Proceedings ISIT 2011, July 31 – August 5, St. Petersburg, Russia, 809–813

Maximization of the information divergence from an exponential family and criticality

František Matúš
Institute of Information Theory and Automation
Academy of Sciences of the Czech Republic
Pod vodárenskou věží 4
182 08 Prague, Czech Republic

Johannes Rauh
Max Planck Institute for Mathematics in the Sciences
Inselstraße 22
04103 Leipzig, Germany

***Abstract*—The problem to maximize the information divergence from an exponential family is compared to the maximization of an entropy-like quantity over the boundary of a polytope. First-order conditions on directional derivatives define critical sets for the two problems. The bijection between the sets of global maximizers in the two problems found earlier is extended here to bijections between the sets of local maximizers and the critical sets. This is based on new inequalities relating the maximized quantities and a reformulation of the first order criticality conditions for the second problem.**

I. Introduction

Let ν be a nonzero measure on a finite set Z with the support $s(\nu) \triangleq \{z \in Z \colon \nu(z) \neq 0\}$. The family of probability measures (pm's) P on Z with $s(P) \subseteq s(\nu)$ is denoted by $\mathcal{P}_{s(\nu)}$. The *information divergence* of a pm from ν is defined by

$$D(P\|\nu) \triangleq \sum_{z\in s(P)} P(z)\,\ln\frac{P(z)}{\nu(z)}\,, \quad P \in \mathcal{P}_{s(\nu)}\,.$$

The *exponential family* $\mathcal{E}_{\nu,f}$ determined by ν and a mapping $f \colon Z \to \mathbb{R}^d$ consists of the pm's $Q_{\nu,f,\vartheta}(z) = e^{\langle\vartheta,f(z)\rangle - t}\,\nu(z)$, $z \in Z$ where $\vartheta \in \mathbb{R}^d$ is a parameter, $\langle\cdot,\cdot\rangle$ is the scalar product and the normalizing constant t depends on ϑ [3], [4], [5], [9]. The information divergence of a pm $P \in \mathcal{P}_{s(\nu)}$ from the family $\mathcal{E}_{\nu,f}$ is defined as the infimum of $D(P\|Q)$ subject to $Q \in \mathcal{E}_{\nu,f}$. It is denoted by $D(P\|\mathcal{E}_{\nu,f})$. The problem to

(A) maximize $D(P\|\mathcal{E}_{\nu,f})$ subject to $P \in \mathcal{P}_{s(\nu)}$

was formulated in [1], investigating the infomax principle in probabilistic models of learning neural networks. For further progress see [13], [2], [10], [11], [12], [14].

A signed measure u on Z decomposes into $u^+ - u^-$ where u^+ and u^- are unique nonnegative measures with disjoint supports. Let $\mathcal{K}_{\nu,f}$ denote the linear space of u that satisfy

$$s(u) \subseteq s(\nu),\ \ \sum_{z\in s(\nu)} f(z)u(z) = 0 \ \text{ and } \ u(s(\nu)) = 0.$$

The set of $u \in \mathcal{K}_{\nu,f}$ decomposing into subprobabilities u^+ and u^- is a polytope. Its relative boundary is denoted by $\mathcal{U}_{\nu,f}$ and consists of $u \in \mathcal{K}_{\nu,f}$ decomposing into pm's u^+ and u^-. For $u \in \mathcal{U}_{\nu,f}$ let

$$\overline{D}(u\|\nu) \triangleq \sum_{z\in s(u)} u(z)\,\ln\frac{|u(z)|}{\nu(z)} = D(u^+\|\nu) - D(u^-\|\nu)$$

The problem to

(B) maximize $\overline{D}(u\|\nu)$ subject to $u \in \mathcal{U}_{\nu,f}$

was introduced in [14] and related to the problem (A) via the mapping

$$\Psi(P) = \frac{P - \pi_{\mathcal{E}}P}{(P-\pi_{\mathcal{E}}P)^+(s(\nu))}\,, \quad P \in \mathcal{P}_{s(\nu)} \setminus cl(\mathcal{E}_{\nu,f})\,.$$

Here, $\pi_{\mathcal{E}}P$ is the *generalized reversed information (rI-) projection* of P to $\mathcal{E}_{\nu,f}$ [6], and $cl(\mathcal{E}_{\nu,f})$ is the closure of $\mathcal{E}_{\nu,f}$ [7]. The mapping Ψ ranges in $\mathcal{U}_{\nu,f}$, see Remark 1. If $cl(\mathcal{E}_{\nu,f})$ differs from $\mathcal{P}_{s(\nu)}$ then Ψ restricts to a bijection between the classes of global maximizers in the problems (A) and (B), and $u \mapsto u^+$ provides its inverse [14, Theorem 4].

Section II recalls criticality conditions in the problems (A) and (B) and presents two theorems. Theorem 1 extends the bijection to the classes of local maximizers, critical and quasi-critical measures. Theorem 2 formulates two new inequalities relating $D(\cdot\|\mathcal{E}_{\nu,f})$ and $\overline{D}(\cdot\|\nu)$. Section III collects notations, known facts and auxiliary statements. Quasi-criticality and criticality are studied in Sections IV and V, respectively. Section VI contains proofs of the two theorems. An illustrating example is presented in Section VII.

