František Matúš

On limiting towards the boundaries of exponential families

Kybernetika, Vol. 51 (2015), No. 5, 725–738

Persistent URL: http://dml.cz/dmlcz/144738

KYBERNETIKA — VOLUME *51* (2015), NUMBER 5, PAGES 725–738

ON LIMITING TOWARDS THE BOUNDARIES OF EXPONENTIAL FAMILIES

František Matúš

This work studies the standard exponential families of probability measures on Euclidean spaces that have finite supports. In such a family parameterized by means, the mean is supposed to move along a segment inside the convex support towards an endpoint on the boundary of the support. Limit behavior of several quantities related to the exponential family is described explicitly. In particular, the variance functions and information divergences are studied around the boundary.

Keywords: exponential family, variance function, Kullback–Leibler divergence, relative entropy, information divergence, mean parametrization, convex support

Classification: 94A17, 62B10, 60A10

1. INTRODUCTION

Let μ be a nonzero Borel measure on $\mathbb{R}^d$ with finite support $\mathsf{s}(\mu)$. The convex hull of $\mathsf{s}(\mu)$ is the polytope called the *convex support* $\mathsf{cs}(\mu)$ of μ. Let $\mathsf{aff}(\mu)$ be the affine hull of $\mathsf{s}(\mu)$ and $\mathsf{lin}(\mu)$ the shift of $\mathsf{aff}(\mu)$ containing the origin. Relative to the topology of $\mathsf{aff}(\mu)$, the interior of $\mathsf{cs}(\mu)$ is denoted by $\mathsf{ri}(\mu)$ and the boundary by $\mathsf{rbd}(\mu)$.

The (full) *exponential family* $\mathcal{E} = \mathcal{E}_\mu$ based on μ and the identity mapping on $\mathbb{R}^d$ consists of the probability measures (pm's) $Q_\vartheta = Q_{\mu,\vartheta}$, $\vartheta \in \mathbb{R}^d$, with μ-densities in the form $x \mapsto e^{\langle\vartheta,x\rangle - \Lambda(\vartheta)}$, $x \in \mathbb{R}^d$. Here, $\langle\cdot,\cdot\rangle$ is the scalar product and $\Lambda = \Lambda_\mu$ the *log-Laplace transform* of μ

$$\vartheta \mapsto \ln \sum\nolimits_{y\in\mathsf{s}(\mu)} e^{\langle\vartheta,y\rangle}\,\mu(y)\,, \qquad \vartheta \in \mathbb{R}^d\,.$$

Two measures Q_ϑ, Q_θ with $\vartheta, \theta \in \mathbb{R}^d$ coincide if and only if $\vartheta - \theta$ is orthogonal to $\mathsf{lin}(\mu)$. Thus, the family is bijectively parameterized as $\mathcal{E} = \{Q_\vartheta \colon \vartheta \in \mathsf{lin}(\mu)\}$. The pm Q_ϑ has the mean

$$\mathsf{m}(Q_\vartheta) \triangleq \sum\nolimits_{x\in\mathsf{s}(\mu)} x \cdot Q_\vartheta(x) = \sum\nolimits_{x\in\mathsf{s}(\mu)} x \cdot e^{\langle\vartheta,x\rangle - \Lambda(\vartheta)}\,\mu(x)$$

that belongs necessarily to $\mathsf{ri}(\mu)$. Actually, $\mathsf{m}(Q_\vartheta)$ equals the gradient $\Lambda'(\vartheta)$ of Λ at ϑ in the Euclidean metric. The means distinguish the pm's from $\mathcal{E}$ and exhaust $\mathsf{ri}(\mu)$. Therefore, $\mathcal{E} = \{Q_{\psi(a)} \colon a \in \mathsf{ri}(\mu)\}$ where $\psi = \psi_\mu$ maps $a \in \mathsf{ri}(\mu)$ into the unique $\vartheta \in \mathsf{lin}(\mu)$ that satisfies $\Lambda'(\vartheta) = a$. For background on exponential families see [2, 3, 4, 8].

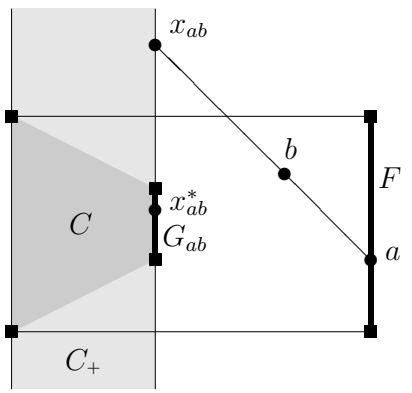
DOI: 10.14736/kyb-2015-5-0725

It follows from the above parameterizations that the mapping $P \mapsto \mathsf{m}(P)$ on the pm's P with mean restricts to a homeomorphism between $\mathcal{E}_\mu$ and $\mathsf{ri}(\mu)$. The mapping provides also a homeomorphism between the closure $\mathsf{cl}(\mathcal{E}_\mu)$ of $\mathcal{E}_\mu$ in variation distance and the convex support $\mathsf{cs}(\mu)$. In fact, $\mathsf{cl}(\mathcal{E}_\mu)$ is union of the exponential families $\mathcal{E}_\nu$ over the restrictions ν of μ to the nonempty faces F of $\mathsf{cs}(\mu)$; the family $\mathcal{E}_\nu$ is denoted preferably by $\mathcal{E}_F$, and the notation inherits to Λ_F, $Q_{F,\vartheta}$, ψ_F, etc. Then, the inversion of the homeomorphism between $\mathsf{cl}(\mathcal{E}_\mu)$ and $\mathsf{cs}(\mu)$ works as follows. Each $a \in \mathsf{cs}(\mu)$ belongs to the relative interior $\mathsf{ri}(F)$ of a unique face F of $\mathsf{cs}(\mu)$ and parameterizes the unique pm $Q_{F,\psi_F(a)}$ from $\mathcal{E}_F$ whose mean is a. The above facts go back to [2, Theorem 9.15] and [4]. For the variation closures of general exponential families see [5].

This work studies the pm's $Q_{\psi(a+\varepsilon(b-a))}$ when $a \in \mathsf{rbd}(\mu)$, $b \in \mathsf{ri}(\mu)$ and $\varepsilon \searrow 0$. By the above discussion, they converge to $Q_{F,\psi_F(a)}$. The goal is to understand the convergence in more detail. In Section 2, the main theorems are formulated and relations to a previous work discussed. Behavior of other mathematical objects related to the family $\mathcal{E}_\mu$, such as the variation function and information divergences, under this directional limiting is investigated as well. Section 3 presents proofs and Section 4 illustrates all results on the example of the multinomial family. The presentation depends heavily on the notation and results of [10] that are not repeated but only referred to here.

2. MAIN RESULTS

It is assumed throughout that $a \in \mathsf{rbd}(\mu)$ and $b \in \mathsf{ri}(\mu)$. There exists a unique proper face F of $\mathsf{cs}(\mu)$ having a in $\mathsf{ri}(F)$. Let C denote the convex hull of $\mathsf{s}(\mu) \setminus F$ and C_+ the polyhedral set $C + \mathsf{lin}(F)$.



μ sits on six points depicted as squares
F is the vertical edge on the right
C is the shaded trapezoid
C_+ is a vertical infinite strip
x_{ab} is a point in C_+
G_{ab} is the vertical edge in the middle
$x^*_{ab} = \mathsf{m}(Q_{G,\vartheta^*})$ is a point in $\mathsf{ri}(G_{ab})$

The reader may wish to follow the presentation in parallel to the example of the multinomial exponential family, see Section 4.

The limiting in $a + \varepsilon(b - a)$ with ε decreasing to zero turns out to be impractical because it gives rise to complicated formulas and more tedious computations. Instead, the scaled limiting in $b_\varepsilon \triangleq a + \varepsilon(x_{ab} - a)$ is considered where x_{ab} denotes the point from C_+ that is closest to a in the direction $b-a$, for existence see [10, Lemma 6.1]. The point $x_{ab} \in C_+$ belongs to the relative interior of a unique face of this polyhedral set.

