Optimality Conditions for Maximizers of the Information Divergence from an Exponential Family

František Matúš
Institute of Information Theory and Automation
Academy of Sciences of the Czech Republic
matus@utia.cas.cz

Abstract

The information divergence of a probability measure P from an exponential family $\mathcal{E}$ over a finite set is defined as infimum of the divergences of P from Q subject to $Q \in \mathcal{E}$. All directional derivatives of the divergence from $\mathcal{E}$ are explicitly found. To this end, behaviour of the conjugate of a log-Laplace transform on the boundary of its domain is analysed. The first order conditions for P to be a maximizer of the divergence from $\mathcal{E}$ are presented, including new ones when P is not projectable to $\mathcal{E}$.

1 Introduction

Let ν be a nonzero measure on a finite set Z and f a mapping from Z into the d-dimensional Euclidean space $\mathbb{R}^d$. The (full) *exponential family* $\mathcal{E} = \mathcal{E}_{\nu,f}$ determined by ν and the directional statistic f, see [7, 5, 6, 11], consists of the probability measures (pm's) $Q_\vartheta = Q_{\nu,f,\vartheta}$, $\vartheta \in \mathbb{R}^d$, given by

$$Q_\vartheta(z) = e^{\langle \vartheta, f(z) \rangle - \Lambda(\vartheta)}\, \nu(z)\,, \quad z \in Z\,,$$

where $\langle \cdot, \cdot \rangle$ is the scalar product on $\mathbb{R}^d$ and

$$\Lambda(\vartheta) = \Lambda_{\nu,f}(\vartheta) = \ln \textstyle\sum_{z\in Z} e^{\langle \vartheta, f(z) \rangle}\, \nu(z)\,.$$

The *information divergence* (relative entropy) of a pm P on Z from ν is

$$D(P\|\nu) = \begin{cases} \sum_{z\in s(P)} P(z) \ln \frac{P(z)}{\nu(z)}\,, & s(P) \subseteq s(\nu)\,, \\ +\infty\,, & \text{otherwise,} \end{cases}$$

This work was supported by Grant Agency of Academy of Sciences of the Czech Republic under Grant IIA 100750603.

AMS 2000 Math. Subject Classification. Primary 94A17; secondary 62B10, 60A10, 52A20.

Key words and phrases. Kullback-Leibler divergence, relative entropy, exponential family, information projection, log-Laplace transform, cumulant generating function, directional derivatives, first order optimality conditions, convex functions, polytopes.

where $\mathsf{s}(\nu) = \{z \in Z \colon \nu(z) > 0\}$ is the support of ν. The information divergence of P from the exponential family $\mathcal{E}$ is defined by

$$D(P\|\mathcal{E}) = \inf_{\vartheta\in\mathbb{R}^d} D(P\|Q_\vartheta)\,. \tag{1}$$

This work studies the maximizers of the function $P \mapsto D(P\|\mathcal{E})$, denoted also by $D(\cdot\|\mathcal{E})$, over the pm's P dominated by ν, thus satisfying $\mathsf{s}(P) \subseteq \mathsf{s}(\nu)$.

Let μ be the f-image ν_f of ν, considered for a Borel measure on $\mathbb{R}^d$. Denoting by $\mathsf{s}(\mu)$ the support $f(\mathsf{s}(\nu))$ of μ, which is the inclusion-minimal closed subset of $\mathbb{R}^d$ of the μ-measure $\mu(\mathbb{R}^d)$,

$$\Lambda(\vartheta) = \Lambda_\mu(\vartheta) = \ln \textstyle\sum_{x\in\mathsf{s}(\mu)} e^{\langle\vartheta,x\rangle}\,\mu(x)$$

whence Λ equals the *log-Laplace transform* (cumulant generating function) of μ. In terms of the *conjugate* Λ^* of Λ [14, §12],

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

the information divergence of a pm P from the exponential family $\mathcal{E}$ rewrites to

$$D(P\|\mathcal{E}) = D(P\|\nu) - \Lambda^*(\mathsf{m}(P_f)) \tag{2}$$

where

$$\mathsf{m}(P_f) = \textstyle\sum_{x\in\mathsf{s}(P_f)} x\,P_f(x) = \sum_{z\in Z} f(z)\,P(z)$$

is the mean of the f-image P_f of P. Hence, $D(\cdot\|\mathcal{E})$ expresses as difference of the strictly convex function $P \mapsto D(P\|\nu)$, denoted by $D(\cdot\|\nu)$, and the function $P \mapsto \Lambda^*(\mathsf{m}(P_f))$, which is convex because Λ^* is convex and $P \mapsto \mathsf{m}(P_f)$ is linear.

From now on assume $\mathsf{s}(\nu) = Z$.

This work is organized as follows. After collecting notations and reviewing necessary known facts in Section 2 directional behavior of the conjugate Λ^* on a boundary of its domain is described in Theorem 3.1 of Section 3. Consequently in Section 4, it is shown relying on (2) that the one-sided directional derivatives of the function $D(\cdot\|\mathcal{E})$ at any pm P exist. They may take the values $\pm\infty$. Explicit formulas for the derivatives are presented in Theorems 4.1 and 4.3. The first order optimality conditions for a pm P to be a maximizer of $D(\cdot\|\mathcal{E})$ emerge by requiring the derivatives not to be positive, see Theorem 5.1 in Section 5. Finally, Section 6 is devoted to a proof of Theorem 3.1.

The maximization of $D(\cdot\|\mathcal{E})$ has emerged in probabilistic models for evolution and learning in neural networks that are based on infomax principles [1, 2]. The divergence from an exponential family can be related to information theoretic measures for interdependence of stochastic units and its maximization reveals stochastic systems with high complexity w.r.t. an exponential family [3]. Dynamical versions of the problem of interactions in recurrent networks appeared in [1, 4, 15]. Two special instances of the maximization (1) are attacked in [13]. Further relations to previous works [1, 12] on this problem are discussed in remarks of Section 5.

2 Preliminaries

This section reviews well-known facts about the log-Laplace transforms, their conjugates and exponential families, and introduces necessary notations. Some of the assertions presented below are valid in general [5, 6, 11], but only positive measures μ on $\mathbb{R}^d$ concentrated on finite sets are considered here.

Let $Q_{\mu,\vartheta}$ denote the pm with μ-density $x \mapsto e^{\langle\vartheta,x\rangle-\Lambda(\vartheta)}$, $\vartheta \in \mathbb{R}^d$, and $\mathcal{E}_\mu$ the family of all such pm's, thus the standard exponential family determined by μ and the identity on $\mathbb{R}^d$.

Fact 2.1. $\mathsf{m}(Q_{\mu,\vartheta}) = \nabla\Lambda(\vartheta)$, $\vartheta \in \mathbb{R}^d$.

In accordance with [14], the affine hull of a set $B \subseteq \mathbb{R}^d$ is denoted by $\mathsf{aff}(B)$, the shift of $\mathsf{aff}(B)$ containing the origin by $\mathsf{lin}(B)$ and the relative interior of B by $\mathsf{ri}(B)$, which is the interior of B in the topology of $\mathsf{aff}(B)$.

Since μ is concentrated on a finite set the convex support $\mathsf{cs}(\mu)$ of μ, which is the inclusion-minimal closed convex subset of $\mathbb{R}^d$ of the μ-measure $\mu(\mathbb{R}^d)$, is the polytope spanned by $\mathsf{s}(\mu)$. For $B = \mathsf{cs}(\mu)$ the above notations are abbreviated to $\mathsf{aff}(\mu)$, $\mathsf{lin}(\mu)$ and $\mathsf{ri}(\mu)$.