II. Main results

Maximizers satisfy first order optimality conditions that are equivalent to the non-positivity of all directional derivatives. By [11, Theorem 5.1 and Remark 5.4], the one-sided directional derivatives of the function $D(\cdot\|\mathcal{E}_{\nu,f})$ are not positive at a pm $P \in \mathcal{P}_{s(\nu)}$ if and only if three conditions are fulfilled. Denoting by Π the generalized *rI*-projection $\pi_{\mathcal{E}}P$ of P to $\mathcal{E} = \mathcal{E}_{\nu,f}$, the first one requires

(A_1) $$P = \Pi^{\,s(P)}$$

where $\Pi^{\,s(P)}(z)$ equals $\Pi(z)/\Pi(s(P))$ for $z \in s(P)$ and 0 otherwise. If $s(\Pi) \neq s(\nu)$ the remaining conditions require

(A_2) $f(s(\Pi))$ and $f(s(\nu) \setminus s(\Pi))$ are contained in two different parallel hyperplanes, respectively,

and

(A_3) $$D(P\|\mathcal{E}) \geqslant \max_{\mathcal{P}_{s(\nu)\setminus s(\Pi)}} D(\cdot\|\mathcal{E}^{\Pi})$$

where $\mathcal{E}^{\Pi}$ consists of the pm's $R^{s(\nu)\setminus s(\Pi)}$ with $R \in \mathcal{E}$ and $R^{s(\Pi)} = \Pi$. The maximization in (A_3) is an instance of that in (A), see Lemma 2.

In the problem (A), a pm in $\mathcal{P}_{s(\nu)} \setminus cl(\mathcal{E}_{\nu,f})$ is *quasi-critical* if it satisfies (A_1) and (A_2), and *critical* if it satisfies also (A_3).

In the problem (B), the first order optimality conditions for signed measures $u \in \mathcal{U}_{\nu,f}$ require, by [14, Proposition 7],

$$(B_0)\quad \begin{aligned}&\textstyle\sum_{z\in s(u)} v(z)\ln\frac{|u(z)|}{\nu(z)} + \sum_{z\in s(v)\setminus s(u)} v(z)\ln\frac{|v(z)|}{\nu(z)}\\ &\leqslant v(s(u^+)\cup[s(v^+)\setminus s(u)])\,\overline{D}(u\|\nu)\,, \qquad v\in\mathcal{K}_{\nu,f}\,,\end{aligned}$$

and

$$(B_2)\qquad v(s(u)) = 0\,, \qquad v\in\mathcal{K}_{\nu,f}\,.$$

In particular, combining the inequalities (B_0) with v and $-v$,

$$(B_1)\quad \begin{aligned}\textstyle\sum_{z\in s(v)} v(z)\ln\frac{|u(z)|}{\nu(z)} &= v(s(u^+))\,\overline{D}(u\|\nu)\,,\\ &v\in\mathcal{K}_{\nu,f},\ s(v)\subseteq s(u)\,.\end{aligned}$$

In the special case $s(u) = s(\nu)$, the conditions (B_0) and (B_1) are equivalent while (B_2) holds by the definition of $\mathcal{K}_{\nu,f}$.

In the problem (B), $u \in \mathcal{U}_{\nu,f}$ is *quasi-critical* if it satisfies (B_1) and (B_2), and *critical* if it satisfies also (B_0).

The main result of this work states correspondences between classes of (quasi-) critical measures and local maximizers.

Theorem 1. *If $cl(\mathcal{E}_{\nu,f}) \neq \mathcal{P}_{s(\nu)}$ then the mappings $P \mapsto \Psi(P)$ on $\mathcal{P}_{s(\nu)} \setminus cl(\mathcal{E}_{\nu,f})$ and $u \mapsto u^+$ on $\mathcal{U}_{\nu,f}$ restrict to mutually inverse bijections between the classes of quasi-critical measures in the problems* (A) *and* (B). *They restrict further to bijections between the classes of critical measures and local maximizers, respectively.*

The assertion on the critical measures relies on a new characterization of the criticality in the problem (B) that parallels (A_1)–(A_3), see Theorem 3 in Section V. When proving the assertions on the maximizers the following two inequalities are crucial.

Theorem 2. *For $\mathcal{E} = \mathcal{E}_{\nu,f}$*

$$(1)\qquad D(P\|\mathcal{E}) \leqslant \ln[1+e^{\overline{D}(\Psi(P)\|\nu)}]\,, \quad P\in\mathcal{P}_{s(\nu)}\setminus cl(\mathcal{E})\,,$$

with the equality if and only if $\Psi(P)^+ = P$, and

$$(2)\qquad \overline{D}(u\|\nu) \leqslant \ln[e^{D(u^+\|\mathcal{E})} - 1]\,, \qquad u\in\mathcal{U}_{\nu,f}\,,$$

with the equality if and only if $\Psi(u^+) = u$.

Combining (1) and (2),

$$D(P\|\mathcal{E}) \leqslant D(\Psi(P)^+\|\mathcal{E})\,, \qquad P\in\mathcal{P}_{s(\nu)}\setminus cl(\mathcal{E})\,,$$

with the equality if and only if $\Psi(P)^+ = P$, and

$$\overline{D}(u\|\nu) \leqslant \overline{D}(\Psi(u^+)\|\nu)\,, \qquad u\in\mathcal{U}_{\nu,f}\,,$$

with the equality if and only if $\Psi(u^+) = u$. As a consequence,

$$(3)\quad \max\nolimits_{\mathcal{P}_{s(\nu)}} D(\cdot\|\mathcal{E}_{\nu,f}) = \ln\left[1+\exp\left(\max\nolimits_{\mathcal{U}_{\nu,f}}\overline{D}(\cdot\|\nu)\right)\right]\,,$$

and the correspondence [14, Theorem 4] between the classes of global maximizers follows.