This face intersects C in $G = G_{ab}$ which is a face of C [10, Lemma 6.3]. The exponential family $\mathcal{E}_G = \{Q_{G,\tau} \colon \tau \in \mathit{lin}(G)\}$ based on the restriction of μ to G plays a crucial role, recall [10, Corollary 6.7] saying

(i) $\vartheta \mapsto \langle \vartheta, x_{ab}\rangle - \Lambda_G(\vartheta)$ attains the maximum over ϑ with $[\vartheta - \psi_F(a)] \perp \mathit{lin}(F)$,

(ii) $\mathit{m}(Q_{G,\vartheta^*})$ does not depend on the choice of a maximizer ϑ^* in (i),

(iii) $x_{ab} - \mathit{m}(Q_{G,\vartheta^*}) \in \mathit{lin}(F)$.

The mean $\mathit{m}(Q_{G,\vartheta^*})$ belongs to $\mathit{ri}(G)$ and is denoted by x^*_{ab}.

Notations for directional derivatives follow standards, for example $\psi'_F(a;\vartheta)$ is the directional derivative of ψ_F at a in a direction $\vartheta \in \mathit{lin}(F)$. The equalities containing the terms $o(\varepsilon^\alpha)$ on the right are 'one-sided', more precisely, $=$ could be replaced by $\in$ when $o(\varepsilon^\alpha)$ were interpreted as a cone of functions, see [7].

First main result exposes an approximation for $Q_{\psi(b_\varepsilon)}$ when ε decreases to 0.

Theorem 2.1. There exists $1<\alpha<2$ such that for $\varepsilon \geqslant 0$ the probability $Q_{\psi(b_\varepsilon)}(y)$ equals

$$\begin{array}{ll} Q_{F,\psi_F(a)}(y) \cdot \big[1 - \varepsilon + \varepsilon \,\langle \psi'_F(a; x_{ab} - x^*_{ab}), y - a\rangle\big] + o(\varepsilon^\alpha)\,, & y \in \mathit{s}(\mu) \cap F\,, \\ \varepsilon \cdot Q_{G,\psi_G(x^*_{ab})}(y) + o(\varepsilon^\alpha)\,, & y \in \mathit{s}(\mu) \cap G\,, \\ o(\varepsilon^\alpha)\,, & y \in \mathit{s}(\mu) \setminus (F \cup G)\,. \end{array}$$

Roughly speaking, $Q_{\psi(b_\varepsilon)}(y)$ is affine in ε on F, decreases linearly to zero in ε on G and is negligible outside $F \cup G$.

The variance function $V = V_\mu$, see [8], of the family $\mathcal{E}_\mu$ maps $z \in \mathit{ri}(\mu)$ to the covariance $V(z)$ of the pm $Q_{\psi(z)}$, the very pm of $\mathcal{E}_\mu$ having the mean z. This covariance is a symmetric bilinear form

$$(\tau, \varsigma) \mapsto \sum\nolimits_{y \in \mathit{s}(\mu)} \langle \tau, y - z\rangle \langle \varsigma, y - z\rangle \cdot Q_{\psi(z)}(y)\,, \quad \tau, \varsigma \in \mathbb{R}^d\,.$$

The above sum is denoted by ${}^\tau V^\varsigma(z)$. The covariance is a convex combination of the elementary forms $U(z)\colon (\tau,\varsigma) \mapsto \langle \tau, z\rangle\langle \varsigma, z\rangle$, $z \in \mathbb{R}^d$. The norm $\|W\|$ of a bilinear form W is defined as the maximum of ${}^\tau W^\tau$ over $\tau \in \mathbb{R}^d$ with $\|\tau\| \leqslant 1$.

The variance function V on the segment between a and b can be approximated around the boundary of $\mathit{cs}(\mu)$ as follows.

Theorem 2.2. There exists $1 < \alpha < 2$ such that for $\varepsilon \geqslant 0$ the forms $V(b_\varepsilon)$ and

$$V_F(a) + \varepsilon \left[\, V'_F(a; x_{ab} - x^*_{ab}) - V_F(a) + V_G(x^*_{ab}) + U(x^*_{ab} - a)\,\right]$$

differ in the norm as $o(\varepsilon^\alpha)$.

The information divergence (relative entropy) of a pm P from μ is given by

$$D(P\|\mu) = \sum\nolimits_{y \in \mathit{s}(P)} P(y)\, \ln \frac{P(y)}{\mu(y)}$$

assuming $\mathit{s}(P) \subseteq \mathit{s}(\mu)$. An approximation for divergences involving $Q_{\psi(b_\varepsilon)}(y)$ is a consequence of Theorem 2.1.

Corollary 2.3. If P is a pm with $\mathsf{s}(P) \subseteq \mathsf{s}(\mu) \cap F$ then

$$D(P\|Q_{\psi(b_\varepsilon)}) = D(P\|Q_{F,\psi_F(a)}) + \varepsilon - \varepsilon \cdot \langle \psi_F'(a; x_{ab} - x_{ab}^*), \mathsf{m}(P) - a\rangle + o(\varepsilon^\alpha)\,.$$

If $P = Q_{F,\psi_F(a)}$ then the divergences $D(Q_{F,\psi_F(a)}\|Q_{\psi(b_\varepsilon)})$ behave for ε small as $\varepsilon + o(\varepsilon^\alpha)$. This is in contrast to the well-known quadratic approximation

$$D(Q_{\psi(x)}\|Q_{\psi(x+\varepsilon(y-x))}) = \tfrac{1}{2}\,\varepsilon^2\,\langle \psi'(x; y-x), y-x\rangle + O(\varepsilon^3)\,, \quad x, y \in \mathsf{ri}(\mu)\,.$$

Theorem 2.4. There exists $1 < \beta < 2$ such that for $\tau \in \mathbb{R}^d$ and $\varepsilon \geqslant 0$

$$\begin{aligned} D(Q_{\psi(b_\varepsilon)}\|Q_\tau) = (1-\varepsilon)D(Q_{F,\psi_F(a)}\|Q_\tau) + \varepsilon D(Q_{G,\psi_G(x_{\bar{a}b}^*)}\|Q_\tau) \\ +\ h(\varepsilon) + \varepsilon \cdot \langle \psi_F(a) - \tau, x_{ab} - x_{ab}^*\rangle + o(\varepsilon^\beta) \end{aligned}$$

where $h(\varepsilon) = \varepsilon \ln \varepsilon + (1-\varepsilon)\ln(1-\varepsilon)$, $0 \leqslant \varepsilon \leqslant 1$.

A single previous result that preceded the above theorems dealt with the conjugate function Λ^* of the log-Laplace transform Λ [15, Section 12],

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

It is finite on $\mathsf{cs}(\mu)$ and $+\infty$ otherwise. By [10, Theorem 3.1],

$$\Lambda^*(a + \varepsilon(b-a)) = \Lambda^*(a) + h(r\varepsilon) + s\varepsilon + o(\varepsilon)$$

where $r > 0$ and $s \in \mathbb{R}$ are explicitly available. The term $o(\varepsilon)$ will be improved in Theorem 3.4 to $o(\varepsilon^\alpha)$ with some $\alpha > 1$. This equality was a crucial tool to compute all directional derivatives of the divergence distance from exponential families. Maximization of such distances goes back to [1] and is relevant when studying probabilistic models for evolution and learning neural networks, based on infomax principles. For further insight and references see [9, 12, 13, 14].

3. PROOFS OF THEOREMS

This section collects proofs of Theorems 2.1, 2.2 and 2.4, the proof of Corollary 2.3, and supporting lemmas. In addition, [10, Theorem 3.1] is improved as promised to in Introduction, see Theorem 3.4.

The convex hull of $F \cup G$ is denoted by A and $\mathsf{s}(\mu) \setminus A$ by B. By [10, Lemma 6.5], there exist parallel hyperplanes H_F and H_G such that $F \subseteq H_F$, $G \subseteq H_G$ and H_G separates F from B strictly. In other words, there exists 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. Scaling τ if necessary, $s_F - s_G = 1$.