Fact 2.2. *The restriction of $\nabla\Lambda$ to $\mathsf{lin}(\mu)$ is injective and onto $\mathsf{ri}(\mu)$.*

Since Λ is smooth this restriction is a diffeomorphism. Its inverse is denoted in the sequel by $\psi = \psi_\mu$.

The orthogonal complement of a linear subspace E of $\mathbb{R}^d$ is denoted by $E^\perp$.

Fact 2.3. *The equality $Q_{\mu,\vartheta} = Q_{\mu,\theta}$ holds if and only if $\vartheta - \theta \in \mathsf{lin}(\mu)^\perp$.*

It follows that the mean parametrization $a \mapsto Q_{\mu,\psi(a)}$ of $\mathcal{E}_\mu$ by the elements of $\mathsf{ri}(\mu)$ is bijective.

Fact 2.4. *Each function $\langle\cdot,a\rangle-\Lambda$, $a \in \mathsf{aff}(\mu)$, is constant on $c+\mathsf{lin}(\mu)^\perp$, $c \in \mathbb{R}^d$.*

Fact 2.5. *If $a \in \mathsf{ri}(\mu)$ then $\Lambda^*(a) = \langle\psi(a),a\rangle - \Lambda(\psi(a))$.*

The following assertion is a consequence of [8, Lemma 6].

Fact 2.6. *If $a \in \mathsf{cs}(\mu) \setminus \mathsf{ri}(\mu)$ then $+\infty > \Lambda^*(a) > \langle\vartheta,a\rangle - \Lambda(\vartheta)$, $\vartheta \in \mathbb{R}^d$.*

Hence, the convex conjugate Λ^* is finite on the polytope $\mathsf{cs}(\mu)$, thus continuous. This and (2) imply that the function $D(\cdot\|\nu)$ is continuous and, in turn, has a global maximizer.

A consequence of above facts is stated for convenience.

Fact 2.7. *If $\mathsf{m}(Q_{\mu,\vartheta}) = a$ then $\Lambda^*(a) = \langle\vartheta,a\rangle - \Lambda(\vartheta)$.*

Fact 2.8. *If $a \in \mathsf{ri}(\mu)$ then for $b \in \mathsf{cs}(\mu)$*

$$\Lambda^*(b) = \Lambda^*(a) + \langle\psi(a), b-a\rangle + o(\|b-a\|)\,.$$

If $A \subseteq \mathbb{R}^d$ then $B \mapsto \mu(B \cap A)$ is the restriction of μ by A. Let Λ_μ, ψ_μ, $Q_{\mu,\vartheta}$, etc., with μ replaced by its restriction be denoted as Λ_A, ψ_A, $Q_{A,\vartheta}$, etc., provided the restriction is nonzero.

Fact 2.9. *If $a \in F$ for a face F of $cs(\mu)$ then $\Lambda^*_\mu(a) = \Lambda^*_F(a)$.*

The following assertion is a special instance of [10, Theorem 4.1].

Fact 2.10. *If $a \in ri(F)$ for a face F of $cs(\mu)$ then*

$$\Lambda^*(a) - \big[\langle\vartheta, a\rangle - \Lambda(\vartheta)\big] \geqslant D(Q_{F,\psi_F(a)} \| Q_{\mu,\vartheta}), \qquad \vartheta \in \mathbb{R}^d.$$

Suppose in the remaining part of this section that $\mu = \nu_f$ as in the introduction. Then $Q_{\mu,\vartheta}$ is the f-image of $Q_{\nu,f,\vartheta}$, $\vartheta \in \mathbb{R}^d$. Taking the f-images of pm's from $\mathcal{E}_{\nu,f}$ is a bijection onto $\mathcal{E}_\mu$. For a face F of $cs(\mu)$ the pm $Q_{F,\theta}$ is the f-image of the pm $Q_{Y,f,\theta}$, $\theta \in \mathbb{R}^d$, where the latter denotes the pm obtained from $Q_{\nu,f,\theta}$ when ν is replaced by its restriction to $Y = f^{-1}(F)$. Taking the f-images of pm's from $\mathcal{E}_{Y,f}$ is a bijection onto $\mathcal{E}_F$. Note that $D(Q_{F,\theta} \| Q_{\mu,\vartheta})$ equals $D(Q_{Y,f,\theta} \| Q_{\nu,f,\vartheta})$, using that f is sufficient. This, Fact 2.10 and [8, Theorem 1] combine to the following assertion.

Fact 2.11. *If P is any pm on Z with $a = m(P_f)$ in $ri(F)$ for a face F of $cs(\mu)$ and $\mathcal{E} = \mathcal{E}_{\nu,f}$ then $\Pi_{P\to\mathcal{E}} = Q_{f^{-1}(F),f,\psi_F(a)}$ is the unique pm satisfying the Pythagorean inequality*

$$D(P\|Q) \geqslant D(P\|\mathcal{E}) + D(\Pi_{P\to\mathcal{E}}\|Q), \qquad Q \in \mathcal{E}. \tag{3}$$

The infimum in (1) is attained by some ϑ if and only if $a = m(P_f)$ belongs to $ri(\mu)$ in which case $\Pi_{P\to\mathcal{E}} = Q_{\nu,f,\psi(a)}$; this pm is called the *reverse information (rI-) projection* of P on $\mathcal{E}$ in [8]. If the infimum is not attained then P is not rI-projectable to $\mathcal{E}$ and $\Pi_{P\to\mathcal{E}}$ is the generalized rI-projection.

Though we do not need it in the sequel let us remark that the (variation) closure of $\mathcal{E}_{\nu,f}$ resp. $\mathcal{E}_\mu$ is equal to union of the families $\mathcal{E}_{f^{-1}(F),f}$ resp. $\mathcal{E}_F$ over the faces F of $cs(\mu)$, for a general result see [9]. The closures are bijectively parameterized by means of pm's exhausting $cs(\mu)$. It is also not difficult to deduce that for $\mathcal{E} = \mathcal{E}_{\nu,f}$ and a pm P

$$D(P\|\mathcal{E}) = D(P\|cl(\mathcal{E})) = \min\nolimits_{Q\in cl(\mathcal{E})} D(P\|Q).$$

where the minimum is attained uniquely by $Q = \Pi_{P\to\mathcal{E}}$.

Given a pm P on Z and a set $Y \subseteq Z$ with $P(Y) > 0$ let P^Y denote the pm, called *truncation* in [7], given by $P^Y(z) = P(z)/P(Y)$ for $z \in Y$ and $P^Y(z) = 0$ otherwise. Note that the set $\{Q \in \mathcal{E}_{\nu,f} \colon Q^Y = P^Y\}$, though not given via a directional statistic, is a full exponential family provided it is nonempty. The same holds for $\{Q_{\nu,f,\vartheta} \colon \vartheta \in E\}$ whenever E is a linear subspace of $\mathbb{R}^d$.

3 On the conjugate of log-Laplace transform

In this section, μ is a positive measure on $\mathbb{R}^d$ concentrated on a finite set.

Each point a of the polytope $\mathsf{cs}(\mu)$ belongs to the relative interior $\mathsf{ri}(F)$ of a unique face F of $\mathsf{cs}(\mu)$. If $b \in F$ then Facts 2.8 and 2.9 combine to

$$\Lambda^*(a+\varepsilon(b-a)) = \Lambda^*(a) + \varepsilon\,\langle \psi_F(a), b-a\rangle + o(\varepsilon)\,, \tag{4}$$

describing the directional behavior of the function $\varepsilon \mapsto \Lambda^*(a+\varepsilon(b-a))$ in a neighborhood of 0.

Let C denote the convex hull of $\mathsf{s}(\mu)\setminus F$ and $C_+ = C + \mathsf{lin}(F)$.

