IEEE Transactions on Information Theory **49** (2003) 1474–1490

Information projections revisited

Imre Csiszár * *and František Matúš* †

Abstract. The goal of this paper is to complete results available about I-projections, reverse I-projections, and their generalized versions, with focus on linear and exponential families. Pythagorean-like identities and inequalities are revisited and generalized, and generalized maximum-likelihood (ML) estimates for exponential families are introduced. The main tool is a new concept of extension of exponential families, based on our earlier results on convex cores of measures.

1. Introduction

For two probability measures (pm's) P, Q on the same measurable space $(X, \mathcal{X})$, the information divergence (I-divergence, relative entropy) of P from Q is defined by

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

1.1. **Information projections.** The infimum of $D(P\|Q)$ for P in a set $\mathcal{S}$ of pm's is denoted by $D(\mathcal{S}\|Q)$. If a unique minimizer exists it is called the *I-projection* of Q to $\mathcal{S}$. If every sequence P_n in $\mathcal{S}$ satisfying $D(P_n\|Q) \to D(\mathcal{S}\|Q)$ converges in a specified sense to a unique pm, not necessarily in $\mathcal{S}$, this pm is called the *generalized I-projection* of Q to $\mathcal{S}$. Similarly, the infimum of $D(P\|Q)$ for Q in a set $\mathcal{S}$ of pm's is denoted by $D(P\|\mathcal{S})$, and a unique minimizer, if exists, is called the *reverse I-projection* (rI-projection) of P to $\mathcal{S}$. If every sequence Q_n in $\mathcal{S}$ satisfying $D(P\|Q_n) \to D(P\|\mathcal{S})$ converges to a unique pm, not necessarily in $\mathcal{S}$, this pm is called the *generalized rI-projection* of Q to $\mathcal{S}$.

This work was supported by the HSSS programme of ESF, by the Volkswagen-Stiftung (RiP program in Oberwolfach), by the Hungarian National Foundation for Scientific Research under Grants T 26041, T 32323 and T 40719, and by Grant Agency of Academy of Sciences of the Czech Republic under Grant A1075104. The material of this paper was presented in part at the *IEEE International Symposium on Information Theory*, Sorrento, Italy, June 25–30, 2000, and at the *IEEE International Symposium on Information Theory*, Lausanne, Switzerland, June 30-July 5, 2002.

AMS 2000 Mathematics Subject Classification. Primary 60A10, 94A17; secondary 60E10, 62B10, 62F10, 62H12.

Key words and phrases. Convex core, exponential family, I-projection, Kullback-Leibler divergence, maximum likelihood (ML), Pythagorean identity.

*I. Csiszár is with A. Rényi Institute of Mathematics, Hungarian Academy of Sciences, H-1364 Budapest, P.O.Box 127, Hungary (e-mail: `csiszar@renyi.hu`).

†F. Matúš is with Institute of Information Theory and Automation, the Academy of Sciences of the Czech Republic, Pod vodárenskou věží 4, 182 08 Prague, Czech Republic (e-mail: `matus@utia.cas.cz`).

Such projections, particularly to linear and exponential families of pm's, occur in various problems of probability and statistics. They are intimately related to large deviation theory and maximum-likelihood (ML) estimation. Previous works studying these projections include Chentsov [6], Csiszár [10, 11], Topsøe [22], etc., see the review of prior results in Subsection 1.3. Our goal here is to complete the existing theory and to show how known results generalize if certain regularity conditions (such as steepness of exponential families) are omitted. Various subtle points will be clarified, including corrections of some errors in [6], a key work on the subject. We shall also address the question how the possible nonexistence of ML estimates can be remedied, in cases when the likelihood function is bounded.

A set $\mathcal{S}$ of pm's is called *log-convex* if it contains all log-convex combinations of pairs of not mutually singular pm's in $\mathcal{S}$. These log-convex combinations are defined, for not mutually singular pm's P and Q with densities p and q with respect to a dominating measure μ, as the pm's $\overline{P^tQ^{1-t}}$ with μ-densities $p^tq^{1-t}/\int p^tq^{1-t}d\mu$, $0<t<1$. Examples of log-convex sets comprise exponential families and their extensions introduced in Section 2. Log-convex sets of mutually absolutely continuous pm's are the "geodesically convex" sets of Chentsov [6]. We are not aware of references to log-convex sets of not mutually absolutely continuous pm's.

Generalized projections exist to convex and log-convex sets of pm's due to the following theorem. Here, *I-convergence* or *rI-convergence* of a sequence of pm's R_n to a pm R means that $D(R_n\|R)\to 0$ or $D(R\|R_n)\to 0$, respectively. Each of these convergences is stronger than convergence in variation distance, due to Pinsker's inequality.

Theorem 1. *For any pm Q and convex set $\mathcal{S}$ of pm's such that $D(\mathcal{S}\|Q)$ is finite, there exists a unique pm, denoted by $\Pi_{\mathcal{S}\leftarrow Q}$, that satisfies*

$$D(P\|Q)\geqslant D(P\|\Pi_{\mathcal{S}\leftarrow Q})+D(\mathcal{S}\|Q)\,,\qquad P\in\mathcal{S}. \tag{1}$$

This $\Pi_{\mathcal{S}\leftarrow Q}$ is the generalized I-projection of Q to $\mathcal{S}$: every sequence P_n in $\mathcal{S}$ satisfying $D(P_n\|Q)\to D(\mathcal{S}\|Q)$ I-converges to $\Pi_{\mathcal{S}\leftarrow Q}$.

For any pm P and log-convex set $\mathcal{S}$ of pm's such that $D(P\|\mathcal{S})$ is finite, there exists a unique pm, denoted by $\Pi_{P\to\mathcal{S}}$, that satisfies

$$D(P\|Q)\geqslant D(P\|\mathcal{S})+D(\Pi_{P\to\mathcal{S}}\|Q)\,,\qquad Q\in\mathcal{S}. \tag{2}$$

This $\Pi_{P\to\mathcal{S}}$ is the generalized rI-projection of P to $\mathcal{S}$: every sequence Q_n in $\mathcal{S}$ satisfying $D(P\|Q_n)\to D(P\|\mathcal{S})$ rI-convergences to $\Pi_{P\to\mathcal{S}}$.

The "convex" part of Theorem 1 was proved by Topsøe [22, Theorem 8], using a refinement of a geometric idea of Csiszár [10]. The "log-convex part" of Theorem 1 is new but can be proved by the same technique, see Appendix A. A key ingredient is the identity

$$tD(P\|Q)+(1-t)D(P\|R)=D(P\|\overline{Q^tR^{1-t}})-\ln\int q^tr^{1-t}\,d\mu \tag{3}$$

(valid for $0<t<1$, any pm P, and not mutually singular pm's Q and R with densities q and r w.r.t. a dominating measure μ) combined with its specialization

$$-\ln\int q^tr^{1-t}\,d\mu=tD(\overline{Q^tR^{1-t}}\|Q)+(1-t)D(\overline{Q^tR^{1-t}}\|R) \tag{4}$$

obtained by taking $P = \overline{Q^t R^{1-t}}$ in (3). Note that substituting (4) into (3) yields a log-convex analogue of the parallelogram identity of Euclidean geometry, the special case $m = 1$ of [6, Lemma 20.5, p. 296].

Remark 1. As a direct consequence of (1) and (2), if the minimum of $D(P\|Q)$ subject to $P \in \mathcal{S}$ or $Q \in \mathcal{S}$ is attained, the minimizer is unique and equals $\Pi_{\mathcal{S}\leftarrow Q}$ or $\Pi_{P\to\mathcal{S}}$, respectively. Actually, the minimum is attained, thus the I- or rI-projection exists, if and only if $\Pi_{\mathcal{S}\leftarrow Q}$ respectively $\Pi_{P\to\mathcal{S}}$ belongs to $\mathcal{S}$ (and then the generalized projection is a true projection). The "if" part follows from Theorem 1 because the lower semicontinuity of I-divergence implies

$$D(\mathcal{S}\|Q) \geqslant D(\Pi_{\mathcal{S}\leftarrow Q}\|Q)\ , \qquad \text{and} \qquad D(P\|\mathcal{S}) \geqslant D(P\|\Pi_{P\to\mathcal{S}})\ .$$

Remark 2. Inequality (1) shows that the generalized I-projection $\Pi_{\mathcal{S}\leftarrow Q}$ belongs to the *I-closure* of $\mathcal{S}$ defined as $\mathit{cl}_I(\mathcal{S}) = \{R : D(\mathcal{S}\|R) = 0\}$. Similarly, inequality (2) entails that the generalized rI-projection $\Pi_{P\to\mathcal{S}}$ belongs to the *reverse I-closure* (rI-closure) of $\mathcal{S}$ defined as $\mathit{cl}_{rI}(\mathcal{S}) = \{R : D(R\|\mathcal{S}) = 0\}$. Calling a set of pm's *I-closed* or *rI-closed* if it equals its own I- or rI-closure, it follows that under the conditions of Theorem 1, I- or rI-projections to $\mathcal{S}$ always exist if $\mathcal{S}$ is I- or rI-closed (in particular, if $\mathcal{S}$ is variation closed). It should be noted that the generalized I-projection $\Pi_{\mathcal{S}\leftarrow Q}$ can be different from $\Pi_{\mathit{cl}_I(\mathcal{S})\leftarrow Q}$ even if $\mathcal{S}$ is a linear family of pm's, see Example 3 in Section 7. On the other hand, if $\mathcal{S}$ is an exponential family, generalized rI-projections to $\mathcal{S}$ equal true rI-projections to $\mathit{cl}_{rI}(\mathcal{S})$, provided the projected pm has a mean, see Corollary 9 in Section 5.

1.2. **Structure of the paper.** We will work mostly with pm's on $\mathbb{R}^d$, with linear families $\mathcal{L} = \mathcal{L}_a$ of pm's that have a given mean $a \in \mathbb{R}^d$, and with standard, full exponential families $\mathcal{E}$, cf. [4]. More general situations can be reduced to this one and are postponed to Section 8.

The main new tools in this paper are convex cores of measures we have introduced in [13], and extensions of exponential families whose definition relies upon the geometric concept of face of a convex core. This extension concept is defined, and its basic properties established, in Section 2. Some simple auxiliary results are collected in Section 3.

For linear and exponential families both parts of Theorem 1 can be improved, and in a sense merged together. Section 4 elaborates upon the first "convex part" of Theorem 1 when $\mathcal{S} = \mathcal{L}$ is a linear family. The main result there, Theorem 3, is a general form of the "Pythagorean Theorem for I-divergences" that requires no regularity conditions other than an obvious finiteness assumption. This neat general form relies substantially on the concept of extension of an exponential family. We note that the proof of Theorem 3 does not actually use Theorem 1.

Section 5 is devoted to the second "log-convex part" of Theorem 1 when $\mathcal{S} = \mathcal{E}$ is an exponential family. An essential role is played by the fact that the rI-closure of $\mathcal{E}$ is always contained, perhaps strictly, in the extension of $\mathcal{E}$, a consequence of Theorem 2 in Section 2. In special cases, results of Sections 4 and 5 overlap, see Theorem 5 in Section 5.

ML estimation in exponential families is closely related to rI-projections. Section 6 addresses this subject, including the question how nonexistence of an ML estimate can be remedied. Here, the "log-convex" part of Theorem 1 and a variation on it play a key role.

In Section 7 we present three examples illustrating our results and showing that certain "irregular" cases may in fact occur.

The straightforward extensions of our results to more general linear and exponential families on arbitrary measurable spaces are discussed in Section 8. As an example, we work out in detail perhaps surprising implications of our results for exponential families on $\mathbb{R}$ with directional statistic $f(x) = (x, x^2, \ldots, x^d)$. Proofs are postponed to Appendix C.

Appendix A contains the proofs of two key theorems not proved in the text and the completion of another proof. In Appendix B, auxiliary results additional to those in Section 3 are presented.

Since several previous results on convex cores are needed in this paper, a friendly introduction into the topic in Section 2 is complemented by Appendix D containing those assertions of [13] that are used throughout in proofs. Full familiarity with [13] is not a prerequisite for understanding of this paper.

1.3. **Previous results.** The I- and rI-convergences of pm's are special cases of more general convergence concepts studied in Csiszár [9]. By the results there, "information neighbourhoods" of the form $\{Q : D(Q\|P) < \epsilon\}$ or $\{Q : D(P\|Q) < \epsilon\}$ do not define a topology for pm's. In particular, I/rI-closures are not topological closure operations, that is, I/rI-closures of sets of pm's need not be I/rI-closed. Previously, Csiszár [8] showed that, even for pm's on a countable set X, no topology exists in which the convergence of nets were equivalent to their I-convergence. Recently, however, Harremoës [15] showed that a topology for pm's does exist in which the convergence of sequences is equivalent to their I-convergence, and I-closures equal sequential closures in that topology. For rI-convergence, the situation is similar.

It has been known for a long time that I-divergence admits a "geometric" interpretation as an analogue of squared Euclidean distance. The first appearance of the "Pythagorean theorem for I-divergences" we are aware of is in Chentsov [5], see also the collection [7, pp. 218-225]. There, implicitly, (10) below (with $\mathit{ext}(\mathcal{E})$ replaced by $\mathcal{E}$) is established when $\mathcal{L} \cap \mathcal{E} \neq \emptyset$. A version of the Pythagorean identity, as in Corollary 7, not requiring $\mathcal{L} \cap \mathcal{E} \neq \emptyset$, goes back to Csiszár [10]. A first systematic development of I-divergence geometry, in particular, the study of information-theoretic projections and closures, appears in Chentsov's book [6]. There, as distinct from our work, differential geometric ideas also play an essential role. For further developments of the differential geometric approach see Amari [1] and the references therein.

The existence of generalized I-projections to convex sets of pm's is implicit in [10]; the full first part of Theorem 1 is due to Topsøe [22]. For its extensions to more general measures of distance see Csiszár [12]. Generalized rI-projections to convex (rather than log-convex) sets of pm's appear in Barron [3]. These, in accordance with [12], may be of total mass less than 1, unlike the generalized projections in Theorem 1. Special cases of our general Pythagorean theorem (Theorem 3) and its corollaries, other than the simple ones already mentioned, appear explicitly or implicitly in Jupp and Mardia [17] and Csiszár [11]. In particular, the generalized I-projection of a pm Q to a linear family $\mathcal{L}$ was shown in [11, Theorem 2] to belong to an exponential family based on a suitable restriction of Q; the proof there admits (in retrospect, using results of [13]) identification of that family as a component of the extension of $\mathcal{E}$.

A concept of extension of exponential families, similar to but not the same as ours, was devised by Chentsov [6, p. 315]. He extended the family $\mathcal{E}$ based on a measure μ by a "boundary at infinity" that consisted of members of exponential families based on restrictions of μ to "ponderable faces" of the convex support of μ. However, some crucial assertions in [6] about an exponential family completed by this boundary, such as Lemma 23.7 (rI-closedness) and Theorem 23.3 (existence of rI-projections), are false, see Example 1 in Section 7, or [13, Example 3]. The reason is that the boundary at infinity in [6] is too small (the ponderable faces of the convex support correspond to the exposed faces of the convex core, see [13, Lemma 11]). Extensions of exponential families were also considered by Barndorff-Nielsen [2, pp. 154-5] and Brown [4, pp. 191-201]; the former for μ with finite support, the latter for μ with at most countable support satisfying additional assumptions. These two extensions are special cases of our definition.

The information-theoretic view of ML estimation in exponential families goes back to Kullback [16], see also Chentsov [6]. A first connection with generalized I-projections was made by Jupp and Mardia [17]. The focus in our paper is on the situation when no ML estimate exists. For this case, the question "whether it is possible to enlarge the family in a natural way such that the ML estimate becomes defined with probability one" was answered in the affirmative by Barndorff-Nielsen [2, pp. 154-5] when μ has finite support.

Though not directly related to our work, we mention that the limiting behaviour of the variance function of a steep exponential family has been studied by Masmoudi [18]. For recent results on ML estimation in multidimensional exponential families we refer to Miao and Hahn [19]. The readers' attention is drawn also to Morozova and Chentsov [20] and Chentsov's collected works [7].

2. Extensions of exponential families

From now on, all measures we consider are finite Borel measures on $\mathbb{R}^d$, with a few exceptions stated explicitly.

The *mean* $a \in \mathbb{R}^d$ of a pm P is the coordinatewise integral $\int x\, P(dx)$, provided that each coordinate function x_i is P-integrable; otherwise, P does not have a mean. A *linear family* of pm's is defined as

$$\mathcal{L} = \mathcal{L}_a = \{\text{all pm's with a mean } a\}$$

where $a \in \mathbb{R}^d$ is the *mean* of $\mathcal{L}$.

The (standard, full) *exponential family* $\mathcal{E}$ based on a nonzero measure μ on $\mathbb{R}^d$ is the set of all pm's Q equivalent to μ such that $\ln \frac{dQ}{d\mu}$ is an affine function. In parametric representation

$$\mathcal{E} = \mathcal{E}_\mu = \left\{ Q_\vartheta \,:\, \frac{dQ_\vartheta}{d\mu}(x) = e^{\langle \vartheta, x\rangle - \Lambda(\vartheta)} \,,\ \vartheta \in \mathsf{dom}(\Lambda) \right\}$$

where

$$\Lambda(\vartheta) = \Lambda_\mu(\vartheta) = \ln \int_{\mathbb{R}^d} e^{\langle \vartheta, x\rangle}\, \mu(dx)$$

and

$$\mathsf{dom}(\Lambda) = \{\vartheta \,:\, \Lambda(\vartheta) < \infty\} \,.$$

Note that μ is permitted to be concentrated on an affine subspace of $\mathbb{R}^d$, thus the above parametrization need not be one-to-one. Also, no assumptions are made on the richness of $\mathit{dom}(\Lambda)$. In the literature, the finiteness of the underlying measure μ is rarely required; our finiteness assumption does not restrict generality but excludes the trivial case $\mathcal{E} = \emptyset$. It could even be assumed that μ is a pm, since clearly $\mathcal{E} = \mathcal{E}_Q$ for each $Q \in \mathcal{E}$.

Exponential families are log-convex, the log-convex combinations of Q_ϑ and Q_τ are the pm's $Q_{t\vartheta+(1-t)\tau}$, $0 < t < 1$.