Let P_ε abbreviate in this section $Q_{\psi(b_\varepsilon)} = Q_{\mu,\psi_\mu(b_\varepsilon)}$. For $\varepsilon > 0$ sufficiently small b_ε belongs to $\mathsf{ri}(A)$ [10, Lemma 6.9] which is assumed in the sequel. Let ϑ_ε denote $\psi_A(b_\varepsilon)$ and θ_ε be the orthogonal projection of ϑ_ε to $\mathsf{lin}(F) + \mathsf{lin}(G)$. The pm's $Q_{F,\vartheta_\varepsilon}$ and $Q_{G,\vartheta_\varepsilon}$ are restrictions of $Q_{A,\vartheta_\varepsilon}$, itself a restriction of P_ε, and can be parameterized also by θ_ε.

A minor improvement of [10, Lemma 6.12], having $\gamma = 1$, is needed.

Lemma 3.1. $\Lambda_\mu^*(b_\varepsilon) = \Lambda_A^*(b_\varepsilon) + o(\varepsilon^\gamma)$ for $\gamma < 1 + s_G - s_B$.

P r o o f . By [10, Corollary 6.11], θ_ε converges and hence

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

This is combined with the inequalities

$$0 \geqslant \Lambda^*_\mu(b_\varepsilon) - \Lambda^*_A(b_\varepsilon) \geqslant -\varepsilon\, e^{r_\varepsilon(s_B - s_G) + \Lambda_B(\theta_\varepsilon) - \Lambda_G(\theta_\varepsilon)}$$

shown in the course of proving [10, Lemma 6.12]. □

The upper bound on γ in Lemma 3.1 depends on the geometry of $\mathsf{s}(\mu)$.

P r o o f . [Proof of Theorem 2.1]

1. *The case* $y \in B = \mathsf{s}(\mu) \setminus (F \cup G)$. The divergence $D(Q_{A,\vartheta_\varepsilon} \| P_\varepsilon)$ equals

$$\begin{aligned}&\textstyle\sum_{y \in \mathsf{s}(\mu) \cap A} Q_{A,\vartheta_\varepsilon}(y) \left[\langle \vartheta_\varepsilon, y\rangle - \Lambda_A(\vartheta_\varepsilon) - \langle \psi(b_\varepsilon), y\rangle + \Lambda_\mu(\psi(b_\varepsilon))\right]\\ &\quad = \langle \vartheta_\varepsilon - \psi(b_\varepsilon), b_\varepsilon\rangle - \Lambda_A(\vartheta_\varepsilon) + \Lambda_\mu(\psi(b_\varepsilon)) = \Lambda^*_A(b_\varepsilon) - \Lambda^*_\mu(b_\varepsilon)\end{aligned}$$

using that $Q_{A,\vartheta_\varepsilon}$ has the mean b_ε. By Lemma 3.1, $D(Q_{A,\vartheta_\varepsilon} \| P_\varepsilon) = o(\varepsilon^\gamma)$ where $\gamma > 1$. This dominates the divergence between the pm's $(1,0)$ and $(P_\varepsilon(A), P_\varepsilon(B))$, equal to $-\ln P_\varepsilon(A)$. In turn, $P_\varepsilon(B) = o(\varepsilon^\gamma)$, which is equivalent to $Q_{\psi(b_\varepsilon)}(y) = o(\varepsilon^\gamma)$ for $y \in B$.

2. *The case* $y \in G$. The pm $Q_{A,\vartheta_\varepsilon}$ is a convex combination of its restrictions Q_{F,θ_ε} and Q_{G,θ_ε} whose means are denoted by $c_{F,\varepsilon}$ and $c_{G,\varepsilon}$, respectively. Then $\mathsf{m}(Q_{A,\vartheta_\varepsilon}) = b_\varepsilon$ is a convex combination of $c_{F,\varepsilon} \in H_F$ and $c_{G,\varepsilon} \in H_G$. Since $b_\varepsilon = (1-\varepsilon)a + \varepsilon x_{ab}$ where $a \in F \subseteq H_F$ and $x_{ab} = x^*_{ab} + (x_{ab} - x^*_{ab}) \in G + \mathsf{lin}(F) \subseteq H_G$ the mean $\mathsf{m}(Q_{A,\vartheta_\varepsilon})$ equals $(1-\varepsilon)c_{F,\varepsilon} + \varepsilon c_{G,\varepsilon}$. Consequently,

$$Q_{A,\vartheta_\varepsilon} = (1-\varepsilon)Q_{F,\theta_\varepsilon} + \varepsilon Q_{G,\theta_\varepsilon}\,.$$

The pm P_ε is the convex combination of $Q_{F,\psi(b_\varepsilon)}$, $Q_{G,\psi(b_\varepsilon)}$ and $Q_{B,\psi(b_\varepsilon)}$ with the weights $P_\varepsilon(F)$, $P_\varepsilon(G)$ and $P_\varepsilon(B)$, respectively. Let $d_{F,\varepsilon}$, $d_{G,\varepsilon}$ and $d_{B,\varepsilon}$ denote the corresponding means*. Thus, b_ε is their convex combination with the same weights. Since $P_\varepsilon(B) = o(\varepsilon^\gamma)$, $d_{F,\varepsilon} \in H_F$ and $d_{G,\varepsilon} \in H_G$ the strict separation implies that

$$P_\varepsilon = [1 - \varepsilon + o(\varepsilon^\gamma)]\, Q_{F,\psi(b_\varepsilon)} + [\varepsilon + o(\varepsilon^\gamma)]\, Q_{G,\psi(b_\varepsilon)} + o(\varepsilon^\gamma)\, Q_{B,\psi(b_\varepsilon)}\,.$$

It follows that $D(Q_{A,\theta_\varepsilon} \| P_\varepsilon)$ is equal to the sum of

$$\textstyle\sum_{y \in \mathsf{s}(\mu) \cap F} (1-\varepsilon) Q_{F,\theta_\varepsilon}(y)\, \ln \frac{(1-\varepsilon)Q_{F,\theta_\varepsilon}(y)}{[1-\varepsilon+o(\varepsilon^\gamma)]Q_{F,\psi(b_\varepsilon)}(y)}$$

and

$$\textstyle\sum_{y \in \mathsf{s}(\mu) \cap G} \varepsilon Q_{G,\theta_\varepsilon}(y)\, \ln \frac{\varepsilon Q_{G,\theta_\varepsilon}(y)}{[\varepsilon+o(\varepsilon^\gamma)]Q_{G,\psi(b_\varepsilon)}(y)}\,.$$

Since the divergence is of the order $o(\varepsilon^\gamma)$ by first part of the proof,

$$(1-\varepsilon)D(Q_{F,\theta_\varepsilon} \| Q_{F,\psi(b_\varepsilon)}) + \varepsilon D(Q_{G,\theta_\varepsilon} \| Q_{G,\psi(b_\varepsilon)}) = o(\varepsilon^\gamma)\,,$$

assuming $\gamma < 2$. Then, $D(Q_{G,\theta_\varepsilon} \| Q_{G,\psi(b_\varepsilon)}) = o(\varepsilon^{\gamma-1})$.

* There is an innocent collision with the notation for the dimension d of the ambient space $\mathbb{R}^d$.

When proving [10, Lemma 6.10] the equality

$$D(Q_{G,\vartheta_\varepsilon}\|Q_{G,\vartheta^*}) + D(Q_{G,\vartheta^*}\|Q_{G,\vartheta_\varepsilon}) = \langle \psi_F(c_{F,\varepsilon}) - \psi_F(a), c_{G,\varepsilon} - x^*_{ab}\rangle$$

emerged for a maximizer ϑ^* from (i). Here, ϑ_ε can be replaced by θ_ε. Since

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

it follows from

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

that $\|c_{F,\varepsilon} - a\| = O(\varepsilon)$[†]. Knowing that $c_{G,\varepsilon}$ converges to x^*_{ab} [10, Lemma 6.10] and ψ_F is smooth on $\mathit{ri}(F)$, it follows that the above sum of two divergences is of the order $o(\varepsilon)$.