If $b \in \mathsf{cs}(\mu)\setminus F$ then it is not difficult to see that there exists a positive t such that $a+t(b-a)$ belongs to C_+ and $a \notin C_+$, see Lemmas 6.1 and 6.2. Then such a minimal $t>0$ exists. Denote this number by t_{ab} and the nearest point $a+t_{ab}(b-a)$ of C_+ from a in the direction $b-a$ by x_{ab}.

Let $\Xi = \psi_F(a) + \mathsf{lin}(F)^\perp$ and

$$\Psi^*_{C,\Xi}(x) = \sup\nolimits_{\theta\in\Xi}\,\big[\,\langle\theta,x\rangle - \Lambda_C(\theta)\,\big]\,, \qquad x \in \mathbb{R}^d\,.$$

By Lemma 6.8 and Fact 2.5, $\Psi^*_{C,\Xi}(x_{ab})$ is finite.

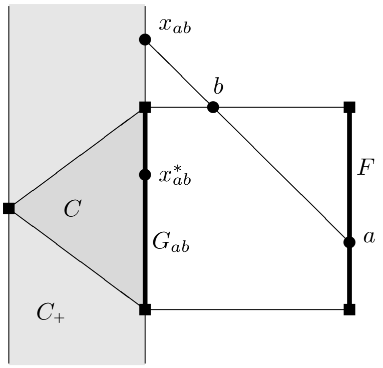
Theorem 3.1. *If $a \in \mathsf{ri}(F)$ for a face F of $\mathsf{cs}(\mu)$, $b \in \mathsf{cs}(\mu)\setminus F$ and $\varepsilon > 0$ then*

$$\Lambda^*(a+\varepsilon\,t_{ab}\,(b-a)) = \Lambda^*(a) + h(\varepsilon) + \varepsilon\,\big[\,\Psi^*_{C,\Xi}(x_{ab}) - \Lambda^*(a)\,\big] + o(\varepsilon)$$

where $h(\varepsilon) = \varepsilon\ln\varepsilon + (1-\varepsilon)\ln(1-\varepsilon)$.

The proof of Theorem 3.1, preceded by several lemmas, is presented in Section 6.

The following figure illustrates the notations presented above or used later in proofs: the support of μ consists of five black squares, F is the vertical edge of the pentagon $\mathsf{cs}(\mu)$, C is a triangle and C_+ an infinite strip.



4 Directional derivatives of $D(\cdot\|\mathcal{E})$

In this section, $\mu = \nu_f$, $\mathcal{E} = \mathcal{E}_{\nu,f}$, P and R are pm's on Z, $a = \mathsf{m}(P_f)$ belongs to $\mathsf{ri}(F)$ for a face F of $\mathsf{cs}(\mu)$, $\vartheta = \psi_F(a)$, $b = \mathsf{m}(R_f)$, and $r = R(Z \setminus \mathsf{s}(P))$.

As well-known, the one-sided directional derivative of $D(\cdot\|\mathcal{E})$ at P in the direction $R - P$ is given by

$$\lim_{\varepsilon\to 0+} \tfrac{1}{\varepsilon} \left[D(P + \varepsilon(R-P)\|\mathcal{E}) - D(P\|\mathcal{E}) \right]$$

provided the limit, finite or infinite, exists. If P dominates R then the limiting $\varepsilon \to 0$ makes sense and gives rise to a two-sided derivative.

Theorem 4.1. *If $b \in F$ and $r = 0$ then the two-sided derivative of $D(\cdot\|\mathcal{E})$ at P in the direction $R - P$ equals*

$$\textstyle\sum_{z\in\mathsf{s}(P)} [R(z) - P(z)] \ln \frac{P(z)}{e^{\langle\vartheta, f(z)\rangle}\nu(z)} \,. \tag{5}$$

If $b \in F$ and $r > 0$ then the one-sided directional derivative of $D(\cdot\|\mathcal{E})$ at P in the direction $R - P$ equals $-\infty$.

If $b \notin F$ then this derivative is equal to

$$\begin{cases} +\infty\,, & rt_{ab} < 1\,,\\ -\infty\,, & rt_{ab} > 1\,,\\ T - r\left[\Psi^*_{C,\Xi}(x_{ab}) - \Lambda^*(a) + \ln r\right], & rt_{ab} = 1\,, \end{cases}$$

where

$$T = \textstyle\sum_{z\in\mathsf{s}(R)\setminus\mathsf{s}(P)} R(z) \ln \frac{R(z)}{\nu(z)} + \sum_{z\in\mathsf{s}(P)} [R(z) - P(z)] \ln \frac{P(z)}{\nu(z)} \,.$$

A proof invokes the following simple assertion, demonstrated for readers convenience at the end of Section 6.

Lemma 4.2. *If $\varepsilon > 0$ then*

$$D(P + \varepsilon(R-P)\|\nu) = D(P\|\nu) + h(\varepsilon)\, r + \varepsilon\, T + o(\varepsilon)\,.$$

If additionally $r = 0$ then this holds also for $\varepsilon \leqslant 0$ with the $h(\varepsilon)$-term omitted.

Proof of Theorem 4.1. If $b \in F$ and $r = 0$ then $\mathsf{s}(R) \subseteq \mathsf{s}(P)$, and on account of (2) the derivative equals the difference of coefficients at the ε-terms in Lemma 4.2 and (4)

$$\textstyle\sum_{z\in\mathsf{s}(P)} [R(z) - P(z)] \ln \frac{P(z)}{\nu(z)} - \langle\vartheta, b - a\rangle$$

which rewrites to (5).

By the same argument, if $b \in F$ and $r > 0$ then the one-sided derivative is equal to $-\infty$, due to the nonzero term $h(\varepsilon)\, r$ in Lemma 4.2.

If $b \notin F$ then the formula of Theorem 3.1 is equivalent to

$$\Lambda^*(a + \varepsilon\,(b-a)) = \Lambda^*(a) + h(\varepsilon)\, \frac{1}{t_{ab}} + \frac{\varepsilon}{t_{ab}} \left[\Psi^*_{C,\Xi}(x_{ab}) - \Lambda^*(a) + \ln \frac{1}{t_{ab}}\right] + o(\varepsilon)\,.$$

This, (2) and Lemma 4.2 imply the last assertion. □

The case $b \notin F$ in Theorem 4.1 can be further simplified when assuming that there exist two different parallel hyperplanes H_F and H_C such that $H_F \supseteq F$ and $H_C \supseteq C$, where C is the convex hull of $\mathsf{s}(\mu) \setminus F$. This implies obviously that $F_+ = F + \mathit{lin}(C)$ and $C_+ = C + \mathit{lin}(F)$ are disjoint. The implication can be reversed: by [14, Corollary 19.3.3] disjointness of the polyhedral sets F_+ and C_+ makes it possible to separate them strongly by a hyperplane H, and then $\mathit{lin}(H)$ contains $\mathit{lin}(F)$ and $\mathit{lin}(C)$ whence shifts of H contain F and C.

Theorem 4.3. *If $b \notin F$ and $F_+ \cap C_+ = \emptyset$ then $rt_{ab} \geqslant 1$. The equality holds here if and only if $R(f^{-1}(F) \setminus \mathsf{s}(P)) = 0$ in which case the one-sided directional derivative of $D(\cdot\|\mathcal{E})$ at P in the direction $R - P$ is equal to*

$$r\big[\,D(R^Y\|\mathcal{F}) - D(P\|\mathcal{E})\,\big] + (1{-}r){\textstyle\sum_{z\in\mathsf{s}(P)}}\,[R^{\mathsf{s}(P)}(z) - P(z)]\ln\tfrac{P(z)}{e^{\langle\vartheta,f(z)\rangle}\nu(z)} \qquad (6)$$

where $Y = f^{-1}(C)$, $\mathcal{F}$ is the exponential family consisting of $Q_{Y,f,\theta}$, $\theta \in \Xi$, and the truncation $R^{\mathsf{s}(P)}$ is well-defined if $r < 1$.