III. Preliminaries

This section reviews known properties of the exponential families $\mathcal{E}_{\nu,f}$ and generalized *rI*-projections, see [6], [7], [8].

The closure $cl(\mathcal{E})$ of a family $\mathcal{E} = \mathcal{E}_{\nu,f}$ equals the union of the components $\{Q^{f^{-1}(F)} : Q\in\mathcal{E}\}$ over the nonempty faces F of the convex hull of $f(s(\nu))$. Such a component is the exponential family determined by $\nu^{f^{-1}(F)}$ and f.

If $P \in \mathcal{P}_{s(\nu)}$ then $cl(\mathcal{E})$ intersects $P + \mathcal{K}_{\nu,f}$ in a unique pm called the generalized *rI*-projection $\pi_{\mathcal{E}}P$ of P to $\mathcal{E}$ [6]. The mapping $P \mapsto \pi_{\mathcal{E}}P$ is continuous.

Remark 1. The *rI*-projection $\pi_{\mathcal{E}}P$ coincides with P if and only if P belongs to $cl(\mathcal{E})$. Thus, the mapping Ψ is well-defined on $\mathcal{P}_{s(\nu)} \setminus cl(\mathcal{E})$. It is continuous and ranges in $\mathcal{U}_{\nu,f}$ because $\Psi(P)$ is proportional to $P - \pi_{\mathcal{E}}P \in \mathcal{K}_{\nu,f}$. In the trivial case $cl(\mathcal{E}) = \mathcal{P}_{s(\nu)}$, excluded in Theorem 1, the function $D(\cdot\|\mathcal{E})$ vanishes identically on $\mathcal{P}_{s(\nu)}$ while $\mathcal{U}_{\nu,f} = \emptyset$.

Given a family $\mathcal{E} = \mathcal{E}_{\nu,f}$, the Pythagorean identity

$$(4)\qquad D(P\|\mu) = D(P\|\pi_{\mathcal{E}}P) + D(\pi_{\mathcal{E}}P\|\mu)$$

holds for all $P \in \mathcal{P}_{s(\nu)}$ and all measures μ such that $\mu/\mu(Z)$ belongs to $\mathcal{E}$. Hence, $D(P\|\mathcal{E}) = D(P\|\pi_{\mathcal{E}}P)$ is less than $D(P\|Q)$ for any $Q \in \mathcal{E}$ different from $\pi_{\mathcal{E}}P$.

Two pm's in $\mathcal{P}_{s(\nu)}$ have the same generalized *rI*-projection if and only if their difference belongs to $\mathcal{K}_{\nu,f}$. In particular, if $u = u^+ - u^- \in \mathcal{U}_{\nu,f}$ then $\pi_{\mathcal{E}}u^+ = \pi_{\mathcal{E}}u^-$. This implies that $cl(\mathcal{E})$ is disjoint with $\mathcal{U}^+_{\nu,f} = \{u^+ : u\in\mathcal{U}_{\nu,f}\}$ and $\Psi(u^+)$ in Theorem 2 is well-defined.

Remark 2. If a measure μ has the same support as ν and $g\colon Z\to\mathbb{R}^e$ then $\mathcal{E}_{\mu,g} \subseteq \mathcal{E}_{\nu,f}$ if and only if $\mu/\mu(Z) \in \mathcal{E}_{\nu,f}$ and there exists an affine mapping $A\colon \mathbb{R}^d \to \mathbb{R}^e$ such that $g = Af$ on $s(\nu)$. In this case, $\mathcal{K}_{\mu,g} \supseteq \mathcal{K}_{\nu,f}$. This implies that $\mathcal{K}_{\nu,f}$ and $\mathcal{U}_{\nu,f}$ depend on ν, f only through the exponential family $\mathcal{E}_{\nu,f}$. If $u \in \mathcal{U}_{\nu,f}$ then eqs. (4) with $P = u^\pm$ and $\pi_{\mathcal{E}}u^+ = \pi_{\mathcal{E}}u^-$ imply $\overline{D}(u\|\nu) = \overline{D}(u\|\mu)$. Thus, the function $\overline{D}(\cdot\|\nu)$ on $\mathcal{U}_{\nu,f}$ depends on ν only through the family $\mathcal{E}_{\nu,f}$.

Remark 3. If w is a signed measure with $s(w) \subseteq s(\nu)$ and $\sum_{z\in s(w)} w(z)v(z) = 0$ for $v \in \mathcal{K}_{\nu,f}$ with $s(v) \subseteq s(w)$ then $w = \langle\tau, f\rangle + r$ on $s(w)$ for some $\tau\in\mathbb{R}^d$ and $r\in\mathbb{R}$.

Lemma 1. *If pm's P and Q in $\mathcal{P}_{s(\nu)}$ have disjoint supports and $R_t = tP + (1-t)Q$, $0 \leqslant t \leqslant 1$, then $D(R_t\|\nu)$ is minimal if only if $t = [1+e^{\overline{D}(P-Q\|\nu)}]^{-1}$, in which case*

$$(5)\qquad \ln\tfrac{t}{1-t} = -\overline{D}(P-Q\|\nu)\,,$$

$$(6)\qquad D(R_t\|\nu) = -\ln[e^{-D(P\|\nu)} + e^{-D(Q\|\nu)}]\,,$$

$$(7)\qquad D(P\|R_t) = -\ln t\,.$$

Proof: Since P and Q have disjoint supports

$$D(R_t\|\nu) = D(Q\|\nu) + t\overline{D}(P-Q\|\nu) + t\ln t + (1-t)\ln(1-t)\,,$$

and the proof is completed by calculus. ∎

Let $lin_f(\nu)$ denote the linear subspace of $\mathbb{R}^d$ spanned by the differences $f(y) - f(z)$ for $y, z \in s(\nu)$.