By triangle inequality for the variation norms and Pinsker inequality,

$$\begin{aligned}\|Q_{G,\psi(b_\varepsilon)} - Q_{G,\vartheta^*}\| &\leqslant \|Q_{G,\psi(b_\varepsilon)} - Q_{G,\theta_\varepsilon}\| + \|Q_{G,\theta_\varepsilon} - Q_{G,\vartheta^*}\| \\ &\leqslant \sqrt{2D(Q_{G,\theta_\varepsilon}\|Q_{G,\psi(b_\varepsilon)})} + \sqrt{2D(Q_{G,\theta_\varepsilon}\|Q_{G,\vartheta^*})} = o(\varepsilon^{(\gamma-1)/2}) + o(\varepsilon^{1/2})\,.\end{aligned}$$

Having $\gamma < 2$, $\|Q_{G,\psi(b_\varepsilon)} - Q_{G,\vartheta^*}\|$ is of the order $o(\varepsilon^{(\gamma-1)/2})$, and so are the other two total variations above. Since Q_{G,ϑ^*} equals $Q_{G,\psi_G(x^*_{ab})}$ by (ii), if $y \in \mathsf{s}(\mu) \cap G$ then

$$P_\varepsilon(y) = [\varepsilon + o(\varepsilon^\gamma)]\, Q_{G,\psi(b_\varepsilon)}(y) = \varepsilon Q_{G,\psi_G(x^*_{ab})}(y) + o(\varepsilon^\beta)$$

where $\gamma > \beta = \frac{\gamma+1}{2} > 1$.

3. *The case $y \in F$.* Knowing about the order of convergence in variation of $Q_{G,\psi(b_\varepsilon)}$ to $Q_{G,\psi_G(x^*_{ab})}$, the norm of $d_{G,\varepsilon} - x^*_{ab}$ is of the order $o(\varepsilon^{(\gamma-1)/2})$. The equations

$$(1-\varepsilon)a + \varepsilon x_{ab} = b_\varepsilon = [1 - \varepsilon + o(\varepsilon^\gamma)]\, d_{F,\varepsilon} + [\varepsilon + o(\varepsilon^\gamma)]\, d_{G,\varepsilon} + o(\varepsilon^\gamma)\, d_{B,\varepsilon}$$

imply

$$(1-\varepsilon)(d_{F,\varepsilon} - a) = \varepsilon(x_{ab} - d_{G,\varepsilon}) + o(\varepsilon^\gamma)d_{F,\varepsilon} + o(\varepsilon^\gamma)d_{G,\varepsilon} + o(\varepsilon^\gamma)d_{B,\varepsilon}\,.$$

Hence,

$$\|(1-\varepsilon)(d_{F,\varepsilon} - a) - \varepsilon(x_{ab} - x^*_{ab})\| \leqslant \varepsilon\|x^*_{ab} - d_{G,\varepsilon}\| + o(\varepsilon^\gamma) = o(\varepsilon^\gamma)\,.$$

Therefore, $\|d_{F,\varepsilon} - a\| = \frac{\varepsilon}{1-\varepsilon}\|x_{ab} - x^*_{ab}\| + o(\varepsilon^\gamma)$. In particular, $\|d_{F,\varepsilon} - a\|$ is of the order $O(\varepsilon)$.

Having $y \in \mathsf{s}(\mu) \cap F$, since $Q_{F,\psi(b_\varepsilon)}(y)$ coincides with $Q_{F,\psi_F(d_{F,\varepsilon})}(y)$

$$\ln \frac{Q_{F,\psi(b_\varepsilon)}(y)}{Q_{F,\psi_F(a)}(y)} = \langle \psi_F(d_{F,\varepsilon}) - \psi_F(a), y\rangle - \big[\Lambda_F(\psi_F(d_{F,\varepsilon})) - \Lambda_F(\psi_F(a))\big]\,.$$

Recall the known facts that $\psi_F\colon \mathit{ri}(F) \to \mathit{lin}(F)$ and Λ_F have Taylor expansions, and the directional derivative $\Lambda'_F(\psi_F(a);\vartheta)$ equals $\langle \vartheta, a\rangle$, $\vartheta \in \mathit{lin}(F)$. It follows that the above bracket equals

$$\langle \psi_F(d_{F,\varepsilon}) - \psi_F(a), a\rangle + O(\|\psi_F(d_{F,\varepsilon}) - \psi_F(a)\|^2)\,.$$

[†]The author apologizes to the readers of [10] where on p. 744, l–9 it is erroneously stated that $\|c_{F,\varepsilon} - a\|$ is of the order $o(\varepsilon)$.

The O-term is actually $O(\|d_{F,\varepsilon} - a\|^2)$, thus $O(\varepsilon^2)$. Hence

$$\ln \frac{Q_{F,\psi(b_\varepsilon)}(y)}{Q_{F,\psi_F(a)}(y)} = \langle \psi_F(d_{F,\varepsilon}) - \psi_F(a), y - a\rangle + O(\varepsilon^2)\,.$$

The scalar product expands to $\langle \psi'_F(a; d_{F,\varepsilon} - a), y - a\rangle + O(\|(d_{F,\varepsilon} - a)^2\|)$ where the O-term simplifies to $O(\varepsilon^2)$. Since $d_{F,\varepsilon} - a$ and $\frac{\varepsilon}{1-\varepsilon}(x_{ab} - x^*_{ab})$ are $o(\varepsilon^\gamma)$ apart in the norm, having $\gamma < 2$,

$$\ln \frac{Q_{F,\psi(b_\varepsilon)}(y)}{Q_{F,\psi_F(a)}(y)} = \tfrac{\varepsilon}{1-\varepsilon}\langle \psi'_F(a; x_{ab} - x^*_{ab}), y - a\rangle + o(\varepsilon^\gamma)\,.$$

Actually, the ratio $\frac{\varepsilon}{1-\varepsilon}$ can be replaced by ε. It follows that

$$\begin{aligned} P_\varepsilon(y) = Q_{\psi(b_\varepsilon)}(y) &= [1 - \varepsilon + o(\varepsilon^\gamma)]\, Q_{F,\psi(b_\varepsilon)}(y) \\ &= [1 - \varepsilon + o(\varepsilon^\gamma)]\, Q_{F,\psi_F(a)}(y)\, \exp\left[\varepsilon\langle \psi'_F(a; x_{ab} - x^*_{ab}), y - a\rangle + o(\varepsilon^\gamma)\right]. \end{aligned}$$

Expanding the exponential,

$$Q_{\psi(b_\varepsilon)}(y) = Q_{F,\psi_F(a)}(y) \cdot \left[1 - \varepsilon + \varepsilon\,\langle \psi'_F(a; x_{ab} - x^*_{ab}), y - a\rangle\right] + o(\varepsilon^\gamma)\,.$$

The assertion of theorem holds with $\alpha = \beta$ which is strictly between one and two. □

The following lemmas are consequences of known results [8], but no explicit reference seems to be available. Direct proofs are by calculus.

Lemma 3.2. For $c \in \mathit{ri}(\mu)$ and $\vartheta \in \mathit{lin}(\mu)$

$$V'(c; \vartheta) = \textstyle\sum_{y \in \mathit{s}(\mu)} U(y - c) \cdot \langle \psi'(c; \vartheta), y - c\rangle \cdot Q_{\psi(c)}(y)\,.$$

Lemma 3.3. If $c \in \mathit{ri}(\mu)$ and $\tau, \varsigma \in \mathbb{R}^d$ then $V(c)$ maps $(\psi'(c; \tau), \varsigma)$ to $\langle \varsigma, \tau\rangle$.

Additionally to the forms $U(z)$, let $W(y, z)$, $y, z \in \mathbb{R}^d$, denote the symmetric bilinear form $(\tau, \varsigma) \mapsto \langle \tau, y\rangle\langle \varsigma, z\rangle + \langle \tau, z\rangle\langle \varsigma, y\rangle$.