Proof. The first assumption implies $R_f(F) < 1$ whence $s = R(Y)$ is positive. Then, $R = sR^Y + (1-s)Q$ for the truncation R^Y and a pm Q concentrated on $f^{-1}(F) = Z \setminus Y$. Thus, $b = \mathsf{m}(R_f)$ equals $sc + (1-s)a'$ where $c = \mathsf{m}(R^Y_f) \in C$ and $a' = \mathsf{m}(Q_f) \in F$. Rewrite $a + \frac{1}{s}(b-a)$ to $c + \frac{1-s}{s}(a'-a)$ to conclude that it belongs to C_+. The second assumption implies that $a \in F$ and C_+ are contained in two parallel hyperplanes whence $a{+}t(b{-}a) \in C_+$ for a unique t. Then, $t_{ab} = \frac{1}{s}$ and $rt_{ab} \geqslant 1$ follows from the obvious inequality $r \geqslant s$. The second assertion obtains from the equivalence of $r = s$ and $R(f^{-1}(F) \setminus \mathsf{s}(P)) = 0$. Under this equality

$$T = rD(R^Y\|\nu) + r\ln r - rD(P\|\nu) + {\textstyle\sum_{z\in\mathsf{s}(P)}}\,[R(z) - (1{-}r)P(z)]\ln\tfrac{P(z)}{\nu(z)}$$

and the derivative equals

$$r\big[\,D(R^Y\|\nu) {-} \Psi^*_{C,\Xi}(x_{ab}) {-} D(P\|\mathcal{E})\,\big] + (1{-}r){\textstyle\sum_{z\in\mathsf{s}(P)}}\,[R^{\mathsf{s}(P)}(z) - P(z)]\ln\tfrac{P(z)}{\nu(z)}$$

where the truncation $R^{\mathsf{s}(P)}$ is well-defined if $r < 1$. Since $x_{ab} = c + \frac{1-r}{r}(a'-a)$, $a' - a \in \mathit{lin}(F)$, and a' is the mean of the f-image of $R^{\mathsf{s}(P)} = Q$ provided $r < 1$,

$$\begin{aligned} r\Psi^*_{C,\Xi}(x_{ab}) &= r\Psi^*_{C,\Xi}(c) + (1{-}r)\,\langle\vartheta, a' - a\rangle \\ &= r\Psi^*_{C,\Xi}(c) + (1{-}r)\,{\textstyle\sum_{z\in\mathsf{s}(P)}}\,[R^{\mathsf{s}(P)}(z) - P(z)]\langle\vartheta, f(z)\rangle\,. \end{aligned}$$

Using also the analogue of (2)

$$D(R^Y\|\mathcal{F}) = \inf\nolimits_{\theta\in\Xi} D(R^Y\|Q_{Y,f,\theta}) = D(R^Y\|\nu) - \Psi^*_{C,\Xi}(c)$$

the above expression for the derivative rewrites to (6). □

Sometimes the above simplification of Theorem 4.1 is not available but such situations are not encountered later due to the following observation proved at the end of Section 6.

Lemma 4.4. *If $F_+ \cap C_+ \neq \emptyset$ then for some pm Q concentrated on $Z \setminus f^{-1}(F)$ the derivative of $D(\cdot\|\mathcal{E})$ at P in the direction $Q - P$ is $+\infty$.*

5 Optimality conditions

The results on derivatives of the function $D(\cdot\|\mathcal{E})$ presented in the previous section imply first order conditions for a pm to be a maximizer of this function.

Theorem 5.1. *If $\mathcal{E}=\mathcal{E}_{\nu,f}$ and P is a maximizer of the function $D(\cdot\|\mathcal{E})$ then P is equal to the truncation $\Pi^{s(P)}_{P\to\mathcal{E}}$ of the rI-projection of P to $\mathcal{E}$. If additionally P is not rI-projectable to $\mathcal{E}$, thus $Y=Z\setminus s(\Pi_{P\to\mathcal{E}})$ is nonempty, then $f(Y)\subseteq H_Y$ and $f(Z\setminus Y)\subseteq H_{Z\setminus Y}$ for two different parallel hyperplanes H_Y and $H_{Z\setminus Y}$, and*

$$D(P\|\mathcal{E})\geqslant\max\left\{D(R\|\mathcal{E}^P)\colon R \text{ is a pm on } Z \text{ with } s(R)\subseteq Y\right\}$$

where $\mathcal{E}^P$ is the exponential family of those truncations Q^Y that arise from $Q\in\mathcal{E}$ with $Q^{Z\setminus Y}$ equal to $\Pi_{P\to\mathcal{E}}$.

Proof. Using the notation of Section 4 and Fact 2.11, the support f $\Pi=\Pi_{P\to\mathcal{E}}$ is equal to $f^{-1}(F)=Z\setminus Y$ and $\Pi=Q_{Z\setminus Y,f,\vartheta}$. Since P is a maximizer two-sided derivatives of $D(\cdot\|\mathcal{E})$ at P vanish, and by Theorem 4.1,

$$\textstyle\sum_{z\in s(P)}\,[R(z)-P(z)]\,\ln\frac{P(z)}{\Pi(z)}=0$$

for all R dominated by P. This implies $P=\Pi^{s(P)}_{P\to\mathcal{E}}$. Moreover, if the maximizer P is not rI-projectable then no derivative is $+\infty$ whence Theorem 4.1 and Lemma 4.4 imply the containment in hyperplanes. By Theorem 4.3, for all pm's R satisfying $r=R(Y)=1$, (6) cannot be positive, and thus $D(P\|\mathcal{E})\geqslant D(R\|\mathcal{F})$. It suffices to observe that $\mathcal{F}=\mathcal{E}^P$. To this end, observe that the truncation of $Q_{\nu,f,\theta}$ to $Z\setminus Y$ is $Q_{Z\setminus Y,f,\theta}$. On account of Fact 2.3, this equals Π if and only if $\theta-\vartheta\in lin(F)^{\perp}$, thus $\theta\in\Xi$. □

Remark 5.2. It is not difficult to reverse argumentation in the previous proof and show that if the conditions of Theorem 5.1 hold for a pm P then no derivative of $D(\cdot\|\mathcal{E})$ at P is positive.

Remark 5.3. The condition $P=\Pi^{s(P)}_{P\to\mathcal{E}}$ goes back to [1, Proposition 3.1] under the assumption that P is rI-projectable on $\mathcal{E}$.

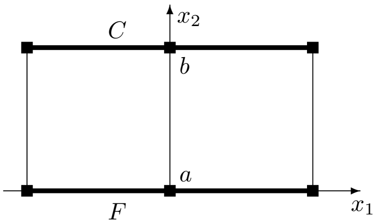
Another necessary condition can be formulated as follows.

Proposition 5.4. *If P is a maximizer of $D(\cdot\|\mathcal{E})$ then f restricted to $s(P)$ is injective and $f(s(P))$ is affine independent.*

Proof. On account of (2), the function $R\mapsto D(R\|\mathcal{E})$ is strictly convex on the polytope $\{R\colon m(R_f)=a\}$ where $a=m(P_f)$. Since P is a maximizer is must be an extreme point of this polytope by [14, Theorem 32.1]. This implies the assertions. □

Remark 5.5. As a consequence, the cardinality of $s(P)$ is at most the affine dimension of F where F is the face of $cs(\nu_f)$ with $m(P_f)\in ri(F)$. This implies that this cardinality is bounded above by 1 plus the dimension of $\mathcal{E}$, as observed in [1, Proposition 3.2] for rI-projectable pm's and in [12, Corollary 2] without this assumption.