The *convex core* $\mathit{cc}(\mu)$ of a measure μ is the intersection of all convex Borel sets of full μ-measure. Equivalently, $\mathit{cc}(\mu)$ is the set of means of pm's dominated by μ [13, Theorem 3] (for this, and other references to [13], see Appendix D). By [13, Lemma 1], the closure of $\mathit{cc}(\mu)$ equals the well-known *convex support* $\mathit{cs}(\mu)$ of μ, the intersection of all convex closed sets of full μ-measure, see [6], [2]. Therefore, the relative interiors of $\mathit{cc}(\mu)$ and $\mathit{cs}(\mu)$ coincide. Note that $\mathit{cc}(\mu)$, unlike $\mathit{cs}(\mu)$, need not be of full μ-measure.

For the exponential family $\mathcal{E}$ based on μ, each $Q \in \mathcal{E}$ has the same convex core as μ. Hence we write $\mathit{cc}(\mathcal{E})$ for $\mathit{cc}(\mu)$, and speak about the convex core of $\mathcal{E}$; we do the same for convex support.

A *face* of a convex set C in $\mathbb{R}^d$ is a convex set $F \subseteq C$ that contains each line segment $ab = \{ta+(1-t)b\colon 0 \leqslant t \leqslant 1\}$ in C that has an interior point $ta+(1-t)b$, $0 < t < 1$, in F. In this paper we exclude the trivial face $F = \emptyset$ and by face we always mean a nonempty face. The *proper faces* are those distinct from C itself. The intersection of C with one of its supporting hyperplanes is always a face; these, and C itself, are called *exposed faces*. Each proper face of C is a subset of a proper exposed face [21, Theorem 11.6]. Each point in C belongs to the relative interior $\mathit{ri}(F)$ of exactly one face F of C [21, Theorem 18.2].

Given an exponential family $\mathcal{E} = \mathcal{E}_\mu$, and a face F of $\mathit{cc}(\mathcal{E}) = \mathit{cc}(\mu)$, let $\mathcal{E}^F$ denote the exponential family based on $\mu^{\mathit{cl}(F)}$, the restriction of μ to the closure of F. Since $\mu^{\mathit{cl}(F)}$ has the convex core F [13, Lemma 3], $F \neq \emptyset$ implies that $\mu^{\mathit{cl}(F)}$ is a nonzero measure. Thus, $\mathcal{E}^F$ is well defined, $\mathit{cc}(\mathcal{E}^F) = F$ and by [13, Lemma 3], $\mathit{cs}(\mathcal{E}^F) = \mathit{cl}(F)$. A member $Q_{F,\vartheta}$ of $\mathcal{E}^F$ has μ-density equal to $e^{\langle\vartheta,x\rangle - \Lambda_F(\vartheta)}$ on $\mathit{cl}(F)$ and 0 otherwise. Here,

$$\Lambda_F(\vartheta) = \ln \int_{\mathit{cl}(F)} e^{\langle\vartheta,x\rangle}\,\mu(dx)$$

and ϑ belongs to $\mathit{dom}(\Lambda_F) = \{\vartheta\colon \Lambda_F(\vartheta) < \infty\}$. Clearly, $\mathcal{E}^F$ contains all pm's $Q(\cdot|\mathit{cl}(F))$ obtained by normalizing the restrictions $Q^{\mathit{cl}(F)}$, $Q \in \mathcal{E}$; they exhaust $\mathcal{E}^F$ if $\mathit{dom}(\Lambda) = \mathbb{R}^d$ but in general they do not.

The family $\mathcal{E}^F = \mathcal{E}_{\mu^{\mathit{cl}(F)}}$ is a *component* in the disjoint union

$$\mathit{ext}(\mathcal{E}) = \bigcup \left\{\mathcal{E}^F\colon F \text{ face of } \mathit{cc}(\mathcal{E})\right\}$$

to be called the *extension* of $\mathcal{E}$. The union is at most countable since $\mathit{cc}(\mu)$ has at most a countable number of faces [13, Theorem 1]. A pm $Q \in \mathit{ext}(\mathcal{E})$ belongs to the component $\mathcal{E}^F$ with $F = \mathit{cc}(Q)$.

Remark 3. Clearly $\mathcal{E} \subseteq \mathit{ext}(\mathcal{E})$. The equality holds if and only if $\mathit{cc}(\mathcal{E})$ is relatively open (it has no proper faces), or equivalently, if every nontrivial supporting hyperplane of $\mathit{cs}(\mu)$ has μ-measure zero [13, Corollary 2].

Theorem 2. *The extension $\mathit{ext}(\mathcal{E})$ of an exponential family $\mathcal{E}$ is log-convex and rI-closed.*

The proof will be given in Appendix A.

Corollary 1. *For each pm P and exponential family $\mathcal{E}$ such that $D(P\|\mathsf{ext}(\mathcal{E}))$ is finite, the rI-projection of P to $\mathcal{S} = \mathsf{ext}(\mathcal{E})$ exists, and equals $\Pi_{P\to\mathcal{S}}$ of Theorem 1.*

Corollary 2. *For an exponential family $\mathcal{E}$ the following assertions are equivalent:*

(i) $\mathcal{E} = \mathsf{ext}(\mathcal{E})$;

(ii) every pm P with $D(P\|\mathcal{E})$ finite has rI-projection to $\mathcal{E}$;

(iii) $\mathcal{E}$ is rI-closed.

Proofs of Corollaries 1 and 2. Corollary 1 is immediate from Theorems 1, 2 and Remark 2, and it gives *(i)⇒(ii)* in Corollary 2. As for *(ii)⇒(iii)*, note that no pm P in $\mathsf{cl}_{rI}(\mathcal{E}) \setminus \mathcal{E}$ can have rI-projection to $\mathcal{E}$.

The implication *(iii)⇒(i)* is a consequence of Lemma 7*(i)*, proved in Appendix B, stating that for each $Q \in \mathcal{E}$ and exposed face F of $\mathsf{cc}(\mathcal{E})$, the pm $Q(\cdot|\mathsf{cl}(F)) \in \mathcal{E}^F$ belongs to $\mathsf{cl}_{rI}(\mathcal{E})$. Since $\mathcal{E}^F$ and $\mathcal{E}$ are disjoint when $F \neq \mathsf{cc}(\mathcal{E})$, the rI-closedness of $\mathcal{E}$ implies that $\mathsf{cc}(\mathcal{E})$ has no proper exposed face. Then, in turn, $\mathsf{cc}(\mathcal{E})$ has no proper face and $\mathcal{E} = \mathsf{ext}(\mathcal{E})$ follows. □

3. Auxiliary results

For a linear family $\mathcal{L}$ and an exponential family $\mathcal{E}$ we denote by $D(\mathcal{L}\|\mathcal{E})$ the infimum of $D(P\|Q)$ subject to $P \in \mathcal{L}$ and $Q \in \mathcal{E}$. Similar notation will be used also for other sets of pm's.

In auxiliary calculations, the quantities $D(P\|\mu)$ and $D(\mathcal{L}\|\mu)$ will be considered also for an arbitrary (positive, finite) measure μ, defined in the same way as if μ were a pm. Then a lower bound to $D(P\|\mu)$ is $-\ln \mu(\mathbb{R}^d)$ rather than 0. We will work also with the (convex) function H defined by

$$H(a) = H_\mu(a) = D(\mathcal{L}_a\|\mu) \ , \qquad a \in \mathbb{R}^d \ . \tag{5}$$

The following lemma is included for reference purposes. It follows directly from [13, Theorem 3], stating that the means of pm's $P \ll \mu$ are in $\mathsf{cc}(\mu)$, and each point in $\mathsf{cc}(\mu)$ is the mean of some $P \ll \mu$ having bounded μ-density.

Lemma 1. *For a linear family $\mathcal{L}$ and finite measure μ, $D(\mathcal{L}\|\mu)$ is finite if and only if the mean of $\mathcal{L}$ is in $\mathsf{cc}(\mu)$. In other words, $\mathsf{dom}(H) = \{a : H(a) < \infty\}$ is equal to $\mathsf{cc}(\mu)$.*

Corollary 3. *For a linear family $\mathcal{L}$ and exponential family $\mathcal{E}$, $D(\mathcal{L}\|\mathcal{E})$ is finite if and only if the mean of $\mathcal{L}$ is in $\mathsf{cc}(\mathcal{E})$. The last condition is necessary (and sufficient) also for the finiteness of $D(\mathcal{L}\|\mathsf{ext}(\mathcal{E}))$.*

Proof. By Lemma 1, $D(\mathcal{L}\|Q)$ is finite if and only if the mean of $\mathcal{L}$ is in $\mathsf{cc}(Q)$. The first assertion follows since $\mathsf{cc}(Q) = \mathsf{cc}(\mathcal{E})$ for each $Q \in \mathcal{E}$, and the second assertion follows since $Q \in \mathsf{ext}(\mathcal{E})$ implies $\mathsf{cc}(Q) \subseteq \mathsf{cc}(\mathcal{E})$. □

Lemma 2. *For any pm P with mean a and exponential family $\mathcal{E} = \mathcal{E}_\mu$*

$$D(P\|Q_\vartheta) = D(P\|\mu) - \langle \vartheta, a\rangle + \Lambda(\vartheta) \ , \qquad Q_\vartheta \in \mathcal{E} \ . \tag{6}$$

Further, if the mean a of P belongs to a face F of $\mathsf{cc}(\mathcal{E})$ then

$$D(P\|Q_{F,\vartheta}) = D(P\|\mu) - \langle \vartheta, a\rangle + \Lambda_F(\vartheta) \ , \qquad Q_{F,\vartheta} \in \mathcal{E}^F \ . \tag{7}$$

Proof. In the nontrivial case $P \ll \mu$, (6) follows by rewriting the left-hand side as

$$\int \ln\left[\frac{dP}{d\mu}\Big/\frac{dQ_\vartheta}{d\mu}\right] dP = D(P\|\mu) - \int \left[\, \langle \vartheta, x\rangle - \Lambda(\vartheta) \,\right] P(dx)\ .$$

Applying (6) to the exponential family $\mathcal{E}^F$, one obtains (7) with $D(P\|\mu)$ replaced by $D(P\|\mu^{\mathit{cl}(F)})$. Since for $P \ll \mu$ with mean in F we have $P \ll \mu^{\mathit{cl}(F)}$ [13, Corollary 6], this replacement does not affect the value of the divergence. ☐

The function Λ is convex by Hölder's inequality. Recall that the convex conjugate f^* of a convex function f on $\mathbb{R}^d$ is defined by

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

(see [21, Section 12]).

Corollary 4. *For each $a \in \mathbb{R}^d$ and $\vartheta \in \mathit{dom}(\Lambda)$*

$$H(a) = D(\mathcal{L}_a\|Q_\vartheta) + \langle \vartheta, a\rangle - \Lambda(\vartheta) = D(\mathcal{L}_a\|\mathcal{E}) + \Lambda^*(a)\ .$$

Proof. Minimization over $P \in \mathcal{L}_a$ in (6) and a trivial rearrangement yield the first equality. If $H(a)$ is finite, it follows that when $D(\mathcal{L}_a\|Q_\vartheta)$ approaches its infimum $D(\mathcal{L}_a\|\mathcal{E})$ subject to $\vartheta \in \mathit{dom}(\Lambda)$ then $\langle \vartheta, a\rangle - \Lambda(\vartheta)$ approaches its supremum $\Lambda^*(a)$; thus the second equality follows. If $H(a) = +\infty$ then $D(\mathcal{L}_a\|Q_\vartheta) = +\infty$ for each $\vartheta \in \mathit{dom}(\Lambda)$, by the first equality (or Lemma 1), and the second equality holds trivially. ☐

The following lemma is less trivial. It is presumably known to experts though we have no direct reference; a related result is [14, Lemma 6.2.13].

Lemma 3. $H^* = \Lambda$.

Proof. For each $\vartheta \in \mathbb{R}^d$ and pm P that has a mean

$$H^*(\vartheta) = \sup_{a\in\mathbb{R}^d} \left[\, \langle \vartheta, a\rangle - H(a) \,\right] \geqslant \int \langle \vartheta, x\rangle\, dP - D(P\|\mu)\ .$$

Letting P be the member with parameter ϑ of the exponential family $\mathcal{E}_{\mu^B}$, based on the restriction of μ to a large ball $B = \{x : \|x\| \leqslant r\} \subset \mathbb{R}^d$, this lower bound is $\ln \int_B e^{\langle \vartheta, x\rangle}\, d\mu$, by trivial calculation. As the ball B can be arbitrarily large, $H^*(\vartheta) \geqslant \Lambda(\vartheta)$ follows.

The opposite inequality $\Lambda(\vartheta) \geqslant H^*(\vartheta)$ is immediate from Corollary 4 that implies $\Lambda(\vartheta) \geqslant \langle \vartheta, a\rangle - H(a)$ for all $a \in \mathbb{R}^d$. ☐

The following lemma extends [2, Theorem 9.13] which is stated for Λ^* only, under the hypotheses that $\mathit{cs}(\mathcal{E})$ and $\mathit{dom}(\Lambda)$ are full dimensional.

Lemma 4. *(i) If $a \in \mathit{ri}(\mathit{cc}(\mu)) = \mathit{ri}(\mathit{cs}(\mu))$ then there exists $\vartheta \in \mathbb{R}^d$ with*

$$\langle \vartheta, a\rangle - \Lambda(\vartheta) = \Lambda^*(a) = H(a).$$

(ii) If $a \notin \mathit{ri}(\mathit{cs}(\mu))$ then for each $\vartheta \in \mathbb{R}^d$

$$\langle \vartheta, a\rangle - \Lambda(\vartheta) < \Lambda^*(a).$$

Proof. *(i)* For a proper convex function f on $\mathbb{R}^d$ and its convex conjugate f^*, the equality $f(x)+f^*(x^*) = \langle x, x^* \rangle$ holds if and only if x^* belongs to the subdifferential $\partial f(x)$ [21, Theorem 23.5 *(a)(d)*], which is a nonempty subset of $\mathbb{R}^d$ if $x \in \mathit{ri}(\mathit{dom}(f))$ [21, Theorem 23.4]. These results are applied to the function $f = H$ with $\mathit{dom}(f) = \mathit{cc}(\mu)$, Lemma 1, whose convex conjugate is $f^* = \Lambda$, Lemma 3. It follows that if a is in $\mathit{ri}(\mathit{cc}(\mu)) = \mathit{ri}(\mathit{cs}(\mu))$ [13, Lemma 1] then $\partial H(a)$ is nonempty and $H(a) + \Lambda(\vartheta) = \langle \vartheta, a \rangle$ for $\vartheta \in \partial H(a)$. This proves *(i)*, since $\langle \vartheta, a \rangle - \Lambda(\vartheta) \leqslant \Lambda^*(a) \leqslant H(a)$, by Corollary 4.

(ii) This is a consequence of Lemma 6 in Appendix B, since $a \notin \mathit{ri}(\mathit{cs}(\mu))$ implies the existence of a closed halfspace containing $\mathit{cs}(\mu)$ whose boundary hyperplane contains a but not the whole $\mathit{cs}(\mu)$. □

Remark 4. A question motivated by large deviations theory is under what conditions does $\Lambda^*(a) = H(a)$ hold for each $a \in \mathbb{R}^d$. Although the subject is not in the scope of this paper, that question will be briefly addressed in Section 5, Remark 10.

Lemma 5. *If a sequence of pm's Q_{ϑ_n} in an exponential family $\mathcal{E}$ rI-converges to some pm $Q = Q_{F,\vartheta}$ in a component $\mathcal{E}^F$ of $\mathit{ext}(\mathcal{E})$ then*

$$\langle \vartheta_n, a \rangle - \Lambda(\vartheta_n) \to \langle \vartheta, a \rangle - \Lambda_F(\vartheta) \,,$$

for all a in the affine hull of F.

Proof. The hypothesis $D(Q \| Q_{\vartheta_n}) \to 0$ implies, by the obvious inequality

$$D(P\|R) \geqslant P(B)\; D(P(\cdot|B)\|R(\cdot|B)) \,, \qquad P(B) > 0 \,, \tag{8}$$

that for each Borel set B of positive Q-measure

$$D(Q(\cdot|B)\|Q_{\vartheta_n}(\cdot|B)) = \int_B \ln \frac{\exp[\,\langle \vartheta, x\rangle - \Lambda_F(\vartheta)\,]}{\exp[\,\langle \vartheta_n, x\rangle - \Lambda(\vartheta_n)\,]}\; dQ(\cdot|B) - \ln \frac{Q(B)}{Q_{\vartheta_n}(B)} \to 0 \,.$$

The logarithmic term vanishes in the limit because Q_{ϑ_n} converges to Q in variation distance. It follows that if the mean of $Q(\cdot|B)$ exists, it belongs to the set A of those $a \in \mathbb{R}^d$ for which $\langle \vartheta_n, a \rangle - \Lambda(\vartheta_n)$ converges to $\langle \vartheta, a \rangle - \Lambda_F(\vartheta)$. Hence, A contains the means of conditionings of Q on arbitrarily small balls with centers in the support of Q (the smallest closed set of Q-measure 1). It is not difficult to see that the affine hull of $\mathit{cs}(Q) = \mathit{cl}(\mathit{cc}(Q)) = \mathit{cl}(F)$ not only contains these means but is even spanned by them. The assertion follows as A is clearly an affine set. □

Corollary 5. *If a sequence of pm's $Q_n \in \mathcal{E}$ rI-converges to some $Q \in \mathit{ext}(\mathcal{E})$ then $D(P\|Q_n)$ tends to $D(P\|Q)$ for each pm P that has a mean.*

Proof. Suppose $Q \in \mathcal{E}^F$, say $Q = Q_{F,\vartheta}$. The assertion is trivial by lower semicontinuity of I-divergence if $D(P\|Q)$ is infinite. If $D(P\|Q)$ is finite, the mean of P belongs to $\mathit{cc}(Q) = F$ by [13, Theorem 3]. Denoting this mean by a, we have

$$D(P\|Q) = D(P\|\mu) + \langle \vartheta, a \rangle - \Lambda_F(\vartheta)$$

by (7), and

$$D(P\|Q_n) = D(P\|\mu) + \langle \vartheta_n, a \rangle - \Lambda(\vartheta_n)$$

by (6), where ϑ_n parametrizes Q_n. Therefore the assertion follows from Lemma 5. □

4. Pythagorean theorem in I-divergence geometry

The main result in this section is a general "Pythagorean identity" for I-divergences that, for $\mathcal{S} = \mathcal{L}$, substantially sharpens the first part of Theorem 1. No regularity assumptions will be used other than a finiteness assumption needed for the problem to be meaningful, see Corollary 3. Recall that each point in the convex set $cc(\mathcal{E})$ belongs to the relative interior $ri(F)$ of exactly one face F of $cc(\mathcal{E})$.