P r o o f . [Proof of Theorem 2.2] By definition,

$$V(b_\varepsilon) = \textstyle\sum_{y \in \mathit{s}(\mu)} U(y - b_\varepsilon) \cdot Q_{\psi(b_\varepsilon)}(y)\,.$$

The elementary form $U(y - b_\varepsilon) = U(y - a - \varepsilon(x_{ab} - a))$ rewrites to

$$U(y - a) - \varepsilon\, W(y - a, x_{ab} - a) + \varepsilon^2\, U(x_{ab} - a)\,.$$

This and Theorem 2.1 imply that $V(b_\varepsilon)$ differs in the norm from

$$\begin{aligned} &\textstyle\sum_{y \in \mathit{s}(\mu) \cap F} \left[U(y - a) - \varepsilon\, W(y - a, x_{ab} - a)\right] \\ &\qquad\qquad \cdot Q_{F,\psi_F(a)}(y) \cdot \left[1 - \varepsilon + \varepsilon\,\langle \psi'_F(a; x_{ab} - x^*_{ab}), y - a\rangle\right] \\ &+ \textstyle\sum_{y \in \mathit{s}(\mu) \cap G} U(y - a) \cdot \varepsilon \cdot Q_{G,\psi_G(x^*_{ab})}(y) \end{aligned}$$

by an $o(\varepsilon^\alpha)$ term, for $1 < \alpha < 2$. The first and second sums are referred to as (I) and (II).

The sum over $y \in \mathsf{s}(\mu) \cap F$ of $W(y-a, x_{ab}-a)$ weighted by $Q_{F,\psi_F(a)}(y)$ vanishes because $Q_{F,\psi_F(a)}$ has the mean a. Hence, (I) recasts to

$$(1-\varepsilon)V_F(a) + \varepsilon \sum\nolimits_{y\in\mathsf{s}(\mu)\cap F} U(y-a)\cdot\langle \psi'_F(a; x_{ab}-x^*_{ab}), y-a\rangle \cdot Q_{F,\psi_F(a)}(y)\,.$$

By Lemma 3.2 applied to the family $\mathcal{E}_F$, the sum above equals the directional derivative $V'_F(a; x_{ab}-x^*_{ab})$. In (II),

$$U(y-a) = U(y-x^*_{ab}) + W(y-x^*_{ab}, x^*_{ab}-a) + U(x^*_{ab}-a)\,.$$

The same argument as above implies that the sum over $y \in \mathsf{s}(\mu)\cap G$ of $W(y-x^*_{ab}, x^*_{ab}-a)$ weighted by $Q_{G,\psi_G(x^*_{ab})}(y)$ is zero. Therefore, $V(b_\varepsilon)$ and

$$(1-\varepsilon)V_F(a) + \varepsilon V'_F(a; x_{ab}-x^*_{ab}) + \varepsilon\, V_G(x^*_{ab}) + \varepsilon U(x^*_{ab}-a)$$

are $o(\varepsilon^\alpha)$ apart in the norm. □

P r o o f . [Proof of Corollary 2.3] It suffices to write

$$D(P\|Q_{\psi(b_\varepsilon)}) = \sum\nolimits_{y\in\mathsf{s}(P)} P(y)\left[\ln\frac{P(y)}{Q_{F,\psi_F(a)}(y)} - \ln\frac{Q_{\psi(b_\varepsilon)}(y)}{Q_{F,\psi(a)}(y)}\right]$$

and expand the logarithm on the right to, by Theorem 2.1,

$$\ln\frac{Q_{\psi(b_\varepsilon)}(y)}{Q_{F,\psi_F(a)}(y)} = -\varepsilon + \varepsilon\langle\psi'_F(a; x_{ab}-x^*_{ab}), y-a\rangle + o(\varepsilon^\alpha)\,, \qquad y\in\mathsf{s}(\mu)\cap F\,.$$

The mean of P emerges after summation. □

P r o o f . [Proof of Theorem 2.4] The divergence $D(Q_{\psi(b_\varepsilon)}\|Q_\tau)$, as a sum over $y\in\mathsf{s}(\mu)$, consists of the three sums according to whether y belongs to F, G or B. They are referred to as (I), (II) and (III), respectively. For the purposes of this proof the expression $\langle\psi'_F(a; x_{ab}-x^*_{ab}), y-a\rangle$ is abbreviated to $\rho(y)$.

The sum (I) is approximated by Theorem 2.1 as

$$\sum\nolimits_{y\in\mathsf{s}(\mu)\cap F} Q_{F,\psi_F(a)}(y)\big[1-\varepsilon+\varepsilon\,\rho(y)+o(\varepsilon^\alpha)\big]\ln\frac{Q_{F,\psi_F(a)}(y)}{Q_\tau(y)}\big[1-\varepsilon+\varepsilon\,\rho(y)+o(\varepsilon^\alpha)\big]\,.$$

Expanding the logarithm of the bracket, this recasts to

$$(1-\varepsilon)D(Q_{F,\psi_F(a)}\|Q_\tau) + \varepsilon\sum\nolimits_{y\in\mathsf{s}(\mu)\cap F} Q_{F,\psi_F(a)}(y)\Big[\rho(y)-1+\rho(y)\ln\frac{Q_{F,\psi_F(a)}(y)}{Q_\tau(y)}\Big]$$

up to an $o(\varepsilon^\alpha)$ term. The sum of $Q_{F,\psi_F(a)}(y)\cdot\rho(y)$ over these y equals zero because the mean of this pm is a. Since the logarithm of the ratio is equal to

$$\langle\psi_F(a)-\tau, y\rangle - \Lambda_F(\psi_F(a)) + \Lambda(\tau)$$

the above sum reduces to

$$-1 + \sum\nolimits_{y\in\mathsf{s}(\mu)\cap F} Q_{F,\psi_F(a)}(y)\cdot\rho(y)\cdot\langle\psi_F(a)-\tau, y\rangle\,.$$

In the scalar product, y can be replaced by $y-a$ whence the above expression rewrites to

$$-1+\left\langle \psi_F(a)-\tau, \sum\nolimits_{y\in \mathsf{s}(\mu)\cap F} Q_{F,\psi_F(a)}(y)\cdot\rho(y)\cdot(y-a)\right\rangle .$$

The scalar product of any $\varsigma\in\mathbb{R}^d$ with this sum is

$$\sum\nolimits_{y\in \mathsf{s}(\mu)\cap F} Q_{F,\psi_F(a)}(y)\cdot\rho(y)\cdot\langle\varsigma, y-a\rangle = {}^{\psi'_F(a;x_{ab}-x^*_{ab})}V^{\varsigma}_F(a)_{=}\langle\varsigma, x_{ab}-x^*_{ab}\rangle$$

on account of Lemma 3.3 applied to the variance function V_F. It follows that the approximation of (I) takes the form

$$(1-\varepsilon)D(Q_{F,\psi_F(a)}\|Q_\tau)+\varepsilon\big[-1+\langle\psi_F(a)-\tau, x_{ab}-x^*_{ab}\rangle\big]+o(\varepsilon^\alpha).$$

The sum (II) rewrites by Theorem 2.1 to

$$\sum\nolimits_{y\in \mathsf{s}(\mu)\cap G}\left[\varepsilon\cdot Q_{G,\psi_G(x^*_{ab})}(y)+o(\varepsilon^\alpha)\right]\ln\frac{Q_{G,\psi_G(x^*_{ab})}(y)}{Q_\tau(y)}\big[\varepsilon+o(\varepsilon^\alpha)\big]$$

and further to

$$\varepsilon D(Q_{G,\psi_G(x^*_{ab})}\|Q_\tau)+o(\varepsilon^\alpha)+[\varepsilon+o(\varepsilon^\alpha)]\ln[\varepsilon+o(\varepsilon^\alpha)].$$

Having $1<\beta<\alpha$, the sum (II) equals

$$\varepsilon D(Q_{G,\psi_G(x^*_{ab})}\|Q_\tau)+\varepsilon\ln\varepsilon+o(\varepsilon^\beta).$$

The sum (III) reduces by Theorem 2.1 to $o(\varepsilon^\alpha)\ln o(\varepsilon^\alpha)$, thus is of the order $o(\varepsilon^\beta)$.