Example 5.6. Let Z consist of six elements depicted as squares in the plane with the origin $a=(0,0)$, $b=(0,1)$, $\nu(z)=1$ for $z\in Z\setminus\{b\}$, $\nu(b)=w>0$ and f be the identity mapping on $\mathbb{R}^2$ restricted to Z, see the following picture.



Since

$$\mathit{m}(Q_{(0,u)}) = b\,\tfrac{(w+2)e^u}{3+(w+2)e^u}\,, \qquad u\in\mathbb{R}\,,$$

where the fraction equals a positive ε if and only if $e^u(w+2)=\frac{3\varepsilon}{1-\varepsilon}$

$$\psi(\varepsilon b) = \big(0\,,\ \ln\tfrac{\varepsilon}{(1-\varepsilon)}\tfrac{3}{w+2}\,\big)\,.$$

By Fact 2.5,

$$\Lambda^*(\varepsilon b) = \varepsilon\,\ln\tfrac{\varepsilon}{(1-\varepsilon)}\tfrac{3}{w+2} - \ln\big[\,3+\tfrac{3\varepsilon}{1-\varepsilon}\,\big] = -\ln 3 + h(\varepsilon) + \varepsilon\ln\tfrac{3}{w+2}$$

which is in accordance with Theorem 3.1 where $t_{ab}=1$, $x_{ab}=b$, $\Lambda^*(a)=-\ln 3$ and $\Psi^*_{\Xi,C}(b)=-\ln(w+2)$. Note that Ξ is the vertical axis and the expression $\langle\theta,b\rangle-\Lambda_C(\theta)$ is constant for $\theta\in\Xi$ by Fact 2.4.

Consider the pm's P and R concentrated on a and b, respectively. By (2),

$$D((1-\varepsilon)P+\varepsilon R\|\mathcal{E}) = \ln 3 + \varepsilon\ln\tfrac{w+2}{3w}\,.$$

This is in accordance with Theorem 4.3 where $r=1$, C is the upper horizontal edge of the rectangle $\mathit{cs}(\nu_f)$, $Y=C\cap Z$ has three elements, $R^Y=R$, $\mathcal{F}$ consists of the single pm $Q_{Y,f,\theta}$ with $\theta=(0,0)$, $D(R\|\mathcal{F})=\ln\frac{w+2}{w}$ and $D(P\|\mathcal{E})=\ln 3$.

Since $\mathit{m}(P_f)$ is outside the interior of $\mathit{cs}(\nu_f)$ the pm P is not rI-projectable on $\mathcal{E}$. Actually, $\Pi_{P\to\mathcal{E}}=Q_{Z\setminus Y,f,(0,0)}$ is uniform on $Z\setminus Y$. The pm P obviously satisfies the first two conditions of Theorem 5.1, having the edges F and C contained in two parallel lines. The third condition requires

$$\ln 3 \geqslant r_1\ln r_1(w+2) + r_2\ln r_2\tfrac{w+2}{w} + r_3\ln r_3(w+2)$$

for all nonnegative r_1, r_2 and r_3 summing to one. This is equivalent to

$$\ln 3 \geqslant \max\big\{\,\ln(w+2), \ln\tfrac{w+2}{w}\,\big\}$$

which holds only for $w=1$. In this case, all known necessary conditions cannot decide whether P is a maximizer or not. On the other hand, it is not difficult to prove that $D(\cdot\|\mathcal{E})\leqslant\ln 3$ so that P is a global maximizer.

6 Proof of Theorem 3.1

Recall the assumptions that μ is a positive measure concentrated on a finite subset of $\mathbb{R}^d$, $a \in \mathit{cs}(\mu)$ and $b \in \mathit{cs}(\mu) \setminus F$ where F is the unique face of $\mathit{cs}(\mu)$ such that $a \in \mathit{ri}(F)$.

Lemma 6.1. *There exists $t > 0$ such that $a + t(b-a) \in C_+$.*

Proof. Write b as $\varepsilon c + (1-\varepsilon)a'$ with $c \in C$, $a' \in F$ and $0 < \varepsilon \leqslant 1$, and then for $t = \frac{1}{\varepsilon}$ express $a + t(b-a)$ as $c + t(1-\varepsilon)(a'-a)$ where the second summand belongs to $\mathit{lin}(F)$. □

Lemma 6.2. *The face F is contained in a hyperplane disjoint with C_+.*

Proof. Since F is a proper face of $\mathit{cs}(\mu)$ there exists a supporting hyperplane H of $\mathit{cs}(\mu)$ such that $H \cap \mathit{cs}(\mu) = F$. The points of $s(\mu) \setminus F$ belong to one of the open halfspaces associated to H. It follows that C_+ is contained in the halfspace, using $\mathit{lin}(F) \subseteq \mathit{lin}(H)$. □

Lemma 6.3. *If G is a face of C_+ then G equals $G + \mathit{lin}(F) = (G \cap C) + \mathit{lin}(F)$, $\mathit{ri}(G) = \mathit{ri}(G \cap C) + \mathit{lin}(F)$ and $G \cap C$ is a face of C.*

Proof. If $g \in G$ then $g \in C_+$, and thus $g = c + c'$ for some $c \in C$ and $c' \in \mathit{lin}(F)$. For $c'' \in \mathit{lin}(F)$ nonzero, g is inside the segment with endpoints $c + c' \pm c''$. Since the endpoints are in C_+ and G is a face of C_+ it contains $c + c' + c'' = g + c''$ for all $c'' \in \mathit{lin}(F)$. Therefore, $G \supseteq G + \mathit{lin}(F)$ and $c \in G \cap C$. This implies $G \subseteq (G \cap C) + \mathit{lin}(F)$, and thus the first assertion holds. The second one follows by [14, Corollary 6.6.2]. If $\varepsilon c' + (1-\varepsilon)c'' \in G \cap C$ for $c', c'' \in C$ and $0 < \varepsilon < 1$ then the $\varepsilon c' + (1-\varepsilon)c'' \in G$ and $c', c'' \in C_+$, and using that G is a face of C_+ it contains c', c''. It follows that $c', c'' \in G \cap C$ whence $G \cap C$ is a face of C. □

By this lemma, if G is the unique face of C_+ that contains x_{ab} in its relative interior then $G \cap C$, denoted in the sequel by G_{ab}, is a face of C.

Corollary 6.4. $x_{ab} \in \mathit{ri}(G_{ab}) + \mathit{lin}(F)$.

Lemma 6.5. *There exist two different parallel hyperplanes H_F, H_G such that $H_F \cap \mathit{cs}(\mu) = F$, $x_{ab} \in H_G$, $H_G \cap C = G_{ab}$ and H_G strongly separates F from $s(\mu) \setminus (F \cup G_{ab})$.*

Proof. The segment with endpoints a and x_{ab} intersects C_+ at its endpoint x_{ab}. By [14, Theorem 20.2] applied to this segment and C_+, there exists a hyperplane H through x_{ab} that separates $a \notin H$ from C_+. On the other hand, $x_{ab} \in \mathit{ri}(G)$ for a unique face G of C_+, and thus there exists a supporting hyperplane K of C_+ that intersects this set in G. Then $H \cap C_+ \supseteq G$ because H contains a point from $\mathit{ri}(G)$.

It follows that there exist nonzero θ, ϑ such that the hyperplanes H and K are defined by the equations $\langle \theta, x - x_{ab} \rangle = 0$ and $\langle \vartheta, x - x_{ab} \rangle = 0$, respectively. In addition, the scalar products vanish for $x \in G$, are nonnegative for $x \in C_+$,

$\langle \vartheta, x - x_{ab} \rangle = 0$ with $x \in C_+$ implies $x \in G$ and $\langle \theta, a - x_{ab} \rangle < 0$. Then the equation $\langle \theta + \varepsilon\vartheta, x - x_{ab} \rangle = 0$ with $\varepsilon > 0$ defines a supporting hyperplane H_ε of C_+ that intersects this set in G. Taking ε sufficiently small, $\langle \theta + \varepsilon\vartheta, a - x_{ab} \rangle < 0$, and thus H_ε separates $a \notin H_\varepsilon$ and C_+.