Theorem 3. *Let $\mathcal{L}$ be a linear family and $\mathcal{E}$ be an exponential family with $D(\mathcal{L}\|ext(\mathcal{E}))$ finite, that is with $cc(\mathcal{E})$ containing the mean of $\mathcal{L}$. Then the intersection $cl_I(\mathcal{L}) \cap ext(\mathcal{E})$ consists of a single pm R, and this pm R satisfies*

$$D(P\|Q) = D(P\|R) + D(\mathcal{L}\|Q)\,, \qquad P \in \mathcal{L}\,,\ Q \in ext(\mathcal{E})\,. \tag{9}$$

Moreover, R belongs to that component $\mathcal{E}^F$ of $ext(\mathcal{E})$ for which the mean of $\mathcal{L}$ is in $ri(F)$, and R is the unique pm satisfying (9).

Remark 5. The pm R belongs to $\mathcal{L}$ if and only if $\mathcal{L}$ intersects $ext(\mathcal{E})$. In that case, (9) with $P = R$ gives $D(R\|Q) = D(\mathcal{L}\|Q)$, and thus (9) rewrites as

$$D(P\|Q) = D(P\|R) + D(R\|Q)\,, \qquad P \in \mathcal{L}\,,\ Q \in ext(\mathcal{E})\,. \tag{10}$$

A direct proof of (10) is straightforward. The difficulty in proving Theorem 3 is due to the fact that $\mathcal{L}$ need not intersect $ext(\mathcal{E})$.

Remark 6. The fact that $cl_I(\mathcal{L}) \cap ext(\mathcal{E})$ is nonempty when $D(\mathcal{L}\|ext(\mathcal{E}))$ is finite, implies that $D(\mathcal{L}\|ext(\mathcal{E}))$ is dichotomic, either infinite or zero.

Proof of Theorem 3. Given $\mathcal{L} = \mathcal{L}_a$ and $\mathcal{E} = \mathcal{E}_\mu$ with $a \in cc(\mu)$, let F be the face of $cc(\mu)$ whose relative interior contains a. Then $D(P\|\mu) = D(P\|\mu^{cl(F)})$ for each pm P with mean a and $D(P\|\mu)$ finite, as in the proof of Lemma 2, and thus $H(a) = H_\mu(a)$ of (5) is equal to $H_{\mu^{cl(F)}}(a)$. Since $cc(\mu^{cl(F)}) = F$ [13, Lemma 3], the hypothesis of Lemma 4*(i)* is satisfied for $\mu^{cl(F)}$ in the role of μ, and it follows that there exists $\vartheta \in \mathbb{R}^d$ satisfying

$$H(a) = H_{\mu^{cl(F)}}(a) = \langle \vartheta, a \rangle - \Lambda_F(\vartheta)\,. \tag{11}$$

We claim that with this ϑ, (9) holds for $R = Q_{F,\vartheta} \in \mathcal{E}^F$.

Fix $Q \in ext(\mathcal{E})$, say $Q = Q_{G,\tau} \in \mathcal{E}^G$, where G is a face of $cc(\mu)$. Then the convex core of Q is $cc(\mathcal{E}^G) = G$. We may assume that G contains a, since, otherwise, $D(\mathcal{L}\|Q)$ is infinite by Lemma 1, and both sides of (9) equal $+\infty$ if $P \in \mathcal{L}$. The face G, containing $a \in ri(F)$, contains the whole face F, by [21, Theorem 18.1].

Next, for any P with mean $a \in F \subseteq G$, Lemma 2 applied to $Q = Q_{G,\tau}$ gives

$$D(P\|Q) = D(P\|\mu) - \langle \tau, a \rangle + \Lambda_G(\tau)$$

whence by minimization over $P \in \mathcal{L}_a$

$$D(\mathcal{L}\|Q) = H(a) - \langle \tau, a \rangle + \Lambda_G(\tau)\,.$$

Similarly, Lemma 2 applied to $R = Q_{F,\vartheta}$ gives

$$D(P\|R) = D(P\|\mu) - \langle \vartheta, a \rangle + \Lambda_F(\vartheta)\,.$$

Comparison of the last three equations and (11) establishes our claim.

By construction, R was in $\mathcal{E}^F \subseteq \mathit{ext}(\mathcal{E})$. Eq. (9) implies that it is also in $\mathit{cl}_I(\mathcal{L})$, since minimizing both sides over $P \in \mathcal{L}$ entails $D(\mathcal{L}\|R) = 0$. To prove the uniqueness assertions, it suffices to show that if $R' \in \mathit{cl}_I(\mathcal{L}) \cap \mathit{ext}(\mathcal{E})$ and R'' satisfies (9) then necessarily $R' = R''$. Now, substituting $Q = R'$ into (9) with R'', we obtain

$$D(P\|R') = D(P\|R'') + D(\mathcal{L}\|R') = D(P\|R'') , \qquad P \in \mathcal{L} .$$

This implies that a sequence of pm's in $\mathcal{L}$ that I-converges to R' (such a sequence exists since $R' \in \mathit{cl}_I(\mathcal{L})$) also I-converges to R''. Hence, $R' = R''$ as claimed. □

An attractive feature of (9) is that it embraces and strengthens instances of inequality (1) with $\mathcal{S} = \mathcal{L}$ and, at the same time, instances of inequality (2) with $\mathcal{S} = \mathit{ext}(\mathcal{E})$ and with $\mathcal{S} = \mathcal{E}$, without relying upon Theorem 1 for existence. (The results in this section do not depend on Theorem 2, either.) To express the feature of Theorem 3, hinted to above and formally stated in the following corollary, the pm R in (9) will be denoted by $\Pi_{\mathcal{L}\leftrightarrow \mathit{ext}(\mathcal{E})}$ in the sequel.

Corollary 6. *For a linear family $\mathcal{L} = \mathcal{L}_a$ and an exponential family $\mathcal{E}$ with $a \in \mathit{cc}(\mathcal{E})$, the pm $\Pi_{\mathcal{L}\leftrightarrow \mathit{ext}(\mathcal{E})}$ is both the generalized I-projection $\Pi_{\mathcal{L}\leftarrow Q}$ to $\mathcal{L}$ of every $Q \in \mathit{ext}(\mathcal{E})$ with $D(\mathcal{L}\|Q)$ finite, and the true rI-projection $\Pi_{P\rightarrow \mathit{ext}(\mathcal{E})}$ to $\mathit{ext}(\mathcal{E})$ of every $P \in \mathcal{L}$ with $D(P\|\mathcal{E})$ finite. Moreover, if $a \in \mathit{ri}(\mathit{cs}(\mathcal{E}))$ then $\Pi_{\mathcal{L}\leftrightarrow \mathit{ext}(\mathcal{E})}$ is the rI-projection $\Pi_{P\rightarrow\mathcal{E}}$ to $\mathcal{E}$ of every $P \in \mathcal{L}$ with $D(P\|\mathcal{E})$ finite.*

Proof. Only the assertions about rI-projections require proof. To this, note that (9) implies by minimizing over $Q \in \mathit{ext}(\mathcal{E})$, and using Remark 6, that

$$D(P\|\mathit{ext}(\mathcal{E})) = D(P\|R) , \qquad P \in \mathcal{L} . \tag{12}$$

By Remark 1 after Theorem 1, the last term in (9) is bounded from below by $D(R\|Q)$, thus we obtain

$$D(P\|Q) \geqslant D(P\|R) + D(R\|Q) = D(P\|\mathit{ext}(\mathcal{E})) + D(R\|Q) , \quad P \in \mathcal{L} ,\ Q \in \mathit{ext}(\mathcal{E}) .$$

This shows that $R = \Pi_{\mathcal{L}\leftrightarrow \mathit{ext}(\mathcal{E})}$ satisfies (2) in the role of $\Pi_{P\rightarrow\mathcal{S}}$, where $\mathcal{S} = \mathit{ext}(\mathcal{E})$, for all $P \in \mathcal{L}$ with $D(P\|\mathit{ext}(\mathcal{E}))$ finite. As $\Pi_{\mathcal{L}\leftrightarrow \mathit{ext}(\mathcal{E})} \in \mathit{ext}(\mathcal{E})$, it is, therefore, the rI-projection of P to $\mathit{ext}(\mathcal{E})$. If $a \in \mathit{ri}(\mathit{cs}(\mathcal{E}))$ then Theorem 3 gives that $\Pi_{\mathcal{L}\leftrightarrow \mathit{ext}(\mathcal{E})}$ belongs to $\mathcal{E}$ itself. Hence, it is the rI-projection of P to $\mathcal{E}$. □

The following corollary of Theorem 3 is immediate.

Corollary 7. *The true I-projection of a pm Q to a linear family $\mathcal{L}$ exists if and only if $\mathcal{L}$ intersects $\mathit{ext}(\mathcal{E}_Q)$. Then this I-projection is $P^* = \Pi_{\mathcal{L}\leftrightarrow \mathit{ext}(\mathcal{E})}$, and satisfies*

$$D(P\|Q) = D(P\|P^*) + D(P^*\|Q) , \qquad P \in \mathcal{L}.$$

An exponential family $\mathcal{E}$ is called *steep* if $\mathit{dom}(\Lambda)$ has nonempty interior, and no pm $Q_\vartheta \in \mathcal{E}$ with ϑ on the boundary of $\mathit{dom}(\Lambda)$ has a mean (this trivially holds if $\mathit{dom}(\Lambda)$ is open). Steepness is a well-known necessary and sufficient condition for each $a \in \mathit{ri}(\mathit{cs}(\mathcal{E}))$ to be the mean of some $Q \in \mathcal{E}$, see [2], [4]. Recalling that for each face F of $\mathit{cc}(\mathcal{E})$, the convex core of the component $\mathcal{E}^F$ of $\mathit{ext}(\mathcal{E})$ is equal to F, and that the relative interiors of convex core and convex support coincide, Corollary 7 and the last necessary and sufficient condition yield the following.

Corollary 8. *A pm Q has I-projection to each linear family with mean in $\mathsf{cc}(Q)$ if and only if each component of $\mathsf{ext}(\mathcal{E}_Q)$ is steep. A sufficient condition for this situation is $\mathsf{dom}(\Lambda_Q) = \mathbb{R}^d$.*

Note that steepness of an exponential family or even openness of $\mathsf{dom}(\Lambda)$ does not guarantee steepness of the components of its extension.

5. Generalized rI-projections to exponential families

This section is devoted to the problem of minimizing $D(P\|Q)$ over Q in an exponential family $\mathcal{E} = \mathcal{E}_\mu$ when P is a pm with mean a. Recall that, by Corollary 3, $a \in \mathsf{cc}(\mathcal{E})$ is a necessary condition for the finiteness of $D(P\|\mathcal{E})$. When $D(P\|\mathcal{E})$ is finite, Lemma 2 implies that $D(P\|\mathcal{E}) = D(P\|\mu) - \Lambda^*(a)$, and the existence of the rI-projection of P to $\mathcal{E}$ is equivalent to $\Lambda^*(a) = \langle\vartheta, a\rangle - \Lambda(\vartheta)$ for some $\vartheta \in \mathsf{dom}(\Lambda)$. The latter takes place if and only if $a \in \mathsf{ri}(\mathsf{cc}(\mathcal{E}))$, by Lemma 4. In this case, $\Pi_{P\to\mathcal{E}}$ equals the pm $R = \Pi_{\mathcal{L}\leftrightarrow\mathsf{ext}(\mathcal{E})}$ of Theorem 3, see Corollary 6. The results in this section cover the case when a belongs to $\mathsf{cc}(\mathcal{E})$ but not to its relative interior.

We shall use without further reference the consequence $\mathsf{cl}_{rI}(\mathcal{E}) \subseteq \mathsf{ext}(\mathcal{E})$ of Theorem 2. A sufficient condition for the equality appears in Lemma 7*(ii)*, see Appendix B; it is satisfied if $\mathsf{dom}(\Lambda) = \mathbb{R}^d$. The case of strict inclusion will be, however, theoretically more interesting.

The following theorem strengthens the log-convex part of Theorem 1 for $\mathcal{S} = \mathcal{E}$. Its proof, unlike that of Theorem 3, will rely upon Theorem 1.

Theorem 4. *For every linear family $\mathcal{L}$ and exponential family $\mathcal{E}$ with $D(\mathcal{L}\|\mathcal{E})$ finite, there exists a unique pm, denoted by $\Pi_{\mathcal{L}\to\mathcal{E}}$, such that*

$$D(P\|Q) \geqslant D(P\|\mathcal{E}) + D(\Pi_{\mathcal{L}\to\mathcal{E}}\|Q)\,, \quad P \in \mathcal{L}\,,\ Q \in \mathsf{cl}_{rI}(\mathcal{E})\,. \tag{13}$$

$\Pi_{\mathcal{L}\to\mathcal{E}}$ belongs to $\mathsf{cl}_{rI}(\mathcal{E})$, and $D(P\|\Pi_{\mathcal{L}\to\mathcal{E}}) = D(P\|\mathcal{E}) = D(P\|\mathsf{cl}_{rI}(\mathcal{E}))$ for each $P \in \mathcal{L}$. Moreover, $D(\mathcal{L}\|\mathcal{E}) = D(\mathcal{L}\|\Pi_{\mathcal{L}\to\mathcal{E}})$, and $\mathsf{cc}(\Pi_{\mathcal{L}\to\mathcal{E}})$ contains the mean of $\mathcal{L}$.

Proof. Given $\mathcal{L} = \mathcal{L}_a$ and $\mathcal{E} = \mathcal{E}_\mu$ with $D(\mathcal{L}\|\mathcal{E})$ finite, fix first a pm $P \in \mathcal{L}$ with $D(P\|\mathcal{E})$ finite. By Theorem 1, applied to $\mathcal{S} = \mathcal{E}$, there exists a unique pm $R = \Pi_{P\to\mathcal{E}} \in \mathsf{cl}_{rI}(\mathcal{E})$ such that

$$D(P\|Q) - D(P\|\mathcal{E}) \geqslant D(R\|Q)\,, \qquad Q \in \mathcal{E}\,.$$

It follows from (6) that the left-hand side of this inequality is constant for $P \in \mathcal{L}$ with $D(P\|\mathcal{E})$ finite (it equals $\Lambda(\vartheta) + \Lambda^*(a) - \langle\vartheta, a\rangle$ if $Q = Q_\vartheta$). Hence the pm R above does not depend on a particular choice of P, and the inequality (13) holds, with this R as $\Pi_{\mathcal{L}\to\mathcal{E}}$, for each $P \in \mathcal{L}$ and $Q \in \mathcal{E}$ (note that (13) trivially holds for $Q \in \mathcal{E}$ if $D(P\|\mathcal{E})$ is infinite). Using Corollary 5 and the lower semicontinuity of I-divergence, the validity of (13) for all $Q \in \mathsf{cl}_{rI}(\mathcal{E})$ is a consequence of its already established, weaker version for $Q \in \mathcal{E}$.

For $P \in \mathcal{L}$, the equality $D(P\|\mathcal{E}) = D(P\|\mathsf{cl}_{rI}(\mathcal{E}))$ follows by minimization in (13) over $Q \in \mathsf{cl}_{rI}(\mathcal{E})$, and $D(P\|\Pi_{\mathcal{L}\to\mathcal{E}}) = D(P\|\mathcal{E})$ holds due to Corollary 5 and $\Pi_{\mathcal{L}\to\mathcal{E}} \in \mathsf{cl}_{rI}(\mathcal{E})$.

Hence, in turn, $D(\mathcal{L}\|\Pi_{\mathcal{L}\to\mathcal{E}}) = D(\mathcal{L}\|\mathcal{E})$ follows and the mean of $\mathcal{L}$ belongs to $\mathsf{cc}(\Pi_{\mathcal{L}\to\mathcal{E}})$ by Lemma 1. $\square$

Remark 7. It remains open whether the rI-closure of an exponential family $\mathcal{E}$ is always log-convex and rI-closed. A partial result towards the rI-closedness is that each pm in the rI-closure of $\mathsf{cl}_{rI}(\mathcal{E})$ that has a mean necessarily belong to $\mathsf{cl}_{rI}(\mathcal{E})$. Indeed, by Theorem 4, $D(P\|\mathsf{cl}_{rI}(\mathcal{E}))=0$ implies $D(P\|\mathcal{E})=0$ for P with a mean.

Corollary 9. *For each pm P having a mean, with $D(P\|\mathcal{E})$ finite, the generalized rI-projection of P to $\mathcal{E}$ and the true rI-projection of P to $\mathsf{cl}_{rI}(\mathcal{E})$ exist and coincide.*

Corollary 10. *If $D(\mathcal{L}\|\mathcal{E})$ is finite,*

$$D(P\|Q)\geqslant D(P\|\Pi_{\mathcal{L}\leftrightarrow\mathsf{ext}(\mathcal{E})})+D(\mathcal{L}\|\mathcal{E})+D(\Pi_{\mathcal{L}\to\mathcal{E}}\|Q)\,,\qquad P\in\mathcal{L}\,,\ Q\in\mathsf{cl}_{rI}(\mathcal{E}). \tag{14}$$

Consequently, if $D(P_n\|Q_n)\to D(\mathcal{L}\|\mathcal{E})$ for pm's $P_n\in\mathcal{L}$ and $Q_n\in\mathcal{E}$ then P_n I-converges to $\Pi_{\mathcal{L}\leftrightarrow\mathsf{ext}(\mathcal{E})}$ and Q_n rI-converges to $\Pi_{\mathcal{L}\to\mathcal{E}}$.

Proofs of Corollaries 9 and 10. Corollary 9 is immediate from Theorem 4. Corollary 10 follows combining (13) and

$$D(P\|\mathcal{E})=D(P\|\Pi_{\mathcal{L}\leftrightarrow\mathsf{ext}(\mathcal{E})})+D(\mathcal{L}\|\mathcal{E})\,,\qquad P\in\mathcal{L}\,, \tag{15}$$

obtained by minimizing both sides of (9) subject to $Q\in\mathcal{E}$. □

Remark 8. The inequality in (13) may be strict. A simple example is provided by the exponential family on the real line based on the pm μ with density $2x^{-3}$ when $x\geqslant 1$, and 0 otherwise, and any linear family $\mathcal{L}$ with mean $a>2$. In this case, $\Pi_{\mathcal{L}\to\mathcal{E}}=\mu$. Note that the mean of $\Pi_{\mathcal{L}\to\mathcal{E}}$ differs from the mean of $\mathcal{L}$.