Combining together the three approximations, $D(Q_{\psi(b_\varepsilon)}\|Q_\tau)$ is equal to

$$(1-\varepsilon)D(Q_{F,\psi_F(a)}\|Q_\tau)+\varepsilon D(Q_{G,\psi_G(x^*_{ab})}\|Q_\tau)-\varepsilon+\varepsilon\ln\varepsilon+\varepsilon\langle\psi_F(a)-\tau, x_{ab}-x^*_{ab}\rangle,$$

up to a term of the order $o(\varepsilon^\beta)$. The assertion of Theorem 2.4 follows using that $-\varepsilon$ equals $(1-\varepsilon)\ln(1-\varepsilon)+O(\varepsilon^2)$. □

Having Lemma 3.1 at disposal, [10, Theorem 3.1] can be strengthened as follows. Its proof is a refinement of the old one. The maximization involved is the same as in (i), see Section 2.

Theorem 3.4. There exists $1<\alpha<2$ such that

$$\Lambda^*(b_\varepsilon)=\Lambda^*(a)+h(\varepsilon)+\varepsilon\left[\Psi^*_{C,\Xi}(x_{ab})-\Lambda^*(a)\right]+o(\varepsilon^\alpha)$$

where $\Psi^*_{C,\Xi}(x_{ab})$ denotes the maximum of $\langle\vartheta, x_{ab}\rangle-\Lambda_G(\vartheta)$ over $\vartheta\in\Xi\triangleq\psi_F(a)+\mathit{lin}(F)^\perp$.

Proof. Since $b_\varepsilon\in \mathit{ri}(A)$ it is possible to write $\Lambda^*_A(b_\varepsilon)=\langle\vartheta_\varepsilon, b_\varepsilon\rangle-\Lambda_A(\vartheta_\varepsilon)$. It follows from [10, eq. (10)] and $Q_{A,\vartheta_\varepsilon}=(1-\varepsilon)Q_{F,\theta_\varepsilon}+\varepsilon Q_{G,\theta_\varepsilon}$ that

$$\begin{aligned}\Lambda_A(\vartheta_\varepsilon)&=\Lambda_F(\vartheta_\varepsilon)-\ln(1-\varepsilon)\\ \Lambda_A(\vartheta_\varepsilon)&=\Lambda_G(\vartheta_\varepsilon)-\ln\varepsilon.\end{aligned}$$

The previous and $b_\varepsilon = (1-\varepsilon)c_{F,\varepsilon} + \varepsilon c_{G,\varepsilon}$ imply

$$\begin{aligned}\Lambda_A^*(b_\varepsilon) &= (1-\varepsilon)\left[\,\langle \vartheta_\varepsilon, c_{F,\varepsilon}\rangle - \Lambda_F(\vartheta_\varepsilon) + \ln(1-\varepsilon)\,\right] \\ &\quad + \varepsilon\left[\,\langle \vartheta_\varepsilon, c_{G,\varepsilon}\rangle - \Lambda_G(\vartheta_\varepsilon) + \ln\varepsilon\,\right] \\ &= h(\varepsilon) + (1-\varepsilon)\,\Lambda_F^*(c_{F,\varepsilon}) + \varepsilon\,\Lambda_G^*(c_{G,\varepsilon})\,.\end{aligned}$$

From the proof of Theorem 2.1, it is known that $\|c_{F,\varepsilon} - a\| = O(\varepsilon)$ and

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

Therefore, $c_{F,\varepsilon} - a$ differs from $\varepsilon(x_{ab} - c_{G,\varepsilon})$ in the norm as $O(\varepsilon^2)$. By Taylor expansion,

$$\Lambda_F^*(c_{F,\varepsilon}) = \Lambda_F^*(a) + \langle \psi_F(a), c_{F,\varepsilon} - a\rangle + O(\|c_{F,\varepsilon} - a\|^2)$$

where the O-term is actually $O(\varepsilon^2)$. The scalar product differs from $\varepsilon\,\langle \psi_F(a), x_{ab} - c_{G,\varepsilon}\rangle$ by an $O(\varepsilon^2)$-term. Hence,

$$\Lambda_A^*(b_\varepsilon) = h(\varepsilon) + (1-\varepsilon)\Lambda_F^*(a) + \varepsilon\,(1-\varepsilon)\,\langle \psi_F(a), x_{ab} - c_{G,\varepsilon}\rangle + \varepsilon\Lambda_G^*(c_{G,\varepsilon}) + O(\varepsilon^2)\,.$$

In the proof of Theorem 2.1 it was found that $\|Q_{G,\theta_\varepsilon} - Q_{G,\psi_G(x^*_{ab})}\|$ is of the order $o(\varepsilon^{(\gamma-1)/2})$ for some $1 < \gamma < 2$. This implies $\|c_{G,\varepsilon} - x^*_{ab}\| = o(\varepsilon^{(\gamma-1)/2})$. By Taylor expansion, $\Lambda_G^*(c_{G,\varepsilon}) = \Lambda_G^*(x^*_{ab}) + O(\|c_{G,\varepsilon} - x^*_{ab}\|) = \Lambda_G^*(x^*_{ab}) + o(\varepsilon^{(\gamma-1)/2})$. It follows that

$$\Lambda_A^*(b_\varepsilon) = h(\varepsilon) + (1-\varepsilon)\Lambda_F^*(a) + \varepsilon\,\langle \psi_F(a), x_{ab} - c_{G,\varepsilon}\rangle + \varepsilon\Lambda_G^*(x^*_{ab}) + o(\varepsilon^\gamma)\,.$$

By [10, Lemma 6.8], the maximum $\Psi^*_{C,\Xi}(x_{ab})$ from (i) equals $\Lambda_G^*(x^*_{ab}) + \langle \psi_F(a), x_{ab} - x^*_{ab}\rangle$. Hence,

$$\Lambda_A^*(b_\varepsilon) = h(\varepsilon) + (1-\varepsilon)\Lambda_F^*(a) + \varepsilon\left[\,\Psi^*_{C,\Xi}(x_{ab}) + \langle \psi_F(a), x^*_{ab} - c_{G,\varepsilon}\rangle\,\right] + o(\varepsilon^\gamma)\,.$$

Here, the scalar product is of the order $o(\varepsilon^{(\gamma-1)/2})$ so that it can drop out. It remains to mention the basic fact $\Lambda^*(a) = \Lambda_F^*(a)$ and take $\alpha = \gamma$. □

4. EXAMPLE: MULTINOMIAL FAMILY

In this section, all results of the paper are illustrated within the multinomial family with the parameters $d \geqslant 1$ and $n \geqslant 1$. Given $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$, let $x_0 = n - (x_1 + \ldots + x_d)$.

The multinomial family can be based on the measure μ concentrated on the set of x's with $x_0, \ldots, x_d$ nonnegative and integer such that $\mu(x) = \frac{n!}{x_0!\,x_1!\,\ldots\,x_d!}$.