Remark 4. Vectors $\vartheta, \theta \in \mathbb{R}^d$ parameterize the same pm in the family $\mathcal{E} = \mathcal{E}_{\nu,f}$, thus $Q_{\nu,f,\vartheta} = Q_{\nu,f,\theta}$, if and only if $\vartheta - \theta$ is orthogonal to $\mathit{lin}_f(\nu)$ [7].

Lemma 2. *If $\Pi \in \mathit{cl}(\mathcal{E}) \setminus \mathcal{E}$ then $\mathcal{E}^{\Pi}$ is the exponential family determined by any $R \in \mathcal{E}^{\Pi}$ and $\boldsymbol{\pi} f$ where $\boldsymbol{\pi}$ is the orthogonal projector on $\mathit{lin}_f(\Pi)^{\perp}$.*

Proof: By assumption, $\Pi = Q_{\nu^{s(\Pi)},f,\theta}$ for some $\theta \in \mathbb{R}^d$. Thus, the family $\mathcal{E}^{\Pi}$ consists of the pm's $Q_{\nu^{s(\nu)\setminus s(\Pi)},f,\vartheta}$ such that $\vartheta \in \mathbb{R}^d$ satisfies $Q_{\nu^{s(\Pi)},f,\vartheta} = Q_{\nu^{s(\Pi)},f,\theta}$. By Remark 4, this equality means that $\vartheta - \theta$ is orthogonal to $\mathit{lin}_f(\Pi)$. Therefore, denoting $Q_{\nu^{s(\nu)\setminus s(\Pi)},f,\theta}$ by R, the family $\mathcal{E}^{\Pi}$ consists of the pm's $Q_{R,\boldsymbol{\pi} f,\vartheta}$ with $\vartheta \in \mathbb{R}^d$. The assertion follows from Remark 2. ∎

IV. Quasi-criticality

This section studies the quasi-critical signed measures in the problems (A) and (B). The family $\mathcal{E}_{\nu,f}$ is denoted by $\mathcal{E}$.

Lemma 3. *If $P \in \mathcal{P}_{s(\nu)} \setminus \mathit{cl}(\mathcal{E})$ satisfies* (A_1) *then*
(i) $\Psi(P)^+ = P$,
(ii) $\pi_{\mathcal{E}} P = rP + (1-r)\Psi(P)^-$ for $r = e^{-D(P\|\mathcal{E})}$,
(iii) $D(P\|\mathcal{E}) = \ln[1 + e^{\overline{D}(\Psi(P)\|\nu)}]$,
and $\Psi(P)$ satisfies (B_1) *in the role of u.*

Proof: Let Π denote $\pi_{\mathcal{E}} P$. By the assumptions, $P \neq \Pi$ and $\Pi(z) = P(z)\Pi(s(P))$, $z \in s(P)$. Hence, $\Pi(s(P)) < 1$, and *(i)* follows from the definition of Ψ. In turn, Π equals $sP + (1-s)\Psi(P)^-$ where $s = \Pi(s(P))$. Since P and $\Psi(P)^-$ have disjoint supports $D(P\|\Pi) = -\ln s$. This and the known equality $D(P\|\Pi) = D(P\|\mathcal{E})$ imply *(ii)* with $r = s$.

Knowing that *(ii)* holds and $P - \Pi \in \mathcal{K}_{\nu,f}$ any pm in the segment between P and $Q = \Psi(P)^-$ has the generalized *rI*-projection to $\mathcal{E}$ equal to Π. By the Pythagorean identity (4), Π minimizes $D(\cdot\|\nu)$ over the segment. Lemma 1 implies that $r = [1 + e^{\overline{D}(P-Q\|\nu)}]^{-1}$. Then, *(iii)* follows from *(i)* and *(ii)*.

Let u denote $\Psi(P)$. If $v \in \mathcal{K}_{\nu,f}$ has $s(v) \subseteq s(u)$ then for δ sufficiently close to zero $P_\delta = \Pi + \delta v$ is a pm and Π is the generalized *rI*-projection of P_δ to $\mathcal{E}$. The Pythagorean equality

$$D(P_\delta\|\nu) = D(P_\delta\|\Pi) + D(\Pi\|\nu)$$

rewrites by *(i)*, *(ii)* and $v(s(u^+)) = -v(s(u^-))$ to

$$\begin{aligned} 0 &= \textstyle\sum_{z\in s(v)} [P_\delta(z) - \Pi(z)] \ln \frac{\Pi(z)}{\nu(z)} \\ &= \delta \textstyle\sum_{z\in s(v)} v(z) \ln \frac{ru^+(z)+(1-r)u^-(z)}{\nu(z)} \\ &= \delta \textstyle\sum_{z\in s(v)} v(z) \ln \frac{|u(z)|}{\nu(z)} + \delta v(s(u^+)) \ln \frac{r}{1-r}\,. \end{aligned}$$

Hence, (B_1) follows on account of (5). ∎

Remark 5. If $\mathcal{U}_{\nu,f} = \{u, -u\}$ then $\mathcal{P}_{s(\nu)} \setminus \mathit{cl}(\mathcal{E})$ is the disjoint union of $\Psi^{-1}(u)$ and $\Psi^{-1}(-u)$. These sets are nonempty and open in $\mathcal{P}_{s(\nu)} \setminus \mathit{cl}(\mathcal{E})$ because Ψ is continuous. Then, they are open in $\mathcal{P}_{s(\nu)}$ whence $\mathit{cl}(\mathcal{E}) \cup \Psi^{-1}(u)$ is compact. Maximizing $D(\cdot\|\mathcal{E})$ on this compact set, the maximum is uniquely attained at u^+. In fact, any global maximizer P of this problem must be in $\Psi^{-1}(u)$. Then, P is a local maximizer in (A). Since P satisfies (A_1) Lemma 3*(i)* implies $P = u^+$.