Remark 9. A sufficient condition for the equality in (13) is $\mathcal{L}\cap\mathsf{cl}_{rI}(\mathcal{E})\neq\emptyset$. Indeed, then the pm $R=\Pi_{\mathcal{L}\leftrightarrow\mathsf{ext}(\mathcal{E})}$ of Theorem 3 is the single member of that intersection, and it satisfies the equality (10) by Remark 5. Minimizing both sides of (10) over $Q\in\mathcal{E}$ gives $D(P\|\mathcal{E})=D(P\|R)$. Then (10) rewrites as (13) with the equality, having $\Pi_{\mathcal{L}\leftrightarrow\mathsf{ext}(\mathcal{E})}$ in the role of $\Pi_{\mathcal{L}\to\mathcal{E}}$.

Note that to guarantee $\mathcal{L}_a\cap\mathsf{cl}_{rI}(\mathcal{E})\neq\emptyset$ for all $a\in\mathsf{cc}(\mathcal{E})$, it does not suffice to assume that all components of $\mathsf{ext}(\mathcal{E})$ are steep (in Example 1 of Section 7, the choice $P=\delta_2$ with mean $a=(0,2)$ renders (13) strict). However, the stronger assumption $\mathsf{dom}(\Lambda)=\mathbb{R}^d$ suffices due to Corollary 8 and Lemma 7*(ii)* in Appendix B.

By Theorem 3, $\mathsf{cl}_I(\mathcal{L}_a)\cap\mathsf{ext}(\mathcal{E})$ is always nonempty for $a\in\mathsf{cc}(\mathcal{E})$, still the intersection of $\mathsf{cl}_I(\mathcal{L}_a)$ with the subset $\mathsf{cl}_{rI}(\mathcal{E})$ of $\mathsf{ext}(\mathcal{E})$ can be empty. Also $\Pi_{\mathcal{L}\leftrightarrow\mathsf{ext}(\mathcal{E})}$ may differ from $\Pi_{\mathcal{L}\to\mathcal{E}}$ in general. Example 1 in Section 7 demonstrates that these irregularities can happen and illustrates differences between Theorem 3 and Theorem 4. The following theorem shows that in this context, various "regularity conditions" are equivalent. There, H denotes the function in (5).

Theorem 5. *If $D(\mathcal{L}\|\mathcal{E})$ is finite then the following assertions are equivalent:*

(i) $\mathsf{cl}_I(\mathcal{L})\cap\mathsf{cl}_{rI}(\mathcal{E})\neq\emptyset$;
(ii) $\Pi_{\mathcal{L}\leftrightarrow\mathsf{ext}(\mathcal{E})}=\Pi_{\mathcal{L}\to\mathcal{E}}$;
(iii) $D(\mathcal{L}\|\mathcal{E})=0$;
(iv) $D(P\|Q)=D(P\|\mathcal{E})+D(\mathcal{L}\|Q)$ *for* $P\in\mathcal{L}$ *and* $Q\in\mathcal{E}$;
(v) *H is lower semicontinuous at a, the mean of $\mathcal{L}$;*

(vi) $H(a) = \Lambda^*(a)$.

In particular, these assertions hold if $a \in \mathit{ri}(\mathit{cs}(\mathcal{E}))$ *or if* $\mathit{cl}_{rI}(\mathcal{E}) = \mathit{ext}(\mathcal{E})$.

Proof. *(i)⇒(iii)*: If $Q \in \mathit{cl}_I(\mathcal{L}) \cap \mathit{cl}_{rI}(\mathcal{E})$ then $D(P_n\|Q) \to 0$ for some sequence $P_n \in \mathcal{L}$ and the implication follows from (14).

(iii)⇒(ii): If $D(\mathcal{L}\|\mathcal{E}) = 0$, that is $D(P_n\|Q_n) \to 0$ for sequences $P_n \in \mathcal{L}$ and $Q_n \in \mathcal{E}$ then, by Corollary 10, also $D(P_n\|\Pi_{\mathcal{L}\leftrightarrow \mathit{ext}(\mathcal{E})}) \to 0$ and $D(\Pi_{\mathcal{L}\to\mathcal{E}}\|Q_n) \to 0$. The three convergences imply $\Pi_{\mathcal{L}\leftrightarrow \mathit{ext}(\mathcal{E})} = \Pi_{\mathcal{L}\to\mathcal{E}}$, using Pinsker's inequality.

(ii)⇒(i): $\Pi_{\mathcal{L}\leftrightarrow \mathit{ext}(\mathcal{E})}$ belongs to $\mathit{cl}_I(\mathcal{L})$ by Theorem 3 and $\Pi_{\mathcal{L}\to\mathcal{E}}$ belongs to $\mathit{cl}_{rI}(\mathcal{E})$ by Theorem 4. If they are equal the intersection in *(i)* is nonempty.

(ii)⇒(iv): If $\Pi_{\mathcal{L}\leftrightarrow \mathit{ext}(\mathcal{E})} = \Pi_{\mathcal{L}\to\mathcal{E}}$ then (9) rewrites as $D(P\|Q) = D(P\|\Pi_{\mathcal{L}\to\mathcal{E}}) + D(\mathcal{L}\|Q)$ where $D(P\|\Pi_{\mathcal{L}\to\mathcal{E}}) = D(P\|\mathcal{E})$ by Theorem 4.

(iv)⇒(iii): By minimization over $Q \in \mathcal{E}$ when $D(P\|\mathcal{E})$ is finite.

(iii)⇔(vi): Obvious from Corollary 4 and the fact that $a \in \mathit{cc}(\mathcal{E})$.

(vi)⇒(v): Since Λ^* is a lower bound to H by Corollary 4, and Λ^*, a convex conjugate function, is lower semicontinous, $a \to a_n$ implies

$$\liminf H(a_n) \geqslant \liminf \Lambda^*(a_n) \geqslant \Lambda^*(a) = H(a)\ .$$

(v)⇒(vi): Lower semicontinuity at a of the convex function H implies $H(a_n) \to H(a)$ if $a_n \to a$ along a line segment in $\mathit{ri}(\mathit{dom}(H))$; hence $H(a) = \Lambda^*(a)$ follows since $H = \Lambda^*$ in the relative interior of $\mathit{dom}(H) = \mathit{cc}(\mu)$, by Lemma 4*(i)*.

If $a \in \mathit{ri}(\mathit{cs}(\mathcal{E}))$ then $\Pi_{\mathcal{L}\leftrightarrow \mathit{ext}(\mathcal{E})} \in \mathcal{E}$ by Theorem 3, and *(i)* holds. If $\mathit{ext}(\mathcal{E})$ coincides with $\mathit{cl}_{rI}(\mathcal{E})$ then *(i)* holds due to Theorem 3. □

Remark 10. If μ is a pm, the large deviation behaviour of the mean, respectively, empirical distribution of an i.i.d. sample from μ is governed by the rate function Λ^*, respectively, $D(\cdot\|\mu)$, see [14, Corollary 6.1.6], [14, Theorem 6.2.10]. Hence, a formal application of the contraction principle suggests that the function H in (5) is equal to Λ^*. The continuity assumption needed for the standard form of the contraction principle [14, Theorem 4.2.1] holds for the present case only if $\mathit{cs}(\mu)$ is bounded. It is therefore of interest under what weaker conditions the equality $H = \Lambda^*$ remains valid. By Theorems 3 and 5*(i)(vi)*, the condition $\mathit{cl}_{rI}(\mathcal{E}) = \mathit{ext}(\mathcal{E})$ is sufficient for $H(a) = \Lambda^*(a)$ when $D(\mathcal{L}_a\|\mathcal{E})$ is finite, that is, see Lemma 1 and Corollary 3, when a is in $\mathit{cc}(\mu) = \mathit{dom}(H)$. However, this condition is not sufficient for $\mathit{dom}(\Lambda^*) = \mathit{cc}(\mu)$, see Example 2 in Section 7. By Proposition 1*(iii)* in Section 6, either of the conditions $\mathit{cc}(\mu) = \mathit{ri}(\mathit{cs}(\mu))$ or $\mathit{dom}(\Lambda) = \mathbb{R}^d$ implies $\mathit{dom}(\Lambda^*) = \mathit{cc}(\mu)$. As either of them implies $\mathit{cl}_{rI}(\mathcal{E}) = \mathit{ext}(\mathcal{E})$ (the latter by Lemma 7*(ii)* in Appendix B), these are sufficient conditions for $H = \Lambda^*$; a necessary and sufficient condition remains elusive.

6. ML ESTIMATES

Let $\mathcal{E} = \mathcal{E}_\mu$ as before. The (normalized) *log-likelihood function* (LLF) associated with a sample $\boldsymbol{x} = (x^1, ..., x^n)$ of size $n \geqslant 1$ from an unknown distribution $Q_{F,\vartheta} \in \mathit{ext}(\mathcal{E})$ can be defined as the function of (F, ϑ) given by $\frac{1}{n}$ times the logarithm of the density

$$\frac{dQ^n_{F,\vartheta}}{d\mu^n} = \begin{cases} \prod_{i=1}^n e^{\langle \vartheta, x^i \rangle - \Lambda_F(\vartheta)}, & \text{if } x^i \in \mathit{cl}(F), 1 \leqslant i \leqslant n \\ 0, & \text{otherwise.} \end{cases} \tag{16}$$

In this paper, however, we will define the LLF for $\mathit{ext}(\mathcal{E})$ as the (extended real-valued) function

$$\ell_a(F, \vartheta) = \begin{cases} \langle \vartheta, a \rangle - \Lambda_F(\vartheta), & \text{if } \mathit{cl}(F) \text{ contains } a \\ -\infty & \text{otherwise,} \end{cases} \tag{17}$$

where $a = \frac{1}{n} \sum_{i=1}^n x^i$ is the sample mean. The two definitions coincide for samples $\boldsymbol{x}$ whose mean belongs to the closure of any face F of $\mathit{cc}(\mathcal{E})$ (if and) only if $x^i \in \mathit{cl}(F)$ for $1 \leqslant i \leqslant n$. Hence, adopting (17) as the definition of LLF is justified by the fact (whose proof is omitted, for brevity) that the set of samples not having the above property has μ^n measure 0.

For $F = \mathit{cc}(\mathcal{E})$ we shall simply write $\ell_a(\vartheta)$ and call it the LLF for $\mathcal{E}$. If the maximum of $\ell_a(\vartheta)$ subject to $\vartheta \in \mathit{dom}(\Lambda)$ is attained and finite, a maximizer ϑ^* will be called a *maximum-likelihood estimate* (MLE) *in* $\mathcal{E}$, from the sample $\boldsymbol{x}$ with mean a. An obvious necessary condition for existence of an MLE is the finiteness of $\sup_\vartheta \ell_a(\vartheta) = \Lambda^*(a)$, that is, $a \in \mathit{dom}(\Lambda^*)$. This condition is, however, not sufficient; by Lemma 4 the necessary and sufficient condition is $a \in \mathit{ri}(\mathit{cs}(\mathcal{E}))$. In this subsection remedies to non-existence of MLE will be offered when $a \in \mathit{dom}(\Lambda^*) \setminus \mathit{ri}(\mathit{cs}(\mathcal{E}))$.

No explicit description of $\mathit{dom}(\Lambda^*)$ seems to be available. Some partial results are given in the following proposition. We note that the three inclusions of *(i)* and *(ii)* may be simultaneously strict, see Example 1 in Section 7. Previous results in this direction are [2, Theorem 9.1*(ii)**], stating that $\mathit{dom}(\Lambda^*)$ contains the interior of $\mathit{cs}(\mu)$, and [2, Theorem 9.5] which is effectively the same as part *(ii)* of the following proposition.

Proposition 1. *(i)* $\mathit{dom}(\Lambda^*) \supseteq \mathit{cc}(\mu)$.

(ii) The intersection of all open half-spaces with full μ-measure contains $\mathit{dom}(\Lambda^*)$ and is contained in $\mathit{cs}(\mu)$.

(iii) $\mathit{dom}(\Lambda^*) = \mathit{cc}(\mu)$ *whenever* $\mathit{cc}(\mu)$ *is relatively open or* $\mathit{dom}(\Lambda) = \mathbb{R}^d$.

Proof. *(i)* This follows from $\Lambda^* \leqslant H$, Corollary 4, and $\mathit{dom}(H) = \mathit{cc}(\mu)$, Lemma 1.

The remaining assertions are consequences of Lemma 6 in Appendix B.

(ii) If a is not in the intersection, there exists an open half-space of full μ-measure whose boundary hyperplane H contains a. Then, $\Lambda^*(a) = +\infty$ by Lemma 6. The intersection is obviously a subset of $\mathit{cs}(\mu)$.

(iii) By part *(i)* it suffices to verify the inclusion $\subseteq$. If $\mathit{cc}(\mu)$ is relatively open then it equals the intersection of all open half-spaces with full μ-measure, and $\mathit{dom}(\Lambda^*) \subseteq \mathit{cc}(\mu)$ follows by *(ii)*. Suppose now that $\mathit{dom}(\Lambda)$ equals $\mathbb{R}^d$. If $\Lambda^*(a)$ is finite then by part *(ii)* either $a \in \mathit{ri}(\mathit{cs}(\mu)) = \mathit{ri}(\mathit{cc}(\mu))$, or else a is on the relative boundary of $\mathit{cs}(\mu)$. In the latter case, in addition, a nontrivial supporting hyperplane H to $\mathit{cs}(\mu)$ at a has positive μ-measure. Then for $F = \mathit{cc}(\mu) \cap H$ we have $F = \mathit{cc}(\mu^H)$ by [13, Lemma 2*(ii)*], and hence $\mathit{cl}(F) = \mathit{cs}(\mu^H)$ by [13, Lemma 1]. Thus

$$\mu(\mathit{cl}(F)) = \mu(H), \tag{18}$$

in particular, F is non-empty and is a proper (exposed) face of $cc(\mu)$. By Lemma 6 and (18) we have $\Lambda^*(a) = \Lambda^*_H(a) = \Lambda^*_F(a)$. Repeating the argument for $\mu^{cl(F)}$ in the role of μ, one concludes that either $a \in ri(cs(\mu^{cl(F)})) = ri(F)$ or else $\Lambda^*_F(a) = \Lambda^*_G(a)$ for a proper face G of F, etc. After at most d steps, a turns out to be in the relative interior of a face of $cc(\mu)$. □

Theorem 6. *For each $a \in dom(\Lambda^*)$ there exists a unique pm R^*_a satisfying*

$$\Lambda_F(\vartheta) + \Lambda^*(a) - \langle \vartheta, a\rangle \geqslant D(R^*_a \| Q_{F,\vartheta}) \tag{19}$$

for each $Q_{F,\vartheta} \in cl_{rI}(\mathcal{E})$ such that the affine hull of F contains a.

Proof. Suppose first that a belongs to $cc(\mu)$. Then a pm P with mean a and $D(P\|\mu)$ finite exists by [13, Theorem 3]. Since $D(P\|Q_\vartheta) = D(P\|\mu) - \langle\vartheta, a\rangle + \Lambda(\vartheta)$ by Lemma 2, the “log-convex part” of Theorem 1 applied to this P and $\mathcal{S} = \mathcal{E}$ implies

$$\Lambda(\vartheta) + \Lambda^*(a) - \langle \vartheta, a\rangle \geqslant D(R^*_a \| Q_\vartheta)\ , \qquad Q_\vartheta \in \mathcal{E}\,, \tag{20}$$

with $R^*_a = \Pi_{P\to\mathcal{E}}$. Any pm $Q_{F,\vartheta} \in cl_{rI}(\mathcal{E})$ is the rI-limit of a sequence Q_{ϑ_n} from $\mathcal{E}$. If the affine hull of F contains a then $\langle\vartheta_n, a\rangle - \Lambda(\vartheta_n)$ converges to $\langle\vartheta, a\rangle - \Lambda_F(\vartheta)$, by Lemma 5. Since $D(R^*_a\|Q_{\vartheta_n})$ cannot be eventually smaller that $D(R^*_a\|Q_{F,\vartheta})$, by lower semicontinuity of divergence, (20) implies (19).

If $a \in dom(\Lambda^*) \setminus cc(\mu)$, the existence of R^*_a satisfying (20) follows by a modification of the proof of Theorem 1, detailed in Appendix A. Then, by the above limiting argument, (19) holds also in this case. The uniqueness of R^*_a is obvious from (19). □

Corollary 11. *For $a \in dom(\Lambda^*)$, if a sequence Q_{ϑ_n} in $\mathcal{E}$ satisfies $\ell_a(\vartheta_n) \to \Lambda^*(a)$ then it rI-converges to R^*_a.*

If an MLE ϑ^* from a sample $\boldsymbol{x}$ with mean a exists, that is, $\ell_a(\vartheta^*) = \Lambda^*(a) < +\infty$, then $a \in dom(\Lambda^*)$ and Theorem 6 implies $D(R^*_a\|Q_{\vartheta^*}) = 0$, thus $R^*_a = Q_{\vartheta^*}$. When the MLE does not exist but $a \in dom(\Lambda^*)$, on account of Corollary 11 it is reasonable to call R^*_a the *generalized* MLE *(*GMLE*)*, from a sample $\boldsymbol{x}$ with mean a. Note that this GMLE is a pm rather than a parameter. As R^*_a belongs to $cl_{rI}(\mathcal{E})$, it can be represented as $R^*_a = Q_{F^*,\vartheta^*}$ where F^* is a unique face of $cc(\mu)$ and $\vartheta^* \in dom(\Lambda_{F^*})$, non-unique in general.

Remark 11. In statistics, approximate MLE's are often used, meaning a parameter ϑ such that the value $\ell_a(\vartheta)$ of the LLF is within ϵ of $\Lambda^*(a)$. By Theorem 6, for any approximate MLE ϑ, the pm Q_ϑ is close to the GMLE R^*_a in the divergence sense $D(R^*_a\|Q_\vartheta) \leqslant \epsilon$.

The following corollary relates the GMLE to an MLE in $cl_{rI}(\mathcal{E})$. That MLE is defined as a maximizer (F^*, ϑ^*) of $\ell_a(F, \vartheta)$ in (17) subject to $Q_{F,\vartheta} \in cl_{rI}(\mathcal{E})$, provided the maximum is attained and finite. We note, however, that sampling from $Q \in \mathcal{E}$ may yield with positive probability a sample $\boldsymbol{x}$ from which the GMLE exists but no MLE in $cl_{rI}(\mathcal{E})$ does, see Example 2 in Section 7.