With such a choice of ε, let $H_G = H_\varepsilon$ and H_F be the shift of H_G containing $a \notin H_G$. By Lemma 6.3, $G = G + \mathit{lin}(F)$, and then $G \subseteq H_G$ implies that $F \subseteq H_F$. By the construction of C, the points of $\mathsf{s}(\mu)$ are either in F or in C, and thus $H_F \cap \mathsf{cs}(\mu) = F$. By the construction of H_ε, $x_{ab} \in H_G$ and $H_G \cap C_+ = G$ which implies $H_G \cap C = G_{ab}$. Then the strict separation takes place. □

Lemma 6.6. *If E is a linear subspace of $\mathbb{R}^d$, $\theta \in E$ and $x \in \mathit{ri}(\mu) + E$ then the function $\vartheta \mapsto \langle \vartheta, x \rangle - \Lambda_\mu(\vartheta)$ has a maximizer ϑ^* over the set $\theta + E^\perp$. The pm Q_{μ,ϑ^*} does not depend on the choice of ϑ^* and $x - \mathsf{m}(Q_{\mu,\vartheta^*}) \in E$.*

Proof. Applying [10, Theorem 3.1] to $\theta + E^\perp$ (in the role of Ξ, with its barrier cone equal to E) the function has a unique maximizer over the orthogonal projection of $\theta + E^\perp$ to $E_{x,\mu} = \mathit{lin}(x - \mathsf{s}(\mu))$. By Fact 2.4, $\langle \vartheta, x \rangle - \Lambda_\mu(\vartheta)$ remains unchanged when ϑ moves orthogonally to $E_{x,\mu}$, containing $\mathit{lin}(\mu)$. It follows that the function has a maximizer ϑ^* over $\theta + E^\perp$ and the difference of two such maximizers is orthogonal to $E_{x,\mu}$. By Fact 2.3, Q_{μ,ϑ^*} is independent of the choice of ϑ^*. By [10, Theorem 3.2], $x - \mathsf{m}(Q_{\mu,\vartheta^*})$ is a normal vector of $\theta + E^\perp$ at ϑ^*, thus belongs to E. □

From now on G_{ab} is abbreviated to G.

Corollary 6.7. *A maximizer ϑ^* of the function $\vartheta \mapsto \langle \vartheta, x_{ab} \rangle - \Lambda_G(\vartheta)$ with ϑ in $\Xi = \psi_F(a) + \mathit{lin}(F)^\perp$ exists, $\mathsf{m}(Q_{G,\vartheta^*})$ does not depend on its choice and $x_{ab} - \mathsf{m}(Q_{G,\vartheta^*}) \in \mathit{lin}(F)$.*

Proof. Lemma 6.6 applies to the restriction of μ to G in the role of μ, the linear space $E = \mathit{lin}(F)$, the element $\theta = \psi_F(a)$ of $\mathit{lin}(F)$ and $x = x_{ab}$, which belongs to $\mathit{ri}(G) + \mathit{lin}(F)$ by Corollary 6.4. □

The mean $\mathsf{m}(Q_{G,\vartheta^*})$, independent of ϑ^*, is denoted in the sequel by x^*_{ab}.

Lemma 6.8. $\Lambda^*_G(x^*_{ab}) + \langle \psi_F(a), x_{ab} - x^*_{ab} \rangle = \Psi^*_{C,\Xi}(x_{ab})$

Proof. By Fact 2.7, applied to $\mathsf{m}(Q_{G,\vartheta^*}) = x^*_{ab}$, where ϑ^* is a maximizer from Corollary 6.7, $\Lambda^*_G(x^*_{ab}) = \langle \vartheta^*, x^*_{ab} \rangle - \Lambda_G(\vartheta^*)$. Since $\vartheta^* - \psi_F(a)$ is orthogonal to $\mathit{lin}(F)$, containing $x_{ab} - x^*_{ab}$,

$$\Lambda^*_G(x^*_{ab}) + \langle \psi_F(a), x_{ab} - x^*_{ab} \rangle = \langle \vartheta^*, x_{ab} \rangle - \Lambda_G(\vartheta^*) \geqslant \langle \vartheta, x_{ab} \rangle - \Lambda_C(\vartheta)\,, \quad \vartheta \in \Xi\,,$$

using $\Lambda_C \geqslant \Lambda_G$. Maximizing over ϑ, $\Psi^*_{C,\Xi}(x_{ab})$ emerges on the right.

On the other hand, Lemma 6.5 implies that there exists nonzero τ orthogonal to $\mathit{lin}(F)$ such that $\langle \tau, x - x_{ab} \rangle \leqslant 0$ holds for $x \in C$, and the equality takes place if and only if $x \in G = G_{ab}$. Hence, $\vartheta^* + t\tau \in \Xi$, $t \in \mathbb{R}$, and

$$\Psi^*_{C,\Xi}(x_{ab}) \geqslant \langle \vartheta^* + t\tau, x_{ab} \rangle - \Lambda_C(\vartheta^* + t\tau) = -\ln \textstyle\sum_{x \in \mathsf{s}(\mu) \setminus F} e^{\langle \vartheta^* + t\tau, x - x_{ab} \rangle}\, \mu(x)$$

where $\langle \vartheta^*, x_{ab} \rangle - \Lambda_G(\vartheta^*)$ emerges on the right when t grows to $+\infty$. □

Let b_ε abbreviate $a + \varepsilon\, t_{ab}(b-a)$, equal to $a + \varepsilon(x_{ab} - a)$. The convex hull of $F \cup G$ is denoted by A.

Lemma 6.9. *If $\varepsilon > 0$ is sufficiently small then $b_\varepsilon \in \mathit{ri}(A)$.*

Proof. By Corollary 6.4, $x_{ab} = c + t(a' - a)$ with $c \in \mathit{ri}(G)$, $a' \in F$ and $t \geqslant 0$. Then

$$b_\varepsilon = a + \varepsilon\bigl(c + t(a'-a) - a\bigr) = (1-\varepsilon)\Bigl[\tfrac{\varepsilon t}{1-\varepsilon}\, a' + \bigl(1 - \tfrac{\varepsilon t}{1-\varepsilon}\bigr)\, a\Bigr] + \varepsilon c\,.$$

For $\varepsilon > 0$ sufficiently small, the bracket is a convex combination of a' and $a \in \mathit{ri}(F)$ whence belongs to $\mathit{ri}(F)$. Then, b_ε is a convex combination of elements from $\mathit{ri}(F)$ and $\mathit{ri}(G)$, and the assertion follows by [14, Theorem 6.9]. □