Lemma 4. *If $P \in \mathcal{P}_{s(\nu)} \setminus \mathit{cl}(\mathcal{E})$ is quasi-critical then* (B_2) *holds for $u = \Psi(P)$.*

Proof: The assertion is trivial unless $s(u)$ is strictly contained in $s(\nu)$. By Lemma 3*(ii)*, $\pi_{\mathcal{E}} P$ and u have the same support. It follows from (A_2) that there exist $\tau \in \mathbb{R}^d$ and $r_1 \neq r_2$ such that $\langle \tau, f \rangle$ equals r_1 on $s(u)$ and r_2 on $s(\nu) \setminus s(u)$. Since $v \in \mathcal{K}_{\nu,f}$ satisfies $\sum_{z\in s(\nu)} f(z)v(z) = 0$ and $v(s(\nu)) = 0$,

$$0 = \textstyle\sum_{z\in s(\nu)} \langle \tau, f(z) \rangle\, v(z) = r_1 \cdot v(s(u)) + r_2 \cdot v(s(\nu) \setminus s(u))$$

and $0 = v(s(u)) + v(s(\nu) \setminus s(u))$. These two equalities imply $v(s(u)) = 0$, thus (B_2) holds. ∎

Remark 6. If $u \in \mathcal{U}_{\nu,f}$ satisfies (B_2) and $s(u) \neq s(\nu)$ then f maps $s(u)$ and $s(\nu) \setminus s(u)$ into different parallel hyperplanes. In fact, (B_2) implies $\sum_{z\in s(w)} w(z)v(z) = 0$ for $v \in \mathcal{K}_{\nu,f}$ where w is the measure given by $w(z) = 1$, $z \in s(u)$, and $w(z) = 0$ otherwise. By Remark 3, there exist $\tau \in \mathbb{R}^d$ and real r such that $\langle \tau, f \rangle + r$ equals 1 on $s(u)$ and 0 on $s(\nu) \setminus s(u)$, which implies the assertion.

Lemma 5. *If $u \in \mathcal{U}_{\nu,f}$ is quasi-critical then*
(i) $\Psi(u^+) = u$,
(ii) $\pi_{\mathcal{E}} u^+ = tu^+ + (1-t)u^-$ for $t = [1 + e^{\overline{D}(u\|\nu)}]^{-1}$,
(iii) $D(u^+\|\mathcal{E}) = \ln[1 + e^{\overline{D}(u\|\nu)}]$,
and u^+ is quasi-critical.

Proof: Let Π denote $tu^+ + (1-t)u^-$ where t is given by *(ii)*. By Lemma 1, Π minimizes $D(\cdot\|\nu)$ over the segment with endpoints $P = u^+$ and $Q = u^-$. For $v \in \mathcal{K}_{\nu,f}$ with $s(v) \subseteq s(u) = s(\Pi)$

$$\begin{aligned} &\textstyle\sum_{z\in s(v)} v(z) \ln \frac{\Pi(z)}{\nu(z)} \\ &\qquad = \textstyle\sum_{z\in s(v)} v(z) \ln \frac{|u(z)|}{\nu(z)} + v(s(u^+)) \ln \frac{t}{1-t}\,. \end{aligned}$$

Here, $\ln \frac{t}{1-t} = -\overline{D}(u\|\nu)$ by (5), and thus the right-hand side vanishes because u satisfies (B_1) by assumption.

Let w be the signed measure given by $w(z) = \ln \frac{\Pi(z)}{\nu(z)}$ for $z \in s(u)$ and $w(z) = 0$ otherwise. Since $\sum_{z\in s(w)} w(z)v(z)$ vanishes for $v \in \mathcal{K}_{\nu,f}$ with $s(v) \subseteq s(w)$, using Remark 3,

$$w(z) = \ln \frac{\Pi(z)}{\nu(z)} = \langle \tau, f(z) \rangle + r\,, \qquad z \in s(\Pi)\,,$$

for some $\tau \in \mathbb{R}^d$ and $r \in \mathbb{R}$. If $s(\Pi) = s(\nu)$ then $\Pi \in \mathcal{E}$. Otherwise, since u satisfies (B_2) it follows from Remark 6 that f maps $s(\Pi)$ and $s(\nu) \setminus s(\Pi)$ into two different parallel hyperplanes. Therefore, $s(\Pi)$ equals $s(\nu) \cap f^{-1}(F)$ where F is the proper face of the convex hull of $f(s(\nu))$ exposed by one of the hyperplanes. In turn, $\Pi \in \mathit{cl}(\mathcal{E})$. Since Π belongs to the segment with the endpoints $u^{\pm}$ and $u \in \mathcal{U}_{\nu,f}$ it coincides with the generalized *rI*-projection of u^+ to $\mathcal{E}$, proving *(ii)*. This implies *(i)*, by the definition of Ψ. Since $D(u^+\|\mathcal{E})$ equals $D(u^+\|\Pi)$, *(iii)* follows from *(ii)*. Then, $P = u^+$ satisfies (A_1) by *(ii)* and (A_2) by Remark 6, thus it is quasi-critical. ∎

Since maximizers are quasi-critical Lemmas 3 and 5 imply the validity of (3) and the correspondence [14, Theorem 4] between the classes of global maximizers in the problems (A) and (B). This was concluded previously from Theorem 2.