For $\vartheta = (\vartheta_1, \ldots, \vartheta_d) \in \mathbb{R}^d$

$$\Lambda(\vartheta) = n\,\ln\left(\,1 + e^{\vartheta_1} + \ldots + e^{\vartheta_d}\,\right) \quad \text{and} \quad \Lambda'(\vartheta) = \frac{n}{1 + e^{\vartheta_1} + \ldots + e^{\vartheta_d}}\left(\,e^{\vartheta_1}, \ldots, e^{\vartheta_d}\,\right)$$

so that for $z = (z_1, \ldots, z_d)$ inside the simplex $\mathsf{cs}(\mu)$, having all $z_0, \ldots, z_d$ positive,

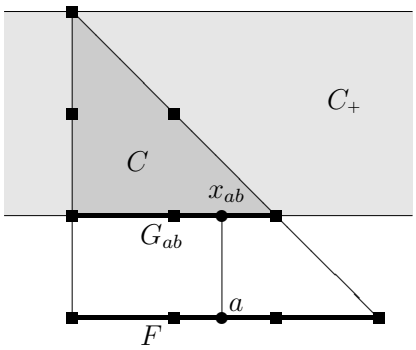
$$\psi(z) = \left(\,\ln\frac{z_1}{z_0}\,,\ \ldots\ ,\ \ln\frac{z_d}{z_0}\,\right).$$

The family consists of the familiar pm's parameterized by means z as

$$Q_{\psi(z)}(y) = \frac{n!}{y_0!\,\ldots\,y_d!}\ z_0^{y_0}\,\ldots\,z_d^{y_d}\,n^{-n}\,, \qquad y \in \mathsf{s}(\mu)\,.$$

The conjugate of Λ is

$$\Lambda^*(z) = \langle \psi(z), z\rangle - \Lambda(\psi(z)) = z_0 \ln z_0 + \ldots + z_d \ln z_d - n \ln n\,.$$



μ sits on ten points depicted as squares
F is the horizontal edge at the bottom
C is the shaded triangle
C_+ is a horizontal infinite strip
$x_{ab} = a + c$ is a point in C_+
G_{ab} is the horizontal edge in the middle

The set $F = \{x \in \mathsf{cs}(\mu)\colon x_d = 0\}$ is a facet of the simplex $\mathsf{cs}(\mu)$. Let a be any point inside F. Similarly as above

$$\psi_F(a) = \big(\, \ln \tfrac{a_1}{a_0}\,,\, \ldots\,,\, \ln \tfrac{a_{d-1}}{a_0}, 0\,\big)$$

and

$$Q_{F,\psi_F(a)}(y) = \tfrac{n!}{y_0! \ldots y_{d-1}!}\; a_0^{y_0} \ldots a_{d-1}^{y_{d-1}}\, n^{-n}\,, \qquad y \in \mathsf{s}(\mu) \cap F\,.$$

The point $b \in \mathsf{ri}(\mu)$ is chosen such that $b - a$ is a positive multiple of $c = (0, \ldots, 0, 1)$. Then $x_{ab} = a + c$ is the closest point of $C_+ = \{x \in \mathbb{R}^d\colon 1 \leqslant x_d \leqslant n\}$ to a in the direction $b - a$. The limiting along $b_\varepsilon = a + \varepsilon c$ to a is actually in the last coordinate. Then, $G = G_{ab}$ is the face $c + \frac{n-1}{n}\, F$ of $C = c + \frac{n-1}{n}\, \mathsf{cs}(\mu)$. In the framework of the exponential family $\mathcal{E}_G$,

$$\Lambda_G(\vartheta) = \ln n + \vartheta_d + (n-1)\, \ln\big(\, 1 + e^{\vartheta_1} + \ldots + e^{\vartheta_{d-1}}\,\big)$$

and

$$\Lambda'_G(\vartheta) = \tfrac{n-1}{1+e^{\vartheta_1}+\ldots+e^{\vartheta_{d-1}}}\, \big(\, e^{\vartheta_1}, \ldots, e^{\vartheta_{d-1}}, 0\,\big) + c\,.$$

Then

$$Q_{G,\vartheta^*}(y) = \tfrac{(n-1)!}{y_0! \ldots y_{d-1}!}\; a_0^{y_0} \ldots a_{d-1}^{y_{d-1}}\, n^{-n+1}\,, \qquad y \in \mathsf{s}(\mu) \cap G\,,$$

and

$$\psi_G(z_1, \ldots, z_{d-1}, 1) = \big(\, \ln \tfrac{z_1}{z_0}\,,\, \ldots\,,\, \ln \tfrac{z_{d-1}}{z_0}, 0\,\big)\,, \qquad (z_1, \ldots, z_{d-1}, 1) \in \mathsf{ri}(G)\,.$$

The maximization of $\langle \vartheta, x_{ab}\rangle - \Lambda_G(\vartheta)$ in (i) is over ϑ that differ from $\psi_F(a)$ only in the last coordinate

$$\max_{\vartheta_d \in \mathbb{R}}\, \langle \psi_F(a), a\rangle + \vartheta_d - \big[\, \ln n + \vartheta_d + (n-1) \ln \tfrac{n}{a_0}\big]\,.$$

Since it does not depend on ϑ_d a maximizer can be chosen as $\vartheta^* = \psi_F(a)$. Then the mean $x^*_{ab} = \mathsf{m}(Q_{G,\vartheta^*})$ is $\frac{n-1}{n}\, a + c$, $\psi_G(x^*_{ab}) = \vartheta^*$ and $x_{ab} - x^*_{ab} = \frac{1}{n}\, a$. It is needed in the sequel also that the maximum in (i) is

$$\Psi^*_{C,\Xi}(x_{ab}) = (a_0 - 1) \ln a_0 + a_1 \ln a_1 + \ldots + a_{d-1} \ln a_{d-1} - n \ln n$$

and

$$\psi'_F(a; x_{ab} - x^*_{ab}) = \tfrac{1}{a_0}(1, 1, \ldots, 1, 0)\,.$$

4.1. Directly from the formula for $Q_{\psi(b_\varepsilon)}$

$$Q_{\psi(b_\varepsilon)}(y) = \tfrac{n!}{y_0!\dots y_{d-1}!}\ (a_0-\varepsilon)^{y_0}\, a_1^{y_1}\ \dots\ a_{d-1}^{y_{d-1}}\, n^{-n}\,, \qquad y \in \mathsf{s}(\mu)\cap F\,,$$

and

$$Q_{\psi(b_\varepsilon)}(y) = \tfrac{n!}{y_0!\dots y_{d-1}!}\ (a_0-\varepsilon)^{y_0}\dots\ a_{d-1}^{y_{d-1}}\varepsilon\, n^{-n}\,, \qquad y \in \mathsf{s}(\mu)\cap G\,.$$

Hence, if $y \in \mathsf{s}(\mu)$ then

$$Q_{\psi(b_\varepsilon)}(y) = \begin{cases} Q_{F,\psi_F(a)}(y)\,(1-\frac{\varepsilon}{a_0})^{y_0}\,, & y\in F\,,\\ Q_{G,\psi_G(x^*_{ab})}(y)\,\varepsilon\,(1-\frac{\varepsilon}{a_0})^{y_0}\,, & y\in G\,,\\ O(\varepsilon^2)\,, & \text{otherwise.}\end{cases}$$

In the first case, $Q_{\psi(b_\varepsilon)}(y) = Q_{F,\psi_F(a)}(y)[1-\varepsilon\frac{y_0}{a_0}] + O(\varepsilon^2)$. Here, the coefficient $-\frac{y_0}{a_0}$ at ε is in accordance with $-1+\langle\psi'_F(a;x_{ab}-x^*_{ab}), y-a\rangle$, as claimed in Theorem 2.1. In the second case, analogously, $Q_{\psi(b_\varepsilon)}(y) = \varepsilon Q_{G,\psi_G(x^*_{ab})}(y) + O(\varepsilon^2)$.

4.2. The covariance $V(z)$, $z\in \mathsf{ri}(\mu)$, is viewed alternatively as the matrix $({}^iV^j(z))_{i,j=1}^d$ with the entries

$${}^iV^j(z) = \textstyle\sum_{y\in\mathsf{s}(\mu)}\ (y_i-z_i)(y_j-z_j)\ \cdot Q_{\psi(z)}(y)\,.$$

By [8, eq. (2.4)], the matrix equals $\Lambda''(\psi(z))$, which implies that ${}^iV^i(z) = z_i - z_i^2/n$ and ${}^iV^j(z) = -z_iz_j/n$ for $i\neq j$. Each entry is a quadratic polynomial in $z_1,\dots,z_d$. To summarize, $V(z)$ can be written as $D(z)-\frac{1}{n}U(z)$ where $D(z)$ is the form given by the diagonal matrix having z on the diagonal. The directional derivative $V'(z;y)$ of V at z in a direction y equals $D(y)-\frac{1}{n}W(y,z)$. In particular, $V'_F(a;x_{ab}-x^*_{ab}) = \frac{1}{n}D(a)-\frac{2}{n^2}U(a)$. The covariance $V_G(z) = D(z-c)-\frac{1}{n-1}U(z-c)$, $z\in\mathsf{ri}(G)$, can be computed analogously. In particular, $V_G(x^*_{ab}) = \frac{n-1}{n}D(a) - \frac{n-1}{n^2}U(a)$.