By Lemma 6.9, if $\varepsilon > 0$ is sufficiently small then $\vartheta_\varepsilon = \psi_A(b_\varepsilon)$ is well-defined. Denote the means of $Q_{F,\vartheta_\varepsilon}$ and $Q_{G,\vartheta_\varepsilon}$ by $c_{F,\varepsilon}$ and $c_{G,\varepsilon}$, respectively. Then

$$\mathit{m}(Q_{A,\theta}) = e^{\Lambda_F(\theta) - \Lambda_A(\theta)} c_{F,\varepsilon} + e^{\Lambda_G(\theta) - \Lambda_A(\theta)} c_{G,\varepsilon}\,, \qquad \theta \in \mathbb{R}^d\,, \tag{7}$$

where the coefficients sum to 1. By Lemma 6.5, two parallel hyperplanes contain the pairs $c_{F,\varepsilon}$, a and $c_{G,\varepsilon}$, x_{ab}, and a geometric argument together with (7) imply that $b_\varepsilon = (1-\varepsilon)a + \varepsilon x_{ab}$ equals $\mathit{m}(Q_{A,\vartheta_\varepsilon}) = (1-\varepsilon)c_{F,\varepsilon} + \varepsilon c_{G,\varepsilon}$. In turn,

$$(1-\varepsilon)(c_{F,\varepsilon} - a) = \varepsilon(x_{ab} - c_{G,\varepsilon}) \tag{8}$$

and

$$\ln(1-\varepsilon) = \Lambda_F(\vartheta_\varepsilon) - \Lambda_A(\vartheta_\varepsilon) \qquad \ln\varepsilon = \Lambda_G(\vartheta_\varepsilon) - \Lambda_A(\vartheta_\varepsilon)\,. \tag{9}$$

Lemma 6.10. *If ε decreases to zero then $c_{F,\varepsilon} \to a$ and $c_{G,\varepsilon} \to x^*_{ab}$.*

Proof. The first convergence is a consequence of (8) and $c_{G,\varepsilon} \in \mathit{ri}(G)$, which is a bounded set. It implies that $\psi_F(c_{F,\varepsilon})$, which is the projection of ϑ_ε to $\mathit{lin}(F)$ by Fact 2.3, converges to $\psi_F(a)$. Hence, for a maximizer ϑ^* from Corollary 6.7

$$\begin{aligned} D(Q_{G,\vartheta_\varepsilon} \| Q_{G,\vartheta^*}) + D(Q_{G,\vartheta^*} \| Q_{G,\vartheta_\varepsilon}) &= \langle \vartheta_\varepsilon - \vartheta^*, \mathit{m}(Q_{G,\vartheta_\varepsilon}) - \mathit{m}(Q_{G,\vartheta^*})\rangle \\ = \langle \vartheta_\varepsilon - \vartheta^*, c_{G,\varepsilon} - x^*_{ab}\rangle &= \langle \psi_F(c_{F,\varepsilon}) - \psi_F(a), c_{G,\varepsilon} - x^*_{ab}\rangle \to 0\,. \end{aligned}$$

For a justification of the last equality observe that $\vartheta_\varepsilon - \psi_F(c_{F,\varepsilon})$ and $\vartheta^* - \psi_F(a)$ are orthogonal to $\mathit{lin}(F)$ while $c_{G,\varepsilon} - x^*_{ab} \in \mathit{lin}(F)$. Note that the latter is sum of $x_{ab} - x^*_{ab}$, belonging to $\mathit{lin}(F)$ by Corollary 6.7, and $c_{G,\varepsilon} - x_{ab}$, proportional to $a - c_{F,\varepsilon} \in \mathit{lin}(F)$ by (8). By Pinsker inequality, $Q_{G,\vartheta_\varepsilon} \to Q_{G,\vartheta^*}$ in variation distance which, in turn, implies $c_{G,\varepsilon} \to x^*_{ab}$. □

Let θ_ε denote the orthogonal projection of ϑ_ε to $\mathit{lin}(F) + \mathit{lin}(G)$.

Corollary 6.11. *If ε decreases to 0 then θ_ε converges.*

Proof. By Fact 2.3, $\psi_F(c_{F,\varepsilon})$ is the orthogonal projection of ϑ_ε to $\mathit{lin}(F)$, converging by Lemma 6.10. The arguments work also when F is replaced by G. □

Lemma 6.12. $\Lambda^*_\mu(b_\varepsilon) = \Lambda^*_A(b_\varepsilon) + o(\varepsilon)$

Proof. The assertion is trivial if $B = \mathsf{s}(\mu) \setminus A$ is empty. Otherwise, Lemma 6.5 implies existence of a nonzero τ such that the function $x \mapsto \langle \tau, x\rangle$ equals a constant s_F on F, a constant $s_G < s_F$ on G and is upper bounded by $s_B < s_G$ on $B = \mathsf{s}(\mu) \setminus A$. Scaling τ if necessary, $s_F - s_G = 1$. Let

$$r_\varepsilon = \Lambda_G(\theta_\varepsilon) - \Lambda_F(\theta_\varepsilon) + \ln \tfrac{1-\varepsilon}{\varepsilon}\,.$$

Since τ is orthogonal to $\mathit{lin}(F) + \mathit{lin}(G)$ the means of $Q_{F,\theta_\varepsilon + r_\varepsilon\tau}$ and $Q_{G,\theta_\varepsilon+r_\varepsilon\tau}$ are equal to $c_{F,\varepsilon}$ and $c_{G,\varepsilon}$, respectively. It follows from (7), with $\theta_\varepsilon + r_\varepsilon\tau$ in the role of θ, that the mean of $Q_{A,\theta_\varepsilon+r_\varepsilon\tau}$ equals $(1-\delta)c_{F,\varepsilon} + \delta c_{G,\varepsilon}$ where

$$\begin{aligned}\ln \tfrac{1-\delta}{\delta} &= \Lambda_F(\theta_\varepsilon + r_\varepsilon\tau) - \Lambda_G(\theta_\varepsilon + r_\varepsilon\tau)\\ &= r_\varepsilon(s_F - s_G) + \Lambda_F(\theta_\varepsilon) - \Lambda_G(\theta_\varepsilon) = \ln\tfrac{1-\varepsilon}{\varepsilon}\end{aligned}$$

by (9) and the choice of r_ε. Therefore, $\delta = \varepsilon$ and $\mathit{m}(Q_{A,\theta_\varepsilon+r_\varepsilon\tau})$ equals the mean b_ε of $Q_{A,\vartheta_\varepsilon}$. This implies

$$\Lambda^*_\mu(b_\varepsilon) \geqslant \langle\theta_\varepsilon + r_\varepsilon\tau, b_\varepsilon\rangle - \Lambda_\mu(\theta_\varepsilon + r_\varepsilon\tau) = \Lambda^*_A(b_\varepsilon) - \Lambda_\mu(\theta_\varepsilon + r_\varepsilon\tau) + \Lambda_A(\theta_\varepsilon + r_\varepsilon\tau)$$

using Fact 2.7. Here,

$$\Lambda_A(\theta_\varepsilon + r_\varepsilon\tau) = \ln\Big[e^{r_\varepsilon s_F + \Lambda_F(\theta_\varepsilon)} + e^{r_\varepsilon s_G + \Lambda_G(\theta_\varepsilon)}\Big]$$

and

$$\Lambda_\mu(\theta_\varepsilon) \leqslant \ln\Big[e^{\Lambda_A(\theta_\varepsilon + r_\varepsilon\tau)} + e^{r_\varepsilon s_B + \Lambda_B(\theta_\varepsilon)}\Big]\,.$$

Hence, the value of $\Lambda^*_\mu - \Lambda^*_A$ at b_ε is at least

$$\begin{aligned}-\ln\Big[1 + \frac{e^{r_\varepsilon s_B + \Lambda_B(\theta_\varepsilon)}}{e^{r_\varepsilon s_F + \Lambda_F(\theta_\varepsilon)} + e^{r_\varepsilon s_G + \Lambda_G(\theta_\varepsilon)}}\Big] &\geqslant -\frac{e^{r_\varepsilon(s_B - s_G) + \Lambda_B(\theta_\varepsilon) - \Lambda_G(\theta_\varepsilon)}}{e^{r_\varepsilon + \Lambda_F(\theta_\varepsilon) - \Lambda_G(\theta_\varepsilon)} + 1}\\ &= -\,\varepsilon\, e^{r_\varepsilon(s_B - s_G) + \Lambda_B(\theta_\varepsilon) - \Lambda_G(\theta_\varepsilon)}\end{aligned}$$

due to the choice of r_ε. By Corollary 6.11, θ_ε converges whence e^{-r_ε} is of the order $O(\varepsilon)$. In turn, $\varepsilon\, e^{r_\varepsilon(s_B - s_G)}$ is of the order $o(\varepsilon)$, on account of $s_B - s_G < 0$. Therefore, a lower bound to $\Lambda^*_\mu(b_\varepsilon) - \Lambda^*_A(b_\varepsilon)$ is of the order $o(\varepsilon)$. The assertion follows by mentioning that $\Lambda^*_\mu \leqslant \Lambda^*_A$. □