V. CRITICALITY

This section presents a reformulation of the first order conditions in the problem (B). Let $\mathcal{E} = \mathcal{E}_{\nu,f}$.

Theorem 3. *A signed measure $u \in \mathcal{U}_{\nu,f}$ is critical in* (B) *if and only if it satisfies* (B_1), (B_2) *and, provided $\mathsf{s}(u) \neq \mathsf{s}(\nu)$,*

$$(B_3) \qquad \overline{D}(u\|\nu) \geqslant \max_{\mathcal{U}_{\lambda,\boldsymbol{\pi} f}} \overline{D}(\cdot\|\lambda)$$

where λ is any pm from $\mathcal{E}^{\pi_{\mathcal{E}} u^+}$ and $\boldsymbol{\pi}$ is the orthogonal projector on $\mathit{lin}_f(\pi_{\mathcal{E}} u^+)^\perp$.

Proof: In the case when $\mathsf{s}(u) = \mathsf{s}(\nu)$, ($B_0$) and ($B_1$) are equivalent while ($B_2$) holds by the definition of $\mathcal{K}_{\nu,f}$ and (B_3) is void. Otherwise, assuming (B_1) and (B_2) it suffices to prove the equivalence of (B_0) and (B_3).

Let Π denote $\pi_{\mathcal{E}} u^+$. By Lemma 5*(ii)*, $\Pi = tu^+ + (1-t)u^-$ where $t = [1 + e^{\overline{D}(u\|\nu)}]^{-1}$. Since the supports of Π and u coincide $\Pi = Q_{\nu^{\mathsf{s}(u)},f,\theta} \in \mathit{cl}(\mathcal{E}) \setminus \mathcal{E}$ for some $\theta \in \mathbb{R}^d$. Lemma 2 implies that $\mathcal{E}^{\pi_{\mathcal{E}} u^+}$ is determined by $R = Q_{\nu^{\mathsf{s}(\nu)\setminus \mathsf{s}(u)},f,\theta}$ and $\boldsymbol{\pi} f$. If $v \in \mathcal{K}_{\nu,f}$ then

$$\begin{aligned}\textstyle\sum_{z\in \mathsf{s}(u)} v(z) \ln \frac{|u(z)|}{\nu(z)} &= \textstyle\sum_{z\in \mathsf{s}(u)} v(z) \ln \frac{\Pi(z)}{\nu(z)} + v(\mathsf{s}(u^+)) \ln \frac{1-t}{t} \\ &= \textstyle\sum_{z\in \mathsf{s}(u)} v(z) \langle \theta, f(z)\rangle + v(\mathsf{s}(u^+))\, \overline{D}(u\|\nu)\end{aligned}$$

by $v(\mathsf{s}(u^+)) = -v(\mathsf{s}(u^-))$ and (5), and also

$$\begin{aligned}\textstyle\sum_{z\in \mathsf{s}(v)\setminus \mathsf{s}(u)} v(z) \ln \frac{|v(z)|}{\nu(z)} &= \textstyle\sum_{z\in \mathsf{s}(v)\setminus \mathsf{s}(u)} v(z) \left[\ln \frac{|v(z)|}{R(z)} + \langle \theta, f(z)\rangle \right].\end{aligned}$$

It follows that

$$\begin{aligned}\textstyle\sum_{z\in \mathsf{s}(u)} v(z) \ln \frac{|u(z)|}{\nu(z)} + \sum_{z\in \mathsf{s}(v)\setminus \mathsf{s}(u)} v(z) \ln \frac{|v(z)|}{\nu(z)} \\ = v(\mathsf{s}(u^+))\, \overline{D}(u\|\nu) + \textstyle\sum_{z\in \mathsf{s}(v)\setminus \mathsf{s}(u)} v(z) \ln \frac{|v(z)|}{R(z)}.\end{aligned}$$

Therefore, (B_0) is equivalent to the inequalities

$$\textstyle\sum_{z\in \mathsf{s}(v)\setminus \mathsf{s}(u)} v(z) \ln \frac{|v(z)|}{R(z)} \leqslant v(\mathsf{s}(v^+) \setminus \mathsf{s}(u))\, \overline{D}(u\|\nu)$$

with $v \in \mathcal{K}_{\nu,f}$. By Lemma 6 proved below, these inequalities rewrite to

$$\textstyle\sum_{z\in \mathsf{s}(w)} w(z) \ln \frac{|w(z)|}{R(z)} \leqslant w(\mathsf{s}(w^+))\, \overline{D}(u\|\nu)\,, \quad w \in \mathcal{K}_{R,\boldsymbol{\pi} f}\,,$$

which is equivalent to (B_3) with $\lambda = R$. It remains to recall that the maximization in (B_3) depends on λ only through the family $\mathcal{E}^{\pi_{\mathcal{E}} u^+}$, see Lemma 2 and Remark 2. ■

The maximization in (B_3) is an instance of that in (B).