It follows that

$$\begin{aligned} V(b_\varepsilon) - V_F(a) &= \big[D(a+\varepsilon c) - \tfrac{1}{n}U(a+\varepsilon c)\big] - \big[D(a)-\tfrac{1}{n}U(a)\big]\\ &= \varepsilon\big[D(c)-\tfrac{1}{n}W(a,c)\big] - \varepsilon^2\tfrac{1}{n}U(c)\,.\end{aligned}$$

On the other hand, Theorem 2.2 implies that the ε-term is

$$\big[\tfrac{1}{n}D(a)-\tfrac{2}{n^2}U(a)\big] - \big[D(a)-\tfrac{1}{n}U(a)\big] + \big[\tfrac{n-1}{n}D(a)-\tfrac{n-1}{n^2}U(a)\big] + U(c-\tfrac{1}{n}a)$$

which is in accordance with $D(c)-\frac{1}{n}W(a,c)$.

4.3. If P is a pm with $\mathsf{s}(P)\subseteq\mathsf{s}(\mu)\cap F$ and the mean $\mathsf{m}(P)=z$ then

$$\begin{aligned} D(P\|Q_{\psi(b_\varepsilon)}) &= D(P\|Q_{F,\psi_F(a)}) - \textstyle\sum_{y\in\mathsf{s}(P)} P(y)\, y_0\, \ln\big(1-\frac{\varepsilon}{a_0}\big)\\ &= D(P\|Q_{F,\psi_F(a)}) + \varepsilon\tfrac{z_0}{a_0} + O(\varepsilon^2)\end{aligned}$$

where the ε-term corresponds to $1-\langle\psi'_F(a;x_{ab}-x^*_{ab}), z-a\rangle$ from Corollary 2.3.

4.4. For $\tau \in \mathbb{R}^d$, up to an $O(\varepsilon^2)$-term,

$$D(Q_{\psi(b_\varepsilon)}\|Q_\tau) = \sum_{y\in \mathsf{s}(\mu)\cap F} Q_{F,\psi_F(a)}(y)[1-\varepsilon\tfrac{y_0}{a_0}]\Big[\ln\,\frac{Q_{F,\psi_F(a)}(y)}{Q_\tau(y)} - \varepsilon\tfrac{y_0}{a_0}\Big]$$
$$+\,\varepsilon\sum_{y\in \mathsf{s}(\mu)\cap G} Q_{G,\psi_G(x^*_{ab})}(y)[1-\varepsilon\tfrac{y_0}{a_0}]\,\Big[\ln\varepsilon + \ln\,\frac{Q_{G,\psi_G(x^*_{ab})}(y)}{Q_\tau(y)} - \varepsilon\tfrac{y_0}{a_0}\Big]\,.$$

This recasts to

$$D(Q_{\psi(b_\varepsilon)}\|Q_\tau) = D(Q_{F,\psi_F(a)}\|Q_\tau) - \varepsilon - \tfrac{\varepsilon}{a_0}\sum_{y\in \mathsf{s}(\mu)\cap F} Q_{F,\psi_F(a)}(y)\,y_0\,\ln\,\frac{Q_{F,\psi_F(a)}(y)}{Q_\tau(y)}$$
$$+\varepsilon\ln\varepsilon + \varepsilon D(Q_{G,\psi_G(x^*_{ab})}\|Q_\tau) + O(\varepsilon^2\ln\varepsilon)\,.$$

To compute the above sum, it is suitable to write out the ratio under the logarithm as

$$a_0\ln\frac{a_0^n\,e^{\Lambda(\tau)}}{n^n} + \sum_{j=1}^{d-1}\,\ln\frac{a_i}{a_0\,e^{\tau_i}}\,\sum_{y\in \mathsf{s}(\mu)\cap F}\,y_0\,y_i\,Q_{F,\psi_F(a)}(y)$$

where the inner sum is $a_0\,a_i\frac{n-1}{n}$, resorting to $V_F(a)$. Hence,

$$D(Q_{\psi(b_\varepsilon)}\|Q_\tau) = (1-\varepsilon)D(Q_{F,\psi_F(a)}\|Q_\tau) + \varepsilon D(Q_{G,\psi_G(x^*_{ab})}\|Q_\tau)$$
$$+h(\varepsilon) + \varepsilon\sum_{j=1}^{d-1}\,\tfrac{1}{n}a_i\,\ln\frac{a_i}{a_0\,e^{\tau_i}} + O(\varepsilon^2\ln\varepsilon)\,.$$

where the sum equals $\langle\psi_F(a)-\tau, x_{ab}-x^*_{ab}\rangle$ in accordance with Theorem 2.4.

4.5. Directly from the formula for Λ^*

$$\Lambda^*(b_\varepsilon) = \Lambda^*(a) + h(\varepsilon) - \varepsilon\ln a_0 + O(\varepsilon^2)\,.$$

The coefficient $-\ln a_0$ at ε is in accordance with $\Psi^*_{C,\Xi}(x_{ab}) - \Lambda^*(a)$ as claimed in Theorem 3.4.

ACKNOWLEDGEMENT

This work was supported by Grant Agency of the Czech Republic under Grant 13-20012S.

(Received March 27, 2014)

REFERENCES

[1] N. Ay: An information-geometric approach to a theory of pragmatic structuring. The Annals of Probability *30* (2002), 416–436. DOI:10.1214/aop/1020107773

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

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

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

[5] I. Csiszár and F. Matúš: Closures of exponential families. The Annals of Probability *33* (2005), 582–600. DOI:10.1214/009117904000000766

[6] I. Csiszár and F. Matúš: Generalized maximum likelihood estimates for exponential families. Probability Theory and Related Fields *141* (2008), 213–246. DOI:10.1007/s00440-007-0084-z

[7] R. Graham, D. Knuth, and O. Patashnik: Concrete Mathematics. Second edition. Addison-Wesley, Reading, Massachusetts 1994, p. 446.

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

[9] F. Matúš and N. Ay: On maximization of the information divergence from an exponential family. In: Proc. WUPES'03 (J. Vejnarová, ed.), University of Economics, Prague 2003, pp. 99–204.

[10] F. Matúš: Optimality conditions for maximizers of the divergence from an EF. Kybernetika *43* (2007), 731–746.

[11] F. Matúš: Divergence from factorizable distributions and matroid representations by partitions. IEEE Trans. Inform. Theory *55* (2009), 5375–5381. DOI:10.1109/tit.2009.2032806

[12] F.Matúš F. and J. Rauh: Maximization of the information divergence from an exponential family and criticality. In: Proc. IEEE ISIT 2011, St. Petersburg 2011, pp. 809–813. DOI:10.1109/isit.2011.6034269

[13] G. Montúfar, J. Rauh J., and N. Ay: Maximal information divergence from statistical models defined by neural networks. In: Proc. GSI 2013, Paris 2013, Lecture Notes in Computer Science *8085* (2013), 759–766. DOI:10.1007/978-3-642-40020-9_85

[14] J. Rauh: Finding the maximizers of the information divergence from an exponential family. IEEE Trans. Inform. Theory *57* (2011), 3236–3247. DOI:10.1109/tit.2011.2136230

[15] R. T. Rockafellar: Convex Analysis. Princeton University Press, 1970. DOI:10.1017/s0013091500010142

František Matúš, Institute of Information Theory and Automation, The Czech Academy of Sciences, Pod Vodárenskou věží 4, 182 08 Praha 8. Czech Republic.
e-mail: matus@utia.cas.cz