Proof o Theorem 3.1. By Lemma 6.12 and Fact 2.9, it suffices to prove that

$$\Lambda^*_A(b_\varepsilon) = \Lambda^*_F(a) + h(\varepsilon) + \varepsilon\,\big[\Psi^*_{C,\Xi}(x_{ab}) - \Lambda^*_F(a)\big] + o(\varepsilon)\,.$$

It follows from Fact 2.7, $b_\varepsilon = (1-\varepsilon)c_{F,\varepsilon} + \varepsilon c_{G,\varepsilon}$ and (9) that

$$\begin{aligned}\Lambda^*_A(b_\varepsilon) &= \langle\vartheta_\varepsilon, b_\varepsilon\rangle - \Lambda_A(\vartheta_\varepsilon)\\ &= (1-\varepsilon)\,\big[\langle\vartheta_\varepsilon, c_{F,\varepsilon}\rangle - \Lambda_F(\vartheta_\varepsilon) + \ln(1-\varepsilon)\big]\\ &\qquad + \varepsilon\,\big[\langle\vartheta_\varepsilon, c_{G,\varepsilon}\rangle - \Lambda_G(\vartheta_\varepsilon) + \ln\varepsilon\big]\\ &= h(\varepsilon) + (1-\varepsilon)\,\Lambda^*_F(c_{F,\varepsilon}) + \varepsilon\,\Lambda^*_G(c_{G,\varepsilon})\,.\end{aligned}$$

By Lemma 6.10 and (8), the norm of $c_{F,\varepsilon} - a \in \mathit{lin}(F)$ is of the order $o(\varepsilon)$. Then, using Fact 2.8,

$$\Lambda_F^*(c_{F,\varepsilon}) = \Lambda_F^*(a) + \langle \psi_F(a), c_{F,\varepsilon} - a \rangle + o(\varepsilon)$$

where the scalar product equals $\varepsilon \, \langle \psi_F(a), x_{ab} - c_{G,\varepsilon} \rangle + o(\varepsilon)$ by (8). Therefore,

$$\Lambda_A^*(b_\varepsilon) = \Lambda_F^*(a) + h(\varepsilon) + \varepsilon \left[\Lambda_G^*(c_{G,\varepsilon}) + \langle \psi_F(a), x_{ab} - c_{G,\varepsilon} \rangle - \Lambda_F^*(a) \right] + o(\varepsilon)\,.$$

This holds also when $c_{G,\varepsilon}$ is replaced by x_{ab}^* because $c_{G,\varepsilon} \to x_{ab}^*$ by Lemma 6.10 and Λ_G^* is continuous on $\mathit{ri}(G)$. Using Lemma 6.8, the assertion follows. □

Proof of Lemma 4.2. Let $P_\varepsilon = P + \varepsilon(R - P)$. Assuming first $\varepsilon > 0$,

$$D(P_\varepsilon \| \nu) = \textstyle\sum_{z \in s(R) \setminus s(P)} \varepsilon \, R(z) \ln \frac{\varepsilon \, R(z)}{\nu(z)} + \sum_{z \in s(P)} P_\varepsilon(z) \ln \frac{P_\varepsilon(z)}{\nu(z)}\,.$$

In the second sum,

$$\ln \tfrac{P_\varepsilon(z)}{\nu(z)} = \ln \tfrac{P(z)}{\nu(z)} + \ln \left[1 + \varepsilon \, \tfrac{R(z) - P(z)}{P(z)} \right] = \ln \tfrac{P(z)}{\nu(z)} + \varepsilon \, \tfrac{R(z) - P(z)}{P(z)} + o(\varepsilon)\,.$$

Hence,

$$\begin{aligned} D(P_\varepsilon \| \nu) = {} & r \, \varepsilon \ln \varepsilon \, + \varepsilon \textstyle\sum_{z \in s(R) \setminus s(P)} R(z) \ln \frac{R(z)}{\nu(z)} \\ & + D(P \| \nu) + \varepsilon \textstyle\sum_{z \in s(P)} [R(z) - P(z)] \left[1 + \ln \frac{P(z)}{\nu(z)}\right] + o(\varepsilon)\,. \end{aligned}$$

This and

$$\varepsilon \textstyle\sum_{z \in s(P)} [R(z) - P(z)] = -r \, \varepsilon = r \, (1 - \varepsilon) \ln(1 - \varepsilon) + o(\varepsilon)$$

imply the first assertion. If $r = 0$ the argumentation goes through also for $\varepsilon \leqslant 0$, omitting corresponding terms. □

Proof of Lemma 4.4. First, it is shown that there exists $c \in C$ such that $t_{ac} < 1$. The assumption implies $a \in \mathit{aff}(C) + \mathit{lin}(F)$. Then $a = tc' + (1 - t)c'' + b'$ for some $c', c'' \in C$, $b' \in \mathit{lin}(F)$ and $t \in \mathbb{R}$. By Lemma 6.2, $a \notin C_+$ whence t is not between 0 and 1. Changing the roles of c' and c'' if necessary it is possible to assume that $t > 1$. Let $c = c''$. It follows that $a + \frac{t-1}{t}(c - a)$ equals $c' + \frac{1}{t}\, b'$ which belongs to C_+. Hence, $t_{ac} \leqslant \frac{t-1}{t} < 1$. Obviously $c = m(Q_f)$ for some pm Q concentrated on $Z \setminus f^{-1}(F)$. Then $f^{-1}(F) \supseteq s(P)$ implies $Q(Z \setminus s(P)) = 1$, and the derivative in the direction $Q - P$ is $+\infty$, by Theorem 4.1. □

References

[1] Ay, N. (2002) An information-geometric approach to a theory of pragmatic structuring. *The Annals of Probability* **30** 416–436.

[2] Ay, N. (2002) Locality of Global Stochastic Interaction in Directed Acyclic Networks. *Neural Computation* **14** 2959–2980.

[3] Ay, N. and Knauf, A. (2006) Maximizing multi-information. (accepted to Kybernetika)

[4] Ay, N. and Wennekers, T. (2003) Dynamical properties of strongly interacting Markov chains. *Neural Networks* **16** 1483–1497.

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

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

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

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

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

[10] Csiszár, I. and Matúš, F. (2006) Generalized maximum likelihood estimates for exponential families. (submitted to Probab. Theory and Related Fields)

[11] Letac, G. (1992) *Lectures on Natural Exponential Families and their Variance Functions.* Monografias de Matemática **50**, Instituto de Matemática Pura e Aplicada, Rio de Janeiro.

[12] Matúš, F. and Ay, N. (2003) On maximization of the information divergence from an exponential family. *Proceedings of WUPES'03* (ed. J. Vejnarová), University of Economics Prague, 199–204.

[13] Matúš, F. (2004) Maximization of information divergences from binary i.i.d. sequences. *Proceedings IPMU 2004*, Perugia, Italy, Vol. 2, 1303–1306.

[14] Rockafellar, R.T. (1970) *Convex Analysis.* Princeton University Press.

[15] Wennekers, T. and Ay, N. (2003) Finite state automata resulting from temporal information maximization. *Theory in Biosciences* **122** 5–18.