The above proof of Theorem 3 refers to the following assertion. The notation is as above.

Lemma 6. *If $u \in \mathcal{U}_{\nu,f}$ satisfies* (B_2) *and R is a pm with support $\mathsf{s}(\nu)\setminus \mathsf{s}(u)$ then $\mathcal{K}_{R,\boldsymbol{\pi} f}$ consists of the signed measures w_v indexed by $v \in \mathcal{K}_{\nu,f}$ that agree with v on $\mathsf{s}(\nu) \setminus \mathsf{s}(u)$ and vanish outside $\mathsf{s}(u)$.*

Proof: By the assumption (B_2), if $v \in \mathcal{K}_{\nu,f}$ then $v(\mathsf{s}(u))$ and $v(\mathsf{s}(\nu) \setminus \mathsf{s}(u))$ vanish. Hence, $\sum_{z\in \mathsf{s}(u)} f(z)v(z)$ belongs to $\mathit{lin}_f(u)$ and $w_v(\mathsf{s}(R)) = 0$. Then $\sum_{z\in \mathsf{s}(\nu)} f(z)v(z) = 0$ implies $\sum_{z\in \mathsf{s}(R)} \boldsymbol{\pi} f(z) \cdot w_v(z) = 0$. Therefore, $w_v \in \mathcal{K}_{R,\boldsymbol{\pi} f}$.

In the opposite direction, if $w \in \mathcal{K}_{R,\boldsymbol{\pi} f}$ then, by definition, $\sum_{z\in \mathsf{s}(R)} \boldsymbol{\pi} f(z) \cdot w(z) = 0$ and $w(\mathsf{s}(R)) = 0$. Thus, the vector $\sum_{z\in \mathsf{s}(R)} f(z)w(z)$ belongs to $\mathit{lin}_f(u)$. Therefore, it can be expressed as $-\sum_{z\in \mathsf{s}(u)} f(z)v(z)$ with some coefficients $v(z)$ summing to zero. Let v be the signed measure on $\mathsf{s}(\nu)$ that coincides with w on $\mathsf{s}(R)$ and is defined on $\mathsf{s}(u)$ by means of the coefficients. Then $v \in \mathcal{K}_{\nu,f}$ and $w = w_v$. ■

VI. PROOFS OF THE MAIN RESULTS

This section presents the proofs of the main theorems formulated in Section II.

Proof of Theorem 2: The dependence of Ψ on $\mathcal{E}$ is made explicit here. For a pm P in $\mathcal{P}_{\mathsf{s}(\nu)} \setminus \mathit{cl}(\mathcal{E})$ let u denote $\Psi_{\mathcal{E}}(P)$. There exists a unique exponential family $\mathcal{F}$ determined by ν and some g such that $\mathcal{U}_{\nu,g} = \{u, -u\}$. Then, $\mathcal{E} \subseteq \mathcal{F}$ and $\pi_{\mathcal{E}} P \in \mathit{cl}(\mathcal{F})$. Hence, $\pi_{\mathcal{F}} P = \pi_{\mathcal{E}} P$, $\Psi_{\mathcal{F}}(P) = u$ and $D(P\|\mathcal{E}) = D(P\|\mathcal{F})$. This number is upper bounded by the maximum of $D(\cdot\|\mathcal{F})$ over $\mathit{cl}(\mathcal{F}) \cup \Psi_{\mathcal{F}}^{-1}(u)$. By Remark 5, the maximum equals $D(u^+\|\mathcal{F})$ and u^+ is the unique maximizer. In addition, u^+ is quasi-critical in (A) with $\mathcal{F}$ in the role of $\mathcal{E}$. By Lemma 3*(iii)*, $D(u^+\|\mathcal{F}) = \ln[1 + \exp(\overline{D}(u\|\nu))]$ whence (1) follows. That inequality is tight if and only if $P = u^+$.

For $u \in \mathcal{U}_{\nu,f}$ let Π denote $\pi_{\mathcal{E}} u^+ = \pi_{\mathcal{E}} u^-$. By Remark 2, $\overline{D}(u\|\nu) = \overline{D}(u\|\Pi)$. Using $D(u^+\|\mathcal{E}) = D(u^+\|\Pi)$, (2) is equivalent to

$$0 \leqslant D(u^+\|\Pi) - \ln[e^{\overline{D}(u\|\Pi)} + 1]\,.$$

Lemma 1 applies to $P = u^+$, $Q = u^-$ and Π in the role of ν. By (6), the right-hand side equals $D(R_t\|\Pi)$ where R_t minimizes $D(\cdot\|\Pi)$ over the segment between P and Q. Hence, (2) holds and is tight if and only if $R_t = \Pi$. The equality rewrites to $\Pi = tu^+ + (1-t)u^-$ where $t = [1 + e^{\overline{D}(u\|\Pi)}]^{-1}$. This is equivalent to $\Psi(u^+) = u$. ■

Proof of Theorem 1: By Lemmas 3 and 4, if P is quasi-critical then $\Psi(P)$ is quasi-critical and $\Psi(P)^+$ equals P. By Lemma 5, if u is quasi-critical then u^+ is quasi-critical and $\Psi(u^+)$ equals u. This proves the first assertion.