Corollary 12. *If (F^*, ϑ^*) is an* MLE *in $cl_{rI}(\mathcal{E})$ from a sample $\boldsymbol{x}$ then the sample mean a belongs to $dom(\Lambda^*)$ and Q_{F^*,ϑ^*} equals the* GMLE *R^*_a from $\boldsymbol{x}$. For $a \in dom(\Lambda^*)$, such an* MLE *exists if and only if $a \in cs(R^*_a)$. This condition holds when $a \in cc(\mathcal{E})$.*

Proof. Suppose $\ell_a(F^*,\vartheta^*)$ with Q_{F^*,ϑ^*} in $\mathit{cl}_{rI}(\mathcal{E})$ is finite and equals the maximum of $\ell_a(F,\vartheta)$ subject to $Q_{F,\vartheta}\in \mathit{cl}_{rI}(\mathcal{E})$. The finiteness implies $\ell_a(F^*,\vartheta^*)=\langle\vartheta^*,a\rangle-\Lambda_{F^*}(\vartheta^*)$ and $a\in \mathit{cl}(F^*)$, by the definition (17) of LLF. Then $\langle\vartheta^*,a\rangle-\Lambda_{F^*}(\vartheta^*)\geqslant \ell_a(\vartheta)$ for $\vartheta\in \mathit{dom}(\Lambda)$ where $\ell_a(\vartheta)=\langle\vartheta,a\rangle-\Lambda(\vartheta)$ because $a\in \mathit{cs}(\mu)$. It follows that $\langle\vartheta^*,a\rangle-\Lambda_{F^*}(\vartheta^*)\geqslant\Lambda^*(a)$, and thus $a\in \mathit{dom}(\Lambda^*)$. Now, for $(F,\vartheta)=(F^*,\vartheta^*)$ the equality takes place in (19); note that the condition for a there is satisfied as $a\in \mathit{cl}(F^*)$. This proves that $Q_{F^*,\vartheta^*}=R^*_a$ and $a\in \mathit{cs}(R^*_a)=\mathit{cl}(F^*)$.

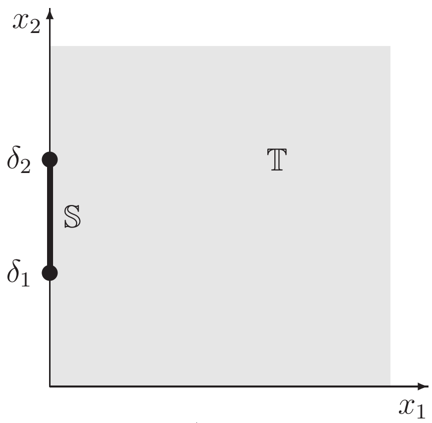
If $a\in \mathit{dom}(\Lambda^*)$ then $Q_{F,\vartheta}\in \mathit{cl}_{rI}(\mathcal{E})$ implies $\ell_a(\vartheta,F)\leqslant\Lambda^*(a)$, by (19). Moreover, using Proposition 1*(ii)*, $\ell_a(\vartheta,\mathit{cc}(\mu))=\langle\vartheta,a\rangle-\Lambda(\vartheta)$, hence the supremum of $\ell_a(\vartheta,F)$ subject to $Q_{F,\vartheta}\in \mathit{cl}_{rI}(\mathcal{E})$ equals the finite number $\Lambda^*(a)$. By Corollary 11, R^*_a is the rI-limit of a sequence $Q_{\vartheta_n}\in\mathcal{E}$ satisfying $\ell_a(\vartheta_n)\to\Lambda^*(a)$. Thus, $R^*_a=Q_{F^*,\vartheta^*}$ for some (F^*,ϑ^*). If, in addition, $a\in \mathit{cs}(R^*_a)=\mathit{cl}(F^*)$ then $\ell_a(F^*,\vartheta^*)=\langle\vartheta^*,a\rangle-\Lambda_{F^*}(\vartheta^*)$, by (17). Lemma 5 implies that $\langle\vartheta_n,a\rangle-\Lambda(\vartheta_n)$ converges to $\langle\vartheta^*,a\rangle-\Lambda_{F^*}(\vartheta^*)$. Hence, $\ell_a(F^*,\vartheta^*)=\Lambda^*(a)$ and (F^*,ϑ^*) is an MLE in $\mathit{cl}_{rI}(\mathcal{E})$.

When $a\in \mathit{cc}(\mathcal{E})$ then $a\in \mathit{dom}(\Lambda^*)$ by Proposition 1*(i)*, and $R^*_a=\Pi_{\mathcal{L}\to\mathcal{E}}$ for $\mathcal{L}=\mathcal{L}_a$, see the proof of Theorem 6. The last assertion of Theorem 4 implies $a\in \mathit{cc}(R^*_a)$. □

Remark 12. Theorem 6 can be easily extended to log-convex subfamilies $\{Q_\vartheta\colon\vartheta\in\Xi\}$ of $\mathcal{E}$, where Ξ is a convex subset of $\mathit{dom}(\Lambda)$, replacing $\Lambda^*(a)$ by $\sup_{\vartheta\in\Xi}\,[\langle\vartheta,a\rangle-\Lambda(\vartheta)]$. Moreover, MLE and GMLE can be considered for log-convex subfamilies of $\mathit{ext}(\mathcal{E})$. We intend to return to this topic elsewhere.

7. Examples

Example 1. Let μ be the measure on the Euclidean plane expressible as sum of the pm δ_1 sitting in the point $(0,1)$, δ_2 sitting in $(0,2)$ and the pm sitting on the open quadrant $\mathbb{T}=\{(x_1,x_2)\in\mathbb{R}^2;\ x_1>0,x_2>0\}$ with density $e^{-x_1-x_2}$ w.r.t. the Lebesgue measure. Then $\mathcal{E}=\mathcal{E}_\mu$ has convex support $\mathit{cs}(\mathcal{E})=\mathit{cl}(\mathbb{T})$ and convex core $\mathit{cc}(\mathcal{E})=\mathbb{T}\cup\mathbb{S}$ where $\mathbb{S}=\{(0,s):\ 1\leqslant s\leqslant 2\}$. This convex core has four faces, namely $\{(0,1)\}$, $\{(0,2)\}$, $\mathbb{S}$ and $\mathit{cc}(\mathcal{E})$. The first two are not exposed.



The extension of $\mathcal{E}$ has four components,

$$\mathit{ext}(\mathcal{E})=\{\delta_1\}\cup\{\delta_2\}\cup\mathcal{E}^{\mathbb{S}}\cup\mathcal{E}$$

where $\mathcal{E}^{\mathbb{S}}=\{(1-t)\delta_1+t\delta_2:\ 0<t<1\}$. By simple calculation

$$\Lambda(\vartheta_1,\vartheta_2)=\ln\left[e^{\vartheta_2}+e^{2\vartheta_2}+\tfrac{1}{1-\vartheta_1}\,\tfrac{1}{1-\vartheta_2}\right]$$

whenever $\vartheta_1 < 1$, $\vartheta_2 < 1$. The two strict inequalities define $\mathit{dom}(\Lambda)$.

By Proposition 1

$$\mathit{cc}(\mathcal{E}) = \mathbb{T} \cup \mathbb{S} \subseteq \mathit{dom}(\Lambda^*) \subseteq \mathbb{T} \cup \{(0,s) : s > 0\}$$

where the union on the right is the intersection of all open half-spaces with full μ-measure. For $a = (0,s)$

$$\Lambda^*(a) = \sup_{\vartheta_1<1,\vartheta_2<1} \left[s\vartheta_2 - \Lambda(\vartheta_1,\vartheta_2)\right] = \sup_{\vartheta_2<1} \left[(s-1)\vartheta_2 - \ln(1+e^{\vartheta_2})\right]$$

whence by simple calculus

$$\Lambda^*(a) = \begin{cases} +\infty\,, & s < 1 \\ (s-1)\ln(s-1) + (2-s)\ln(2-s)\,, & 1 \leqslant s \leqslant \frac{1+2e}{1+e} \\ s - 1 - \ln(1+e)\,, & \frac{1+2e}{1+e} < s\,. \end{cases}$$

This shows that the inclusions of Proposition 1*(i)(ii)* are strict. By Lemma 4, the function H defined in (5) equals Λ^* on $\mathbb{T}$. For $a = (0,s)$

$$H(0,s) = D(\mathcal{L}_a\|\mu) = (s-1)\ln(s-1) + (2-s)\ln(2-s)\,, \qquad 1 \leqslant s \leqslant 2,$$

because $(2-s)\delta_1 + (s-1)\delta_2$ is the only pm in $\mathcal{L}_a$ dominated by μ. Note that $H(a) = \Lambda^*(a)$ only for $s \leqslant \frac{1+2e}{1+e} < 2$.

To determine $\mathit{cl}_{rI}(\mathcal{E})$, note that by Lemma 7*(i)* in Appendix B, all conditioned pm's

$$Q_{(\vartheta_1,\vartheta_2)}(\cdot|\mathbb{S}) = \tfrac{1}{1+e^{\vartheta_2}}\delta_1 + \tfrac{e^{\vartheta_2}}{1+e^{\vartheta_2}}\delta_2\,, \qquad Q_{(\vartheta_1,\vartheta_2)} \in \mathcal{E}$$

belong to $\mathit{cl}_{rI}(\mathcal{E})$. Hence $\{(1-t)\delta_1 + t\delta_2 : 0 < t < \frac{e}{1+e}\} \subseteq \mathit{cl}_{rI}(\mathcal{E})$. The pm's δ_1 and $\frac{1}{1+e}\delta_1 + \frac{e}{1+e}\delta_2$, which are in the rI-closure of the previous set, also belong to $\mathit{cl}_{rI}(\mathcal{E})$, by Remark 7. Further, as $Q\{(0,2)\}/Q\{(0,1)\} = e^{\vartheta_2} < e$ for each $Q \in \mathcal{E}$, no other members of $\mathit{ext}(\mathcal{E}) \setminus \mathcal{E}$ can belong to $\mathit{cl}_{rI}(\mathcal{E})$. Thus,

$$\mathit{cl}_{rI}(\mathcal{E}) = \mathcal{E} \cup \left\{(1-t)\delta_1 + t\delta_2 : 0 \leqslant t \leqslant \tfrac{e}{1+e}\right\}.$$

The pm $\delta_1 \in \mathit{cl}_{rI}(\mathcal{E})$ does not belong to the "boundary at infinity" of $\mathcal{E}$ in the sense of [6]. Hence $R = \delta_1$ is a counterexample to [6, Theorem 23.3] claiming that each pm R with $D(R\|\mathcal{E})$ finite has an rI-projection to $\mathcal{E}$ completed with the boundary, and also to [6, Lemma 23.7] claiming rI-closedness of that completion. (In [6], our $D(R\|P)$ was denoted by $\boldsymbol{I}[\,P|R\,]$, and our rI-closedness was termed $\boldsymbol{I}$-closedness).

To illustrate Theorems 3, 4 and 5, let a linear family $\mathcal{L}$ have mean $a = (0,s)$, with $1 \leqslant s \leqslant 2$ to ensure that a belongs to $\mathit{cc}(\mathcal{E})$. Then $\Pi_{\mathcal{L}\leftrightarrow \mathit{ext}(\mathcal{E})}$ equals $(2-s)\delta_1 + (s-1)\delta_2$. Further $\Pi_{\mathcal{L}\to\mathcal{E}} = \Pi_{\mathcal{L}\leftrightarrow \mathit{ext}(\mathcal{E})}$ for $s \leqslant \frac{1+2e}{1+e}$, while $\Pi_{\mathcal{L}\to\mathcal{E}}$ equals $\frac{1}{1+e}\delta_1 + \frac{e}{1+e}\delta_2$ which is different from $\Pi_{\mathcal{L}\leftrightarrow \mathit{ext}(\mathcal{E})}$ for $s > \frac{1+2e}{1+e}$.

Finally, we discuss maximization of likelihood. The MLE in $\mathcal{E}$ from a sample $\boldsymbol{x}$ with mean $a = (0,s)$ does not exist, as a is on the boundary of $\mathit{cs}(\mathcal{E})$, cf. Lemma 4. The pm R^*_a of Theorem 6 is well-defined when $s \geqslant 1$. It equals $\Pi_{\mathcal{L}\to\mathcal{E}}$ if $a \in \mathit{cc}(\mathcal{E})$, i.e. $s \leqslant 2$, and $\frac{1}{1+e}\delta_1 + \frac{e}{1+e}\delta_2$, otherwise. Thus, the GMLE R^*_a is not in $\mathcal{E}$. By Corollary 12, an MLE (F^*,ϑ^*) in $\mathit{cl}_{rI}(\mathcal{E})$ from $\boldsymbol{x}$ exists if and only if $1 \leqslant s \leqslant 2$, that is, $a \in \mathit{cs}(R^*_a)$, in which case $Q_{F^*,\vartheta^*} = R^*_a$.

Example 2. Let $C \subseteq \mathbb{R}^3$ be the closed cone $\{(x, y, z) \colon z^2 \geqslant x^2 + y^2,\ z \geqslant 0\}$ and ν be the pm on the boundary of C equal to the joint distribution of (X, Y, Z) such that Z is exponentially distributed with density e^{-z}, $z \geqslant 0$, and the conditional distribution of (X, Y) given $Z = z$ is uniform on the circle $\{(x, y) \colon x^2 + y^2 = z^2\}$. Consider $\mathcal{E} = \mathcal{E}_\mu$ with $\mu = \nu + \delta$ where δ is the unit mass at the origin. Then $\mathit{cc}(\mu)$ is the union of $\{(0, 0, 0)\}$, and $\mathit{int}(C)$ and its only proper face is $\{(0, 0, 0)\}$; clearly $\mathit{cs}(\mathcal{E}) = C$. Thus $\mathit{ext}(\mathcal{E}) = \mathit{cl}_{rI}(\mathcal{E}) = \mathcal{E} \cup \{\delta\}$. We claim that $\mathit{dom}(\Lambda^*) = C$. To prove this, by Proposition 1, it suffices to show that $\Lambda^*(a) < +\infty$ for each boundary point of C. Now, since

$$\begin{aligned} \Lambda(\vartheta) &= \ln \Big[1 + \int e^{\vartheta_1 x + \vartheta_2 y + \vartheta_3 z}\, \nu(dx, dy, dz) \Big] \\ &= \ln \Big[1 + \int_0^{+\infty} \int_0^{2\pi} e^{z(\vartheta_1 \cos\varphi + \vartheta_2 \sin\varphi + \vartheta_3 - 1)}\, \tfrac{d\varphi\, dz}{2\pi} \Big] > 0\,, \end{aligned}$$

and $\Lambda(\vartheta) = +\infty$ if $\vartheta_1 \cos\varphi + \vartheta_2 \sin\varphi + \vartheta_3 > 1$ for some φ, it follows that for a boundary point $a = (t\cos\varphi, t\sin\varphi, t)$, $t \geqslant 0$, of C we have

$$\Lambda^*(a) \leqslant \sup_{\vartheta \in \mathit{dom}(\Lambda)} \langle \vartheta, a \rangle = t \cdot \sup_{\vartheta \in \mathit{dom}(\Lambda)} \big(\vartheta_1 \cos\varphi + \vartheta_2 \sin\varphi + \vartheta_3 \big) \leqslant t\,.$$

This establishes the claim.

Having a sample of size $n \geqslant 1$, it is a positive probability event that exactly $n - 1$ elements of the sample are equal to $(0, 0, 0)$. Then the sample mean does not belong to $\mathit{cc}(\mathcal{E})$ but remains on the boundary of C and thus in $\mathit{dom}(\Lambda^*)$. In this case, the GMLE exists and equals δ while no MLE in $\mathit{cl}_{rI}(\mathcal{E})$ exists, by Corollary 12.

Example 3. In this example, all pm's are on the real line $\mathbb{R}$, and $\mathcal{L}$ is the linear family with mean 0. We show that $\mathit{cl}_I(\mathit{cl}_I(\mathcal{L}))$ contains all Gaussian distributions.

By symmetry, let Q be Gaussian with density $q(x)$, mean $m < 0$ and variance σ^2. Consider the pm Q_k with density $q_k(x)$ equal to $q(x)$ when $x < k$, and $q_k(x) = a_k x^{-4}$ when $x \geqslant k$. Then

$$D(Q_k \| Q) = \int_k^{+\infty} q_k(x) \ln \tfrac{q_k(x)}{q(x)}\, dx = a_k \int_k^{+\infty} \Big[\ln a_k - 4 \ln x + \tfrac{1}{2} \ln \big(2\pi\sigma^2 \big) + \tfrac{(x-m)^2}{2\sigma^2} \Big] \tfrac{dx}{x^4}$$

goes to 0 as $k \to \infty$ (note that the normalizing constants a_k are bounded). Hence, the assertion $Q \in \mathit{cl}_I(\mathit{cl}_I(\mathcal{L}))$ follows if we prove that $Q_k \in \mathit{cl}_I(\mathcal{L})$. A little more generally, we show that if a pm R has mean $m < 0$ and density $r(x) = a x^{-4}$ for $x \geqslant \ell$ then $R \in \mathit{cl}_I(\mathcal{L})$. In fact, given such R, for $n > \ell$ there exists $P_n \in \mathcal{L}$ such that $\frac{dP_n}{dR}$ is constant both on $(-\infty, n)$ and on $(n, +\infty)$. The corresponding constants b_n and c_n can be determined from the conditions that P_n is a pm and that its mean is 0. Using the identities

$$\int_n^{+\infty} r(x)\, dx = \tfrac{a}{3n^3}\,, \qquad \int_n^{+\infty} x r(x)\, dx = \tfrac{a}{2n^2}$$

these conditions give the equations

$$b_n \Big(1 - \tfrac{a}{3n^3} \Big) + c_n \tfrac{a}{3n^3} = 1\,, \qquad b_n \Big(m - \tfrac{a}{2n^2} \Big) + c_n \tfrac{a}{2n^2} = 0$$

yielding

$$b_n = \tfrac{3n}{3n - 2m}\,, \qquad c_n = \tfrac{3n}{a}\, \tfrac{a - 2n^2 m}{3n - 2m}\,.$$

For $n \to \infty$ one has

$$D(P_n \| R) = \Big(1 - \tfrac{a}{3n^3} \Big)\, b_n\, \ln b_n + \tfrac{a}{3n^3}\, c_n\, \ln c_n \to 0\,.$$

This establishes $R \in cl_I(\mathcal{L})$, and consequently $Q \in cl_I(cl_I(\mathcal{L}))$.

