathit{\Pi} = Q_{\vartheta^*}$ are presented.

Example 3.5. Let X be the product $E \times F$ of two topological spaces E and F and $\mathcal{U}$ consist of all functions of the form $(x_1, x_2) \mapsto f(x_1) + g(x_2)$ where f and g are bounded and continuous functions on E and F, respectively. Define the linear functional γ on $\mathcal{U}$ by $\gamma(f+g) = \int f \, d\alpha + \int g \, d\beta$ where α and β are given pm's on E and F. If $\boldsymbol{C} = \{\gamma\}$ then the set $\mathcal{S}$ given by (9) consists of all pm's on X whose marginals on E and F are equal to α and β. Since $\mathcal{S}$ is closed in variation distance, the I-projection exists thus the primal problem of minimizing $D(P\|\mu)$ subject to $P \in \mathcal{S}$ has the minimum, whenever $D(\mathcal{S}\|\mu)$ is finite. The dual problem is to maximize the LLF based on $\varphi = \gamma$, viz.

$$f + g \;\;\mapsto\;\; \int_E f \, d\alpha + \int_F g \, d\beta - \ln \int_X e^{f(x_1)+g(x_2)} \, \mu(dx_1, dx_2)\,,$$

over all bounded continuous functions f on E and g on F. Note that $\mathit{\Theta} = \mathcal{U}$ and the exponential family $\mathcal{E} = \mathcal{E}_{\mathit{\Theta}}$ consists of all Borel pm's on X whose μ-density is the product of two continuous functions on E and F that are positive, bounded and bounded away from zero. Typically, the maximum is not attained, nevertheless the GMLE $R^*(\varphi)$ exits whenever the LLF is bounded.

The problem of minimizing $D(P\|\mu)$ subject to marginal constraints on P has been extensively studied, in particular, the question whether the μ-density of the I-projection $\mathit{\Pi}$ can be represented as the product of some function on E and another on F, either measurable or not. This issue is not entered here, see for example [18, 12]. Only the result of [4] is recalled that, under the assumptions that E and F are completely regular, there exists a sequence of pm's $Q_n \in \mathcal{E}$ such that $D(\mathit{\Pi}\|Q_n) \to 0$. Since this implies $D(\mathcal{S}\|\mathcal{E}) = 0$, the duality gap vanishes and $\mathit{\Pi}$ equals $R^*(\varphi)$, by Proposition 3.1.

Finally, the general approach of Léonard to f-divergence minimization has to be mentioned that implies deep results even for the instance of I-divergence, see [14, 15, 16]. This approach involves minimization over more general objects than measures, consisting of an absolutely continuous component which is a measure and a singular component which is not. In this framework, minimizers typically exist (though not uniquely), and the duality gap vanishes. When a minimizer is not a measure, its absolutely continuous component is typically equal to the generalized projection in the sense of this paper.

References

[1] Barndorff-Nielsen, O., *Information and Exponential Families in Statistical Theory.* Wiley, New York, 1978.

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

[3] Brown, L.D., *Fundamentals of Statistical Exponential Families.* Inst. of Math. Statist. Lecture Notes – Monograph Series, Vol. 9, 1986.

[4] Cattiaux P. and Gamboa F., Large deviations and variational theorems for marginal problems. *Bernoulli* **5** (1999) 81–108.

[5] Csiszár I., Information-type measures of difference of probability distributions and indirect observations. *Studia Sci. Math. Hungarica* **2** (1967) 249–318.

[6] Csiszár I., Sanov property, generalized I-projection, and a conditional limit theorem. *Annals of Probability* **12** (1984) 768–793.

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

[8] Csiszár, I. and Matúš, F., Information closure of exponetial families and generalized maximum likelihood estimates, *Proceedings 2002 IEEE Int. Symp. Information Theory*, p. 434.

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

[10] Csiszár, I. and Matúš, F., Closures of exponential families, *Annals of Probability* **33** (2005) 582–600.

[11] Csiszár, I. and Matúš, F., Generalized maximum likelihood estimates for exponential families. Submitted in May 2006.

[12] Fölmer H. and Gantert N., Entropy minimization and Schrödinger processes in infinite dimensions. *Annals of Probability* **25** (1997) 901–926.

[13] Jupp, P.E. and Mardia, K.V., A note on the maximum-entropy principle. *Scand. J. Statist.* **10** (1983) 45-47.

[14] Léonard Ch., Minimizers of energy functionals. *Acta Math. Hungarica* **93** (2001) 281–325.

[15] Léonard Ch., Minimizers of energy functionals under not very integrable constraints. *Journal of Convex Analysis* **10** (2003) 63-88.

[16] Léonard Ch., Dominating points and entropic projections. Preprint 2004, available at http://www.cmap.polytechnique.fr/∼leonard/.

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

[18] Rüschendorf, L. and Thomsen, W., Note on the Schrödinger equation and I-projections. *Statist. Probab. Letters* **17** (1993) 369–375.

[19] Topsoe, F., Information theoretical optimization techniques. *Kybernetika* **15** (1979) 7–17.

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

František Matúš: Institute of Information Theory and Automation, Academy of Sciences of the Czech Republic, Pod vodárenskou věží 4, Prague, 182 08, Czech Republic, matus@utia.cas.cz