If P is critical in (A) then $u = \Psi(P)$ is quasi-critical by the first assertion, and thus critical when $\mathsf{s}(u) = \mathsf{s}(\nu)$. Otherwise, ($A_3$) with $\Pi = \pi_{\mathcal{E}} P$ and (3) with $\mathcal{E}^{\Pi}$ in the role of $\mathcal{E}_{\nu,f}$ combine to

$$D(P\|\mathcal{E}) \geqslant \ln\left[1 + \exp\left(\max_{\mathcal{U}_{\lambda,\boldsymbol{\pi} f}} \overline{D}(\cdot\|\lambda)\right)\right]$$

where $\lambda \in \mathcal{E}^{\Pi}$ and $\boldsymbol{\pi}$ is as in Lemma 2. This inequality and (1) from Theorem 2 entail (B_3) because $\boldsymbol{\pi}$ is at the same time the projector from Theorem 3, using that $\pi_{\mathcal{E}} u^+ = \pi_{\mathcal{E}} P = \Pi$ by Lemma 3*(i)*. Therefore, u is critical in (B) by Theorem 3. If u is critical in (B) then an analogous argumentation provides that u^+ is critical in (A). Details are omitted. This proves the second assertion.

If $u \in \mathcal{U}_{\nu,f}$ is a local maximizer in (B) then u is critical and $\overline{D}(\cdot\|\nu)$ is upper bounded by $\overline{D}(u\|\nu)$ in a neighborhood $\mathcal{V} \subseteq \mathcal{U}_{\nu,f}$ of u. By continuity, $\Psi^{-1}(\mathcal{V})$ is a neighborhood

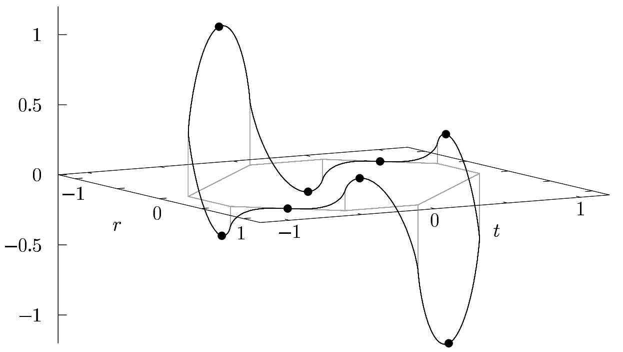

Fig. 1. The function $\overline{D}(\cdot\|\nu)$ from (B) for the binomial family with $n = 3$.

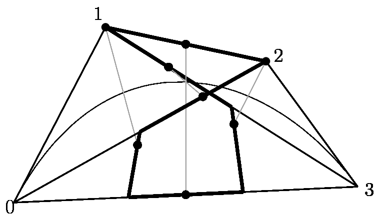

Fig. 2. The binomial family with $n = 3$ is a curve in a tetrahedron. The segments represent $\mathcal{U}^+_{\nu,f}$ and the dots the critical measures in (A).

of u^+, using also that $\Psi(u^+) = u$ by Lemma 5*(i)*. Applying (1) and (2), if $P \in \Psi^{-1}(\mathcal{V})$ then

$$\begin{aligned} D(P\|\mathcal{E}) &\leqslant \ln[1+\exp(\overline{D}(\Psi(P)\|\nu))] \\ &\leqslant \ln[1+\exp(\overline{D}(u\|\nu))] \leqslant D(u^+\|\mathcal{E})\,. \end{aligned}$$

Thus, u^+ is a local maximizer in (A). The opposite implication can be proved analogously, covering the third assertion on the classes of local maximizers. ∎

VII. EXAMPLE

By Theorem 1, if $cl(\mathcal{E}_{\nu,f})$ does not exhaust $\mathcal{P}_{s(\nu)}$ then the maximization in (A) can be restricted to the set $\mathcal{U}^+_{\nu,f}$ which contains all quasi-critical pm's and thus all maximizers in the problem (A). The surjection $u \mapsto u^+$ from $\mathcal{U}_{\nu,f}$ to $\mathcal{U}^+_{\nu,f}$ is not injective in general, as in the example below.

Binomial distributions on $Z = \{0, 1, \ldots, n\}$ form the family $\mathcal{E}_{\nu,f}$ where $\nu(z) = \binom{n}{z}$ and $f(z) = z$ for $z \in Z$. For this family the global maximizers in (A) were studied in [10].

Fig. 1 illustrates the problem (B) when $n = 3$. In this case, $\mathcal{K}_{\nu,f}$ corresponds to $\{(-2t-r, 3t, 3r, -t-2r)\colon t, r \in \mathbb{R}\}$ and $\mathcal{U}_{\nu,f}$ is the boundary of an octagon that is depicted in grey. The graph of $\overline{D}(\cdot\|\nu)$ on $\mathcal{U}_{\nu,f}$ is drawn in the vertical direction as a function of t, r and is linked to the vertices of the octagon. The problem (B) has a unique global maximizer and two additional local ones. The values of $\overline{D}(\cdot\|\nu)$ at the critical measures are marked as black dots.

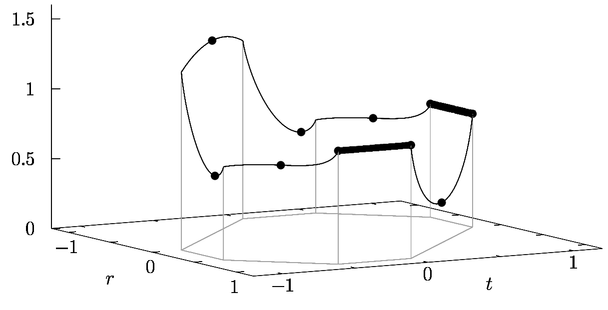

Fig. 3. The function $u \mapsto D(u^+\|\mathcal{