The I-projection of any Gaussian pm Q with mean $m \neq 0$ to $\mathcal{L}$ is obviously the Gaussian pm with mean 0 and the same variance. Hence, $D(\mathcal{L}\|Q) > 0$ and $Q \notin cl_I(\mathcal{L})$. On the other hand, $D(cl_I(\mathcal{L})\|Q) = 0$ and thus the generalized I-projection of Q to the convex set $cl_I(\mathcal{L})$ equals Q itself. This exhibits occurrence of $\Pi_{\mathcal{L}\leftarrow Q} \neq \Pi_{cl_I(\mathcal{L})\leftarrow Q}$.

This example also demonstrates that the identity (9) cannot be extended to $P \in cl_I(\mathcal{L})$. Indeed, take $\mathcal{L} = \mathcal{L}_0$ and $\mathcal{E} = \mathcal{E}_Q$ with Q as above. Then (9) reduces to (10) where R is the Gaussian pm with mean 0 and variance σ^2. For $P = Q_k$ as above, $D(P\|Q)$ can be arbitrarily close to 0, whereas $D(P\|R) + D(R\|Q)$ is bounded from below by $D(R\|Q) > 0$.

8. I-, rI-projections and MLE in more general cases

In this section minimization of $D(P\|Q)$ over sets of pm's P, Q on a measurable space $(X, \mathcal{X})$ is compared with minimization of $D(\tilde{P}\|\tilde{Q})$ over sets of pm's $\tilde{P}, \tilde{Q}$ on $\mathbb{R}^d$. The MLE and GMLE are discussed in this framework, too. The focus is on general linear families

$$\mathcal{L}_{a,f} = \left\{P \text{ pm on } (X, \mathcal{X}) : \int_X f\, dP = a\right\}$$

where $f\colon X \to \mathbb{R}^d$ is a vector-valued measurable function and $a \in \mathbb{R}^d$, and on general exponential families

$$\mathcal{E}_{\mu,f} = \left\{Q_{\vartheta,f} : \frac{dQ_{\vartheta,f}}{d\mu}(x) = e^{\langle\vartheta,f(x)\rangle - \Lambda(\vartheta)}\ ,\ \ \vartheta \in dom(\Lambda)\right\}$$

where now μ is a finite nonzero measure on $(X, \mathcal{X})$ and

$$\Lambda(\vartheta) = \Lambda_{\mu,f}(\vartheta) = \ln \int_X e^{\langle\vartheta,f(x)\rangle}\, \mu(dx)\ .$$

The function f is the directional statistic of this exponential family.

8.1. **Direct generalizations.** It is a well-known fact that the minimization of $D(P\|Q)$ subject to $P \in \mathcal{L}_{a,f}$ or subject to $Q \in \mathcal{E}_{\mu,f}$ can be transferred to the Euclidean space $\mathbb{R}^d$ via the mapping f. To outline this idea, let μ_f denote the f-image of μ and, given a pm $\tilde{P}$ on $\mathbb{R}^d$ dominated by μ_f, let $\tilde{P}_{f^{-1},\mu}$ denote the pm on $(X, \mathcal{X})$ with μ-density $\frac{d\tilde{P}}{d\mu_f}(f(x))$. Recall further that the inequality $D(P\|Q) \geqslant D(P_f\|Q_f)$ holds; the equality takes place for P, Q dominated by μ if and only if $P = (P_f)_{f^{-1},\mu}$ and $Q = (Q_f)_{f^{-1},\mu}$. Since the mapping $Q \to Q_f$ maps $\mathcal{L}_{a,f}$ onto $\mathcal{L}_a$, $D(\mathcal{L}_{a,f}\|Q) \geqslant D(\mathcal{L}_a\|Q_f)$. It is not difficult to see that, actually, the equality takes place here, and when $D(\mathcal{L}_{a,f}\|Q)$ is finite, the generalized I-projection of Q to $\mathcal{L}_{a,f}$ equals $\tilde{P}_{f^{-1},\mu}$ where $\tilde{P}$ is the generalized I-projection of Q_f to $\mathcal{L}_a$. In addition, both I-projections can be true projections only simultaneously. An analogous observation is valid for the minimization of $D(P\|Q)$ over $Q \in \mathcal{E}_{\mu,f}$. Here, f is a sufficient statistic for this family and the mapping $Q \to Q_f$ is even a bijection of $\mathcal{E}_{\mu,f}$ onto the standard family $\mathcal{E}_{\mu_f}$, based on the f-image of μ, and $\tilde{Q} \to \tilde{Q}_{f^{-1},\mu}$ is its inverse. Note that $\Lambda_{\mu,f} = \Lambda_{\mu_f}$.

This simple device of transferring problems to $\mathbb{R}^d$ via the mapping f has been frequently employed to lift results from Euclidean spaces to more general settings. It works for our results as well. As a first example, one can immediately recognize whether the minimization of $D(P\|Q)$ over $P \in \mathcal{L}_{a,f}$ is a feasible problem, that is whether $P \in \mathcal{L}_{a,f}$ with $D(P\|Q)$ finite exists. The necessary and sufficient condition for this is $a \in cc(Q_f)$, by Lemma 1 or [13, Theorem 3].

The key concept in this paper, the extension of a standard exponential family, is generalized by defining $\mathsf{ext}(\mathcal{E}_{\mu,f})$ to be the set of all pm's $\tilde{P}_{f^{-1},\mu}$ where $\tilde{P} \in \mathsf{ext}(\mathcal{E}_{\mu_f})$. Equivalently, $\mathsf{ext}(\mathcal{E}_{\mu,f})$ is the union of its components; for each face F of $\mathsf{cc}(\mu_f)$ a component $\mathcal{E}^F_{\mu,f}$ of $\mathsf{ext}(\mathcal{E}_{\mu,f})$ is the exponential family $\mathcal{E}_{\nu,f}$ with ν equal to the restriction of μ to $f^{-1}(\mathsf{cl}(F)) = \{x \colon f(x) \in \mathsf{cl}(F)\}$, and with the previous directional statistic f.

All results of Sections 4, 5 and 6 admit straightforward generalizations. For example Theorem 3 extends to the following form.

For the set of pm's $\mathcal{L}_{a,f}$ and exponential family $\mathcal{E}_{\mu,f}$ such that $D(\mathcal{L}_{a,f}\|\mathcal{E}_{\mu,f})$ is finite, that is, $a \in \mathsf{cc}(\mu_f)$, the intersection $\mathsf{cl}_I(\mathcal{L}_{a,f}) \cap \mathsf{ext}(\mathcal{E}_{\mu,f})$ consists of exactly one pm R. For this R the Pythagorean identity holds

$$(21) \qquad D(P\|Q) = D(P\|R) + D(\mathcal{L}_{a,f}\|Q)\,, \qquad P \in \mathcal{L}_{a,f},\ Q \in \mathsf{ext}(\mathcal{E}_{\mu,f})\,.$$

The pm R belongs to the component $\mathcal{E}_{\nu,f}$ of $\mathsf{ext}(\mathcal{E}_{\mu,f})$ based on $\nu = \mu^{f^{-1}(\mathsf{cl}(F))}$ where F is the face of $\mathsf{cc}(\mu_f)$ containing a in its relative interior.

The proof is immediate from Theorem 3. Namely, by the assumption, $D(\mathcal{L}_a\|\mathcal{E}_{\mu_f})$ is finite and then the intersection of $\mathsf{cl}_I(\mathcal{L}_a)$ and $\mathsf{ext}(\mathcal{E}_{\mu_f})$ consists of a single pm $\tilde{R}$. The pm R equal to $\tilde{R}_{f^{-1},\mu}$ is easily seen to have the claimed properties. In particular, (21) follows from the Pythagorean identity (9) (with P, Q, R replaced by $\tilde{P}, \tilde{Q}, \tilde{R}$) since for each $P \in \mathcal{L}_{a,f}$ and $Q \in \mathsf{ext}(\mathcal{E}_{\mu,f})$ such that not both $D(P\|Q)$ and $D(P\|R)$ are infinite

$$\begin{aligned} D(P\|Q) - D(P\|R) &= \int_X \ln\left[\frac{dR}{d\mu}(x) \Big/ \frac{dQ}{d\mu}(x)\right] P(dx) = \int_X \ln\left[\frac{d\tilde{R}}{d\mu_f}(f(x)) \Big/ \frac{dQ_f}{d\mu_f}(f(x))\right] P(dx) \\ &= \int_{\mathbb{R}^d} \ln\left[\frac{d\tilde{R}}{dQ_f}(\tilde{x})\right] P_f(d\tilde{x}) = D(P_f\|Q_f) - D(P_f\|\tilde{R}) = D(\mathcal{L}_a\|Q_f)\,. \end{aligned}$$

To sketch the implications of our results for ML estimation in the exponential family $\mathcal{E} = \mathcal{E}_{\mu,f}$, let $\boldsymbol{x} = (x^1, \ldots, x^n) \in X^n$ be an i.i.d. sample drawn from a member of $\mathcal{E}_{\mu,f}$ and $a = \frac{1}{n}\sum_{i=1}^n f(x^i)$ denote the sample mean of the directional statistic f. The LLF is defined similarly to (17) where now F is a face of $\mathsf{cc}(\mu_f)$ and

$$\Lambda_F(\vartheta) = \int_{f^{-1}(\mathsf{cl}(F))} e^{\langle\vartheta, f(x)\rangle}\,\mu(dx) = \int_{\mathsf{cl}(F)} e^{\langle\vartheta,\tilde{x}\rangle}\,\mu_f(d\tilde{x})\,.$$

As an example let us discuss an extension of Theorem 6.

For $a \in \mathsf{dom}(\Lambda^*_{\mu,f})$, there exists a unique pm R^*_a such that for each $Q \in \mathsf{cl}_{rI}(\mathcal{E})$

$$\Lambda_F(\vartheta) + \Lambda^*(a) - \langle\vartheta, a\rangle \geqslant D(R^*_a\|Q)$$

where F is the face of $\mathsf{cc}(\mu_f)$ such that Q belongs to the component $\mathcal{E}^F_{\mu,f}$ of $\mathsf{ext}(\mathcal{E})$ and ϑ parametrizes Q.

The above pm R^*_a equals $\tilde{R}_{f^{-1},\mu}$ where $\tilde{R}$ is the pm playing the role of R^*_a in Theorem 6 for $\mathcal{E}_{\mu_f}$. The analogue of Corollary 11 obviously holds for this R^*_a, hence it can be interpreted as a GMLE.

To apply the results of this paper to the families $\mathcal{L}_{a,f}$ and $\mathcal{E}_{\mu,f}$, the convex core $\mathsf{cc}(\mu_f)$ of the f-image of μ, and its faces have to be determined, which may be nontrivial. If the convex core turns out to be an open set then it equals the domain of Λ^*, by Proposition 1*(iii)*, and, in absence of nontrivial faces, the extension and rI-closure of the corresponding exponential family will be the family itself. In this case, classical results on the information projections and ML estimation suffice.

Another situation that has been well-understood is when X is finite. Then, the image μ_f of μ has finite support and $\mathit{cc}(\mu_f) = \mathit{cs}(\mu_f) = \mathit{dom}(\Lambda^*)$ is a polytope. Obviously, $\mathit{dom}(\Lambda) = \mathbb{R}^d$, and the extension and rI-closure of $\mathcal{E}_{\mu,f}$ coincide with the closure of $\mathcal{E}_{\mu,f}$ viewed as a subset of $\mathbb{R}^X$. Similarly, $\mathcal{L}_{a,f}$ is closed in $\mathbb{R}^X$ and coincides with $\mathit{cl}_I(\mathcal{L}_{a,f})$. Hence the Pythagorean identity (21) takes place with $D(\mathcal{L}_{a,f}\|Q)$ replaced by $D(R\|Q)$, see also (10) in Remark 5. For $a \in \mathit{cs}(\mu_f)$ the GMLE R^*_a exists and coincides with the MLE in $\mathit{cl}_{rI}(\mathcal{E}_{\mu,f})$, see also Barndorff-Nielsen [2, pp. 154-5].

8.2. **Example with moment statistics.** The last part of this section illustrates the above generalizations when $X = \mathbb{R}$ and $f(x) = (x, x^2, \ldots, x^d)$. Let μ be a finite measure on $\mathbb{R}$; denote by S the support of μ (the smallest closed subset of $\mathbb{R}$ with full μ-measure) and by Y the subset of S consisting of the points with positive μ-measure. Then the image μ_f of μ is supported by the subset $f(S)$ of the curve $f(\mathbb{R})$ in $\mathbb{R}^d$, known as the *moment curve*, see [23]. We exclude the case when S is a finite set that is covered by the last paragraph of the previous subsection.

The moment curve intersects any hyperplane $H = \{\tilde{x} \colon \langle\vartheta, \tilde{x}\rangle = r\}$ with nonzero $\vartheta = (\vartheta_1, \ldots, \vartheta_d)$ in at most d points because $f(x) \in H$ implies that x is a real root of the polynomial $\sum_{i=1}^d \vartheta_i x^i - r$ of degree at most d. This fact implies that any $1 \leqslant k \leqslant d+1$ points of the moment curve $f(\mathbb{R})$ are affinely independent, that is, span a simplex of dimension $k-1$. In particular, the convex hull $\mathit{conv}(f(S))$ of $f(S)$ has nonempty interior because S is not finite.

A description of $\mathit{cc}(\mu_f)$ and its faces is summarized bellow; for proofs of the following two propositions see Appendix C.

Proposition 2. *(i) The interior of $\mathit{cc}(\mu_f)$ is equal to that of $\mathit{conv}(f(S))$.*

(ii) Each proper face F of $\mathit{cc}(\mu_f)$ is exposed and equals a simplex $\mathit{conv}(T)$ with $T \subseteq f(Y)$ of cardinality at most d. In addition, $\mu_f^F = \mu_f^T$.

(iii) Each set $T \subseteq f(Y)$ of cardinality $1 \leqslant k \leqslant \frac{d}{2}$ spans a face of $\mathit{cc}(\mu_f)$. If Y is contained in the interior of S then all proper faces of $\mathit{cc}(\mu_f)$ are of this form.

It is an elementary fact that the standard exponential family based on a measure whose support is an affinely independent set $T \subset \mathbb{R}^d$, equals the set of all pm's with support T. Then, it follows from Proposition 2 that each component $\mathcal{E}^F_{\mu,f}$ of the extension of $\mathcal{E}_{\mu,f}$, corresponding to a proper face $F = \mathit{conv}(T)$ of $\mathit{cc}(\mu_f)$, where $T \subseteq f(Y)$, is equal to the set of all pm's with support T. Moreover, if Y is contained in the interior of S then $\mathit{ext}(\mathcal{E}_{\mu,f})$ equals the union of $\mathcal{E}_{\mu,f}$ and the set of all pm's whose support is a subset of size $\leqslant \frac{d}{2}$ of Y.

Proposition 3. $\mathit{cl}_{rI}(\mathcal{E}_{\mu,f}) = \mathit{ext}(\mathcal{E}_{\mu,f})$

Remark 13. The inclusion $\mathit{cc}(\mu_f) \subseteq \mathit{dom}(\Lambda^*)$, see Proposition 1*(i)*, may be strict, giving another example of the equality $H = \Lambda^*$ on $\mathit{cc}(\mu_f) = \mathit{dom}(H)$ but not beyond this set, see Remark 10 in Section 5. Indeed, take $d = 3$, thus $f(x) = (x, x^2, x^3)$, and let μ be the sum of the unit mass at the origin and the measure with Lebesgue density $e^{-|x|^3}$. Then $S = \mathbb{R}$, $Y = \{0\}$, and $\mathit{cc}(\mu_f)$ consists of the interior of $\mathit{conv}(f(\mathbb{R}))$ and the point $(0,0,0)$. In particular, $\mathit{cc}(\mu_f)$ is contained in the open half-space $\{(x_1, x_2, x_3) \colon x_2 > 0\}$ except for the point $(0,0,0)$. Since for the boundary hyperplane H of this half-space we have

$\int_H e^{\langle\vartheta,x\rangle}\,\mu(dx) = 1$, for each $\vartheta \in \mathbb{R}^3$, and $\mathit{dom}(\Lambda)$ is determined by the inequality $|\vartheta_3| < 1$, Lemma 6 applied with $\tau = (0,-1,0)$ and $a = (0,0,t)$ gives $\Lambda^*(a) = \sup_{\vartheta\in \mathit{dom}(\Lambda)}\langle\vartheta, a\rangle = |t|$. Thus $\mathit{dom}(\Lambda^*)$ contains the line $\{(0,0,t)\colon t \in \mathbb{R}\}$ intersecting $\mathit{cc}(\mu_f)$ only in the single point $(0,0,0)$.

The generalization of Theorem 3 given in the previous subsection implies that for Q in $\mathcal{E}_{\mu,f}$ and a in the interior of $\mathit{cc}(\mu_f)$, the generalized I-projection $\Pi_{\mathcal{L}_{a,f}\leftarrow Q}$ belongs to $\mathcal{E}_{\mu,f}$, and equals the rI-projection to $\mathcal{E}_{\mu,f}$ of any $P \in \mathcal{L}_{a,f}$ with $D(P\|\mathcal{E}_{\mu,f})$ finite. If $a \in \mathit{ri}(F)$ for a proper face F of $\mathit{cc}(\mu_f)$, there is only one pm $\tilde{P}$ with mean a and $\tilde{P} \ll \mu_f^F$ (since μ_f^F is concentrated on the vertices of the simplex F). Then only one pm $P \in \mathcal{L}_{a,f}$ exists with $D(P\|Q)$ finite, and it is trivially the I-projection of Q to $\mathcal{L}_{a,f}$.

The GMLE exists if and only if the sample transformed by the directional statistic f has mean $a \in \mathit{dom}(\Lambda^*_{\mu,f})$. For example, if a sample with $a \in \mathit{ri}(F)$ as above is contained in Y, the GMLE equals the empirical distribution of the sample and coincides with the MLE in $\mathit{cl}_{rI}(\mathcal{E}_{\mu,f})$.

When μ is continuous, $Y = \emptyset$, then $\mathit{cc}(\mu_f)$ is open by Proposition 2*(ii)*. Thus, $\mathcal{E}_{\mu,f} = \mathit{ext}(\mathcal{E}_{\mu,f})$, and $\mathit{dom}(\Lambda^*) = \mathit{cc}(\mu_f)$ by Proposition 1*(iii)*. Now, a remarkable feature is that the MLE or GMLE never exists if the sample size is $\leqslant \frac{d}{2}$. This follows from the well-known property of the moment curve that any $k \leqslant \frac{d}{2}$ points of it span a face of the convex hull of the curve, see also the proof of Proposition 2 in Appendix C. Thus, for samples of small sizes the sample mean of f does not belong to $\mathit{cc}(\mu_f)$, μ-a.s., cf. also [13, Example 2]. This includes the well-known special case of nonexistence of MLE in the Gaussian family from a sample of size 1.

Appendix A

This appendix contains the proofs of Theorems 1, 2, and the completion of the proof of Theorem 6.

Proof of Theorem 1, "log-convex part". In this proof, all measures are given on an arbitrary measurable space. Let $D(P\|\mathcal{S})$ be finite for a pm P and a log-convex set $\mathcal{S}$. The only nontrivial assertion to prove is the existence of a pm $\Pi_{P\to\mathcal{S}}$ that satisfies inequality (2) for all $Q \in \mathcal{S}$ with $D(P\|Q)$ finite.

If Q and R are nonsingular pm's in $\mathcal{S}$ then (3) and $\overline{Q^t R^{1-t}} \in \mathcal{S}$ imply

$$tD(P\|Q) + (1-t)D(P\|R) - D(P\|\mathcal{S}) \geqslant -\ln \int q^t r^{1-t}\,d\mu \tag{22}$$

where the right-hand side is nonnegative. Let Q_n be a sequence in $\mathcal{S}$ with $D(P\|Q_n)$ converging to $D(P\|\mathcal{S})$. One can assume all Q_n dominated by a finite measure μ, with densities q_n. Applying (22) with $Q = Q_m$, $R = Q_n$, and $t = \frac{1}{2}$, it follows that

$$\ln \int \sqrt{q_n q_m}\,d\mu \to 0 \quad \text{thus} \quad \int \left[\sqrt{q_m} - \sqrt{q_n}\right]^2 d\mu = 2 - 2\int \sqrt{q_n \cdot q_m}\,d\mu \to 0$$

as $m, n \to \infty$. Hence, $\sqrt{q_n}$ is a Cauchy sequence in $L_2(\mu)$, and, therefore, converges in $L_2(\mu)$, say to $\sqrt{q^*}$. Then q_n converges to q^* in $L_1(\mu)$, and Q_n converges in total variation to the pm Q^* with μ-density q^*. If R_n is another sequence in $\mathcal{S}$ such that $D(P\|R_n)$ converges to $D(P\|\mathcal{S})$ then, by same argument, R_n also converges in total variation. Since the sequences Q_n and R_n can be merged together, the limit of R_n must be equal to Q^*. This Q^* will play the role of $\Pi_{P\to\mathcal{S}}$.

When $D(P\|R)$ is finite, (22) rewrites as

$$D(P\|Q) - D(P\|R) + \tfrac{1}{t}\left[D(P\|R) - D(P\|\mathcal{S})\right] \geqslant -\tfrac{1}{t}\ln \int q^t r^{1-t}\, d\mu\,. \tag{23}$$

Given any $Q \in \mathcal{S}$ with $D(P\|Q)$ finite, Q_n as above, and t_n going to 0 slowly, such that $t_n^{-1}\left[\, D(P\|Q_n) - D(P\|\mathcal{S})\,\right] \to 0$, one has

$$D(P\|Q) - D(P\|\mathcal{S}) \geqslant \liminf_{n\to\infty} -\frac{1}{t_n}\ln\int q^{t_n} q_n^{1-t_n}\, d\mu \geqslant \liminf_{n\to\infty} D(\overline{Q^{t_n}Q_n^{1-t_n}}\|Q) \tag{24}$$

Here, the first inequality follows from (23) with $R = Q_n$, $t = t_n$, and the second inequality from (4). The sequence $R_n = \overline{Q^{t_n}Q_n^{1-t_n}}$ in $\mathcal{S}$ satisfies $D(P\|R_n) \to D(P\|\mathcal{S})$, on account of

$$t_n D(P\|Q) - (1-t_n)D(P\|Q_n) \geqslant D(P\|R_n)\,,$$

a consequence of (3). As shown above, this implies that R_n converges to Q^* in variation distance. Hence, by lower semicontinuity, the right-most lim inf in (24) is bounded from below by $D(Q^*\|Q)$. This proves (2) with $Q^* = \Pi_{P\to\mathcal{S}}$. The last assertion of Theorem 1 obviously follows from (2), and implies the uniqueness of $\Pi_{P\to\mathcal{S}}$. □

Proof of Theorem 2. Supposing $\mathcal{E} = \mathcal{E}_\mu$, two members $P \in \mathcal{E}^F$ and $Q \in \mathcal{E}^G$ of $\mathsf{ext}(\mathcal{E})$, where F and G are faces of $\mathsf{cc}(\mu)$, are not mutually singular if and only if $\mathsf{cl}(F)\cap\mathsf{cl}(G)$ has positive μ-measure. In that case, since $\mu(\mathsf{cl}(F)\cap\mathsf{cl}(G)) = \mu(\mathsf{cl}(F\cap G))$ [13, Corollary 4], the set $F\cap G$ is nonempty, and is a face of $\mathsf{cc}(\mu)$. As the μ-densities of P and Q are equal to $e^{\langle\vartheta,x\rangle-\Lambda_F(\vartheta)}$ and $e^{\langle\tau,x\rangle-\Lambda_G(\tau)}$ on $\mathsf{cl}(F)$, respectively $\mathsf{cl}(G)$, and 0 elsewhere, the μ-density of the log-convex combination $\overline{P^tQ^{1-t}}$ is proportional to $e^{\langle t\vartheta+(1-t)\tau,x\rangle}$ on $\mathsf{cl}(F)\cap\mathsf{cl}(G)$ and 0 elsewhere. Since $\mathsf{cl}(F)\cap\mathsf{cl}(G)$ and its subset $\mathsf{cl}(F\cap G)$ have the same μ-measure, and a density can be arbitrarily changed on a set of measure 0, one can also say that $\overline{P^tQ^{1-t}}$ has μ-density proportional to $e^{\langle t\vartheta+(1-t)\tau,x\rangle}$ on $\mathsf{cl}(F\cap G)$ and 0 elsewhere. Thus, $\overline{P^tQ^{1-t}} \in \mathcal{E}^{F\cap G} \subseteq \mathsf{ext}(\mathcal{E})$, proving the log-convexity of $\mathsf{ext}(\mathcal{E})$.

The rI-closedness of $\mathsf{ext}(\mathcal{E})$ means that pm's P with $D(P\|\mathsf{ext}(\mathcal{E})) = 0$ necessarily belong to $\mathsf{ext}(\mathcal{E})$. For P having a mean, say $P \in \mathcal{L}_a = \mathcal{L}$, the assumption $D(P\|\mathsf{ext}(\mathcal{E})) = 0$ implies by (12) that P equals the pm R from Theorem 3. Thus $P \in \mathsf{ext}(\mathcal{E})$ is a simple consequence of Theorem 3, provided P has a mean. If P with $D(P\|\mathsf{ext}(\mathcal{E})) = 0$ does not have a mean, a truncation argument is needed.

We first claim that a component $\mathcal{E}^F$ of $\mathsf{ext}(\mathcal{E})$ exists such that $D(P\|\mathcal{E}^F) = 0$. To see this, pick any sequence Q_n in $\mathsf{ext}(\mathcal{E})$ with $D(P\|Q_n) \to 0$, and define another sequence R_n recursively, letting $R_1 = Q_1$, and R_n be a log-convex combination of Q_n and R_{n-1}, with $t = t_n \to 1$. Then $D(P\|R_n)$ also converges to 0 (this follows, e.g., from (3)), and by the above proof of log-convexity, the pm's R_n belong to components of $\mathsf{ext}(\mathcal{E})$ that correspond to faces of $\mathsf{cc}(\mathcal{E})$ with $F_n \supseteq F_{n+1}$, $n \geqslant 1$. Our first claim follows since then F_n must be eventually equal to a fixed face F.

Now, since clearly $\mathsf{ext}(\mathcal{E}^F) \subseteq \mathsf{ext}(\mathcal{E})$, it suffices to show that $D(P\|\mathcal{E}) = 0$ implies that P belongs to $\mathsf{ext}(\mathcal{E})$. To this end, denote by B_n the ball $\{x : \|x\| \leqslant n\} \subset \mathbb{R}^d$, and write $P_n = P(\cdot|B_n)$, $\mathcal{E}_n = \mathcal{E}_{\mu^{B_n}}$ for n sufficiently large to make $P(B_n)$ positive. Since $Q(\cdot|B_n) \in \mathcal{E}_n$ if $Q \in \mathcal{E}$, the assumption $D(P\|\mathcal{E}) = 0$ implies $D(P_n\|\mathcal{E}_n) = 0$, using the inequality (8). Since P_n has a mean, it follows that $P_n \in \mathsf{ext}(\mathcal{E}_n)$. In particular, $\mathsf{cc}(P_n)$ is a face F_n of $\mathsf{cc}(\mu^{B_n}) \subseteq B_n$, and (using that the restriction of μ^{B_n} to $\mathsf{cl}(F_n) \subseteq B_n$ equals the restriction of μ to $\mathsf{cl}(F_n)$) the logarithm of the $\mu^{\mathsf{cl}(F_n)}$-density of P_n equals an affine function, $\mu^{\mathsf{cl}(F_n)}$-almost everywhere. Noting that on the set $\mathsf{cs}(P_n) = \mathsf{cl}(F_n) \subseteq B_n$ we have

$$\ln \frac{dP}{d\mu} = \ln \frac{dP_n}{d\mu^{\mathsf{cl}(F_n)}} + \ln P(B_n),$$

it follows that on this set $\ln \frac{dP}{d\mu}$ equals an affine function μ-almost everywhere. It is not hard to see that the union of the increasing sequence of sets $\mathsf{cs}(P_n) = \mathsf{cs}(P^{B_n})$ is equal to $\mathsf{cs}(P)$. Therefore, $\ln \frac{dP}{d\mu}$ equals an affine function on the whole convex support of P, μ-almost everywhere.

To complete the proof, it remains to show that $\mathsf{cc}(P)$ is a face of $\mathsf{cc}(\mu)$. To this end, since $P \ll \mu$ obviously implies $\mathsf{cc}(P) \subseteq \mathsf{cc}(\mu)$, it suffices to verify that each segment ab contained in $\mathsf{cc}(\mu)$ and having an interior point c in $\mathsf{cc}(P)$, must be contained in $\mathsf{cc}(P)$. By [13, Lemma 11], the convex core of any finite measure equals the union of the increasing sequence of convex cores of its restrictions to the balls B_n, $n \geqslant 1$. Hence, for sufficiently large n, the segment ab is contained in $\mathsf{cc}(\mu^{B_n})$, and the point c is contained in $\mathsf{cc}(P^{B_n}) = \mathsf{cc}(P_n)$. As the latter is a face of the former, this implies that ab is contained in $\mathsf{cc}(P_n) \subseteq \mathsf{cc}(P)$. $\square$

Completion of the proof of Theorem 6. We have shown that Theorem 6 is a consequence of (20). In the case $a \in \mathsf{cc}(\mu)$, when a pm P with mean a and finite $D(P\|Q)$ exists, Theorem 1 with this P and $\mathcal{S} = \mathcal{E}$ was applied, and (20) with $R_a^* = \Pi_{P \to \mathcal{S}}$ followed from (2). It remains to show the existence of R_a^* satisfying (20) for all a in $\mathsf{dom}(\Lambda^*)$, not necessarily in $\mathsf{cc}(\mu)$.

To this end, the proof of Theorem 1 for $\mathcal{S} = \mathcal{E}$ can be modified as follows. For a pm P with mean a and $Q = Q_\vartheta$ in $\mathcal{E}$, the divergence $D(P\|Q)$ can be rewritten by (6) as $D(P\|\mu) - \langle \vartheta, a\rangle + \Lambda(\vartheta)$. With this substitution, the starting point (3) of the proof of Theorem 1 takes for $Q = Q_\vartheta$ and $R = Q_\tau$ the form

$$t\,\big[-\langle \vartheta, a\rangle + \Lambda(\vartheta)\big] + (1-t)\,\big[-\langle \tau, a\rangle + \Lambda(\tau)\big] =$$
$$= -\langle t\vartheta + (1-t)\tau, a\rangle + \Lambda(t\vartheta + (1-t)\tau) - \ln \int e^{t[\langle \vartheta, x\rangle - \Lambda(\vartheta)]}\, e^{(1-t)[\langle \tau, x\rangle - \Lambda(\tau)]}\, \mu(dx)$$

since the divergence $D(P\|\mu)$ cancels, provided it is finite. Fortunately, the above identity is obviously valid for arbitrary a. Since all inequalities of the proof of Theorem 1 were consequences of (3), and of its specialization (4) not containing the pm P, their counterparts obtained by replacing $D(P\|Q_\vartheta)$ by $-\langle \vartheta, a\rangle + \Lambda(\vartheta)$ and $D(P\|\mathcal{S})$ by $-\Lambda^*(a)$ hold. (This is as if $D(P\|\mu)$ were canceled in all equations of that proof.) Finally we arrive, exactly as in the above proof, at a pm Q^* equal to the limit in total variation of a sequence $Q_{\vartheta_n} \in \mathcal{E}$ with $\langle \vartheta_n, a\rangle - \Lambda(\vartheta_n) \to \Lambda^*(a)$, such that (20) holds for $R_a^* = Q^*$. $\square$

Appendix B

Lemma 6. *If $\mathsf{cs}(\mu)$ is contained in a half-space $\{x : \langle \tau, x - a\rangle \leqslant 0\}$, $\tau \neq 0$, but not in its boundary hyperplane H, then $\langle \vartheta, a\rangle - \Lambda(\vartheta) < \Lambda^*(a)$ for each $\vartheta \in \mathbb{R}^d$. Moreover,*

$$\Lambda^*(a) = \sup_{\vartheta \in \mathsf{dom}(\Lambda)} \big[\langle \vartheta, a\rangle - \Lambda_H(\vartheta) \big]$$

where

$$\Lambda_H(\vartheta) = \begin{cases} -\infty\,, & \mu(H) = 0 \\ \ln \int_H e^{\langle \vartheta, x\rangle}\, d\mu\,, & \mu(H) > 0\,. \end{cases}$$

In particular, if $\mathsf{dom}(\Lambda) = \mathbb{R}^d$ *then* $\Lambda^*(a) = \Lambda_H^*(a)$ *where* Λ_H^* *denotes the convex conjugate of* Λ_H*.*

Proof. For $\vartheta \in \mathsf{dom}(\Lambda)$ and arbitrary $t > 0$

(25)
$$\langle \vartheta + t\tau, a\rangle - \Lambda(\vartheta + t\tau) = -\ln \int e^{\langle \vartheta, x-a\rangle + t\langle \tau, x-a\rangle}\, \mu(dx)\,.$$

Since $\langle \tau, x-a\rangle \leqslant 0$ μ-a.e., it follows that

$$\langle \vartheta + t\tau, a\rangle - \Lambda(\vartheta + t\tau) \geqslant -\ln \int e^{\langle \vartheta, x-a\rangle}\, \mu(dx) = \langle \vartheta, a\rangle - \Lambda(\vartheta)$$

and then $\vartheta + t\tau \in \mathsf{dom}(\Lambda)$. This inequality is, however, strict because the hyperplane H does not have full μ-measure, proving the first assertion. For t growing to $+\infty$ the integral in (25) decreases to $\int_H e^{\langle \vartheta, x-a\rangle}\, d\mu$, by dominated convergence. It follows that

(26)
$$\langle \vartheta + t\tau, a\rangle - \Lambda(\vartheta + t\tau) \nearrow \langle \vartheta, a\rangle - \Lambda_H(\vartheta) \quad \text{for } t \nearrow +\infty\,.$$

Hence, $\Lambda^*(a) \geqslant \sup_{\vartheta \in \mathsf{dom}(\Lambda)}[\langle \vartheta, a\rangle - \Lambda_H(\vartheta)]$. The opposite inequality is trivial because $\Lambda \geqslant \Lambda_H$. □

Lemma 7. *(i) For an exposed face* F *of* $\mathsf{cc}(\mathcal{E})$*, the pm* $Q(\cdot|\mathsf{cl}(F))$*, obtained by conditioning* $Q \in \mathcal{E}$ *on* $\mathsf{cl}(F)$*, belongs to* $\mathsf{cl}_{rI}(\mathcal{E})$*. Moreover, to each* $Q \in \mathcal{E}$ *there exists a sequence* Q_n *in* $\mathcal{E}$ *that rI-converges to* $Q(\cdot|\mathsf{cl}(F))$ *and satisfies* $Q_n(\cdot|\mathsf{cl}(F)) = Q(\cdot|\mathsf{cl}(F))$*,* $n \geqslant 1$*.*

(ii) For $\mathcal{E}$ *such that*

(27)
$$\mathsf{ext}(\mathcal{E}) = \{Q(\cdot|\mathsf{cl}(F)) : Q \in \mathcal{E},\ F \textit{ face of } \mathsf{cc}(\mathcal{E})\}\,,$$

the assertions of (i) hold also for non-exposed faces of $\mathsf{cc}(\mathcal{E})$*. In particular, then* $\mathsf{ext}(\mathcal{E})$ *equals* $\mathsf{cl}_{rI}(\mathcal{E})$*. A sufficient condition for* (27) *is* $\mathsf{dom}(\Lambda) = \mathbb{R}^d$*.*

Proof. Note first that $Q_n(\cdot|\mathsf{cl}(F)) = Q(\cdot|\mathsf{cl}(F))$ holds if and only if the Q-density of Q_n is constant on $\mathsf{cl}(F)$. Subject to this condition, the rI-convergence of Q_n to $Q(\cdot|\mathsf{cl}(F))$ is equivalent to $Q_n(\mathsf{cl}(F)) \to 1$.

(i) Suppose F is a proper exposed face of $\mathsf{cc}(\mathcal{E})$, say $F = \mathsf{cc}(\mathcal{E}) \cap H$ for a supporting hyperplane H of $\mathsf{cc}(\mathcal{E})$. Then H is the boundary of a closed half-space $\{x : \langle \tau, x-a\rangle \leqslant 0\}$ containing $\mathsf{cs}(\mathcal{E})$ as in Lemma 6. As in the proof of that lemma, $\vartheta + n\tau$ belongs to $\mathsf{dom}(\Lambda)$, $n \geqslant 0$. The Q-density of the pm $Q_{\vartheta+n\tau} \in \mathcal{E}$ is constant on H, namely

$$\frac{\exp[\,\langle \vartheta+n\tau, x\rangle - \Lambda(\vartheta+n\tau)\,]}{\exp[\,\langle \vartheta, x\rangle - \Lambda(\vartheta)\,]} = e^{n\langle \tau, a\rangle - \Lambda(\vartheta+n\tau) + \Lambda(\vartheta)}\,.$$

Using (18), $\mu(\mathsf{cl}(F)) = \mu(H)$, and

$$Q_{\vartheta+n\tau}(\mathsf{cl}(F)) = Q_{\vartheta+n\tau}(H) = \int_H e^{\langle \vartheta+n\tau, x\rangle - \Lambda(\vartheta+n\tau)}\, d\mu = e^{n\langle \tau, a\rangle - \Lambda(\vartheta+n\tau) + \Lambda_H(\vartheta)}$$

converges to 1 as n goes to infinity, by (26).

(ii) We have to show that if $\mathcal{E}$ satisfies the assumption (27) and Q belongs to a component $\mathcal{E}^F$ of $\mathsf{ext}(\mathcal{E})$, there exist pm's $Q_n \in \mathcal{E}$ with Q-densities constant on $\mathsf{cl}(F)$ such that $Q_n(\mathsf{cl}(F)) \to 1$. We prove this by induction on the affine dimension of $\mathsf{cc}(\mathcal{E})$. There is nothing to be proved if this affine dimension is zero. Our induction hypothesis will be

the validity of the assertion for exponential families whose convex core has smaller affine dimension than $\mathsf{cc}(\mathcal{E})$.

If F is exposed, *(i)* applies. For a nonexposed face F of $\mathsf{cc}(\mathcal{E})$, there exists a proper exposed face G of $\mathsf{cc}(\mathcal{E})$ that contains F. The component $\mathcal{E}^G$ of $\mathsf{ext}(\mathcal{E})$ satisfies (27) because $\mathcal{E}$ does. By the induction hypothesis, $P_n(\mathsf{cl}(F)) \to 1$ for some sequence $P_n \in \mathcal{E}^G$ with Q-densities constant on $\mathsf{cl}(F)$. Applying *(i)* to each pm $P_n \in \mathcal{E}^G$, there exists a sequence $R_{n,m} \in \mathcal{E}$, $m \geqslant 1$, with P_n-densities constant on $\mathsf{cl}(G)$, such that $R_{n,m}(\mathsf{cl}(G)) \to 1$ as m tends to ∞. Then the Q-density of every pm $R_{n,m}$ is constant on $\mathsf{cl}(F) \subset \mathsf{cl}(G)$. Given $\varepsilon > 0$, $P_n(\mathsf{cl}(F)) > 1 - \varepsilon$ for n sufficiently large. Since $R_{n,m}$ rI-converges to P_n, as m tends to ∞, $R_{n,m}(\mathsf{cl}(F)) > P_n(\mathsf{cl}(F)) - \varepsilon$ for $m \geqslant m(n)$, sufficiently large. Thus it follows that the desired sequence Q_n exists within the array $R_{n,m} \in \mathcal{E}$. $\square$

Appendix C

The material collected here complements Subsection 8.2.

Proof of Proposition 2. (i) This follows from $\mathsf{cc}(\mu_f) \subseteq \mathsf{conv}(f(S)) \subseteq \mathsf{cs}(\mu_f)$ and the fact that $\mathsf{cc}(\mu_f)$ and $\mathsf{cs}(\mu_f)$ have the same interior, [13, Lemma 1].

(ii) Let F be an exposed proper face of $\mathsf{cc}(\mu_f)$, $F = H \cap \mathsf{cc}(\mu_f) = \mathsf{cc}(\mu_f^H)$, where H is the boundary hyperplane of a closed half-space $\{\tilde{x}\colon\ \langle \vartheta, \tilde{x}\rangle \geqslant r\}$, $\vartheta \neq 0$, that contains $f(S)$. Since at most d points of $f(S) \subseteq f(\mathbb{R})$ can be contained in H, $\mu_f^H = \mu_f^{H\cap f(S)} = \mu_f^T$ for $T = H \cap f(Y)$ of cardinality at most d. This proves that the exposed face $F = \mathsf{cc}(\mu_f^H) = \mathsf{cc}(\mu_f^T)$ equals the simplex $\mathsf{conv}(T)$. It remains to show that each proper face of $\mathsf{cc}(\mu_f)$ is exposed. This will be done when all facets of an exposed face $F = H \cap \mathsf{cc}(\mu_f) = \mathsf{cc}(\mu_f^H)$ as above (that is, the simplices $\mathsf{conv}(T \setminus \{z\})$, $z \in T$) are identified also as exposed faces of $\mathsf{cc}(\mu_f)$.

Now, the condition that the half-space contains $f(S)$ means that

$$g(x) = \sum_{i=1}^{d} \vartheta_i\, x^i - r \geqslant 0\,, \qquad x \in S \tag{28}$$

and, denoting by U the set of those roots of the polynomial g that belong to Y, the set $T = H \cap f(Y)$ of extreme points of F equals $f(U)$. We have to show that to any $y \in U$ there exists a polynomial $\tilde{g}(x)$, also of degree $\leqslant d$ and nonnegative on S, whose roots in Y are exactly the elements of the set $U \setminus \{y\}$.

Suppose $y \in U$ it is a root of multiplicity α of g. If α is even then the polynomial $\tilde{g}(x) = g(x)/(x-y)^\alpha$ is obviously suitable. If α is odd then (28) implies that some open interval with (right or left) endpoint y is disjoint from S. Taking v from such an interval, the polynomial $\tilde{g}(x) = (x-v)g(x)/(x-y)^\alpha$ will be suitable.

(iii) The last argument also shows that roots of g in the interior of S cannot have odd multiplicity. In particular, if Y is a subset of the interior of S then the polynomial g must be divisible by $\prod_{y\in U}(x-y)^2$, hence the cardinality of $T = f(U)$ is $\leqslant \frac{d}{2}$. Thus, in this case all proper faces of $\mathsf{cc}(\mu_f)$ are simplices spanned by $\ell \leqslant \frac{d}{2}$ points in $f(Y)$.

Finally, let U be any subset of size $1 \leqslant \ell \leqslant \frac{d}{2}$ of Y and let

$$g(x) = \prod_{y \in U} (x - y)^2 = \sum_{i=1}^{d} \vartheta_i \, x^i - r$$

(where $r = -\prod_{y \in U} y^2$, and $\vartheta_i = 0$ for $2\ell < i \leqslant d$). Then, clearly, the moment curve $f(\mathbb{R})$ is contained in the closed half-space $\{\tilde{x}\colon \langle \vartheta, \tilde{x} \rangle \geqslant r\}$ whose boundary hyperplane contains exactly the subset $T = f(U)$ of $f(\mathbb{R})$. As seen before, this implies that $\mathit{conv}(T)$ is a face of $\mathit{cc}(\mu_f)$. $\square$

Proof of Proposition 3. The assertion is obviously equivalent to $\mathit{cl}_{rI}(\mathcal{E}_{\mu_f}) = \mathit{ext}(\mathcal{E}_{\mu_f})$. For each proper face G of $\mathit{cc}(\mu_f)$, the conditionings $Q(\cdot|G)$ of pm's $Q \in \mathcal{E}_{\mu_f}$ belong to $\mathit{cl}_{rI}(\mathcal{E}_{\mu_f})$, due to Lemma 7*(i)* in Appendix B, since G is exposed and closed by Proposition 2. Hence it suffices to show that, for any fixed G which by Proposition 2 is a simplex with vertex set $T \subseteq f(Y)$ and $\mu_f^G = \mu_f^T$, the family $\mathcal{F} = \{Q(\cdot|G)\colon Q \in \mathcal{E}_{\mu_f}\}$ contains all pm's with support T. As $\mathcal{F}$ is clearly log-convex, this follows from Lemma 8 below if for each facet F of G there exist pm's $\tilde{Q}_n \in \mathcal{F}$ with $\tilde{Q}_n(F) \to 1$ and $\tilde{Q}_n(\cdot|F)$ not depending on n. By Proposition 2, F is an exposed face of $\mathit{cc}(\mu_f)$, thus Lemma 7*(i)* guarantees the existence of pm's $Q_n \in \mathcal{E}_{\mu_f}$ with $Q_n(F) \to 1$ and $Q_n(\cdot|F)$ not depending on n. Then the pm's $\tilde{Q}_n = Q_n(\cdot|G)$ enjoy the same properties, and this establishes our claim. $\square$

Lemma 8. *Let $\mathcal{F}$ be a nonempty log-convex set of pm's with common support equal to the set T of extreme points of a simplex in $\mathbb{R}^d$. If for each facet F of this simplex there exist pm's $Q_n \in \mathcal{F}$ with $Q_n(F) \to 1$ and $Q_n(\cdot|F)$ not depending on n then $\mathcal{F}$ consists of all pm's with support T.*

Proof. As the standard exponential family $\mathcal{E}$ based on the counting measure on T is the set of all pm's with support T, we have to show that $\mathcal{F} = \mathcal{E}$ or, equivalently, that $\Xi = \{\vartheta \in \mathbb{R}^d\colon Q_\vartheta \in \mathcal{F}\} = \mathbb{R}^d$. It suffices to consider simplices of dimension d. Then, the parametrization $\vartheta \mapsto Q_\vartheta$ is bijective.

Let F be any facet of the simplex, spanned by $T \setminus \{z\}$, $z \in T$, and τ_F be a normal vector to F with $\langle \tau_F, x \rangle = t_F > \langle \tau_F, z \rangle$, $x \in F$. For pm's Q_{ϑ_n}, $\vartheta_n \in \Xi$, in $\mathcal{F}$ satisfying the hypothesis $Q_{\vartheta_n}(F) \to 1$ and $Q_{\vartheta_n}(\cdot|F) = Q_{\vartheta_1}(\cdot|F)$, $n \geqslant 1$,

$$e^{\langle \vartheta_n, x \rangle - \Lambda(\vartheta_n)} = c_n \cdot e^{\langle \vartheta_1, x \rangle - \Lambda(\vartheta_1)}\,, \qquad x \in T \setminus \{z\}$$

where $c_n > 0$ is a constant. This implies that $\langle \vartheta_n - \vartheta_1, x \rangle$ equals a constant when $x \in F$, depending on n. It follows that $\vartheta_n - \vartheta_1$ is a scalar multiple of τ_F, say $\vartheta_n = \vartheta_1 + t_n \tau_F$, $t_n \in \mathbb{R}$, and

$$\frac{Q_{\vartheta_n}(F)}{Q_{\vartheta_n}(z)} = \sum_{x \in T \setminus \{z\}} e^{\langle \vartheta_n, x - z \rangle} = \sum_{x \in T \setminus \{z\}} e^{\langle \vartheta_1, x - z \rangle + t_n \langle \tau_F, x - z \rangle} = e^{t_n (t_F - \langle \tau_F, z \rangle)} \sum_{x \in T \setminus \{z\}} e^{\langle \vartheta_1, x - z \rangle}\,.$$

By the assumption $Q_{\vartheta_n}(F) \to 1$, this ratio goes to infinity, hence $t_n \to +\infty$. Suppose indirectly that the set Ξ, which is convex since $\mathcal{F}$ is log-convex, is a proper subset of $\mathbb{R}^d$. Then Ξ is contained in some half-space $\{\vartheta\colon \langle \vartheta, y \rangle \leqslant r\}$, $y \neq 0$, and from $\vartheta_n = \vartheta_1 + t_n \tau_F \in \Xi$ with $t_n \to +\infty$ it follows that $\langle \tau_F, y \rangle \leqslant 0$. But if some $y \neq 0$ satisfies $\langle \tau_F, y \rangle \leqslant 0$ for each face F of the simplex then, since the simplex is the set of those points x that satisfy the inequalities $\langle \tau_F, y \rangle \leqslant t_F$, the simplex would contain with any of its points x the whole half-line $\{x + sy\colon s \geqslant 0\}$, a contradiction. $\square$

Appendix D

This appendix lists certain results of [13]. The intent is to help the reader to go through our argumentation involving convex cores. The numbering of theorems, lemmas and consequences corresponds to [13]. The symbol Q used originally for a finite Borel measure on $\mathbb{R}^d$ has been replaced by P where it denoted a pm and by μ elsewhere.

Lemma 1. $\mathsf{cl}(\mathsf{cc}(\mu)) = \mathsf{cs}(\mu)$ *and* $\mathsf{ri}(\mathsf{cc}(\mu)) = \mathsf{ri}(\mathsf{cs}(\mu))$.

Lemma 2. *(i) If* C *is a convex Borel set then* $\mathsf{cc}(\mu) \cap C \supseteq \mathsf{cc}(\mu^C)$.
(ii) Let H *be a hyperplane and* $H_<, H_>$ *the open half-spaces determined by* H. *If the intersection* $\mathsf{cc}(\mu) \cap H_>$ *is empty then* $\mathsf{cc}(\mu) \cap H = \mathsf{cc}(\mu^H)$.

Corollary 2. $\mathsf{cc}(\mu) = \mathsf{ri}(\mathsf{cs}(\mu))$ *if and only if each nontrivial supporting hyperplane of* $\mathsf{cs}(\mu)$ *has measure zero.*

Lemma 3. $\mathsf{cc}(\mu^{\mathsf{cl}(F)}) = F$ *for every face* F *of* $\mathsf{cc}(\mu)$.

Corollary 4. $\mu(\mathsf{cl}(F) \cap \mathsf{cl}(G)) = \mu(\mathsf{cl}(F \cap G))$ *for any two faces* F *and* G *of* $\mathsf{cc}(\mu)$.

Theorem 1. *A convex set* $C \subseteq \mathbb{R}^d$ *is the convex core of some finite Borel measure if and only if* C *has at most a countable number of faces.*

Theorem 3. *The convex core of* μ *equals* $\left\{ \int_{\mathbb{R}^d} x\, P(dx) \colon P \text{ a pm dominated by } \mu \right\}$. *Moreover, to each* $a \in \mathsf{cc}(\mu)$ *there exists* $P \ll \mu$ *with mean* a *such that* $\frac{dP}{d\mu}$ *is bounded.*

Lemma 4. *If* $\mathsf{cc}(P)$ *is contained in a convex set* C *and the mean of* P *belongs to a face* F *of* C *then* $\mathsf{cc}(P) \subseteq F$.

Corollary 6. *If the mean of* $P \ll \mu$ *is in a face* F *of* $\mathsf{cc}(\mu)$ *then* $P \ll \mu^{\mathsf{cl}(F)}$.

Lemma 9. *Let* $X_1, X_2, \ldots$ *be i.i.d. random variables with the distribution* P. *Then*

$$Pr\left\{ \tfrac{1}{n}(X_1 + \ldots + X_n) \in \mathsf{ri}(\mathsf{cc}(P)) \right\} = P^{*n}(n\,\mathsf{ri}(\mathsf{cc}(Q))) \to 1 \,, \qquad n \to \infty \,.$$

Here, P^{*n} *is the* n*-th convolution power of* P.

References

[1] Amari, S.-I., Information Geometry on Hierarchy of Probability Distributions, *IEEE Trans.Inform. Theory* **47** pp. 1701-1711, July 2001.

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

[3] Barron, A., Limits of information, Markov chains, and projection, *Proceedings 2000 IEEE International Symposium on Information Theory*, Sorrento, Italy, p. 25.

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

[5] Chentsov, N.N., A nonsymmetric distance between probability distributions, entropy and the Pythagorean theorem. Math. Zametki 4 (1968) pp. 323–332. (in Russian)

[6] Chentsov, N.N., *Statistical Decision Rules and Optimal Inference.* Translations of Mathematical Monographs, Amer. Math. Soc., Providence – Rhode Island 1982. (Russian original: Nauka, Moscow, 1972)

[7] Chentsov, N.N., *Collected Works: Mathematics.* FIZMATLIT, Moscow, 2001. (in Russian)

[8] Csiszár, I., Informationstheoretische Konvergenzbegriffe im Raum der Wahrscheinlichkeitsverteilungen, *Publ. Math. Inst. Hungar. Acad. Sci.* **7** pp. 137–157, 1962.

[9] Csiszár, I., On topological properties of f-divergences, *Stud. Sci. Math. Hungar.* **2** pp. 329–339, 1967.

[10] Csiszár, I., I-divergence geometry of probability distributions and minimization problems, *Ann. Probab.* **3** pp. 146–158, 1975.
[11] Csiszár, I., Sanov property, generalized I-projections, and a conditional limit theeorem, *Ann. Probab.* **12** pp. 768–793, 1984.
[12] Csiszár, I., Generalized projections for nonnegative functions, *Acta Math. Hungar.* **68 (1-2)** pp. 161–185, 1995.
[13] Csiszár, I. and Matúš, F., Convex cores of measures on $\mathbb{R}^d$, *Stud. Sci. Math. Hungar.* **38** pp. 177–190, 2001.
[14] Dembo, A. and Zeitouni, O., *Large Deviations Techniques and Applications.* Springer, New York, 1998.
[15] Harremoës, P., The Information Topology, *Proc. 2002 IEEE International Symposium on Information Theory*, p. 431, Lausanne, Switzerland.
[16] Kullback, S., *Information Theory and Statistics.* Wiley, New York, 1959.
[17] Jupp, P.E. and Mardia, K.V., A note on the Maximum-Entropy Principle, *Scandinavian J. Statistics* **10** pp. 45–47, 1983.
[18] Masmoudi, A., Structure d'une fonction variance multidimensionnelle sur la frontière du domaine des moyennes, PhD. Dissertation, University of Sfax, Tunisia, 2000.
[19] Miao, W. and Hahn, M., Existence of maximum likelihood estimates for multidimensional exponential families, *Scandinavian J. Statistics* **24** pp. 371–386, 1997.
[20] Morozova, E.A. and Chentsov N.N., Natural geometry of families of probability laws, *Advances of Science and Technics (Itogi Nauki i Techniki), Ser. Contemporary Problems of Mathematics* **83** pp. 133-265, 1991. (in Russian)
[21] Rockafellar, R.T., *Convex Analysis.* Princeton University Press, 1970.
[22] Topsøe, F., Information theoretical optimization techniques, *Kybernetika* **15** pp. 7–17, 1979.
[23] Ziegler, G.M., *Lectures on Polytopes.* Springer-Verlag, New York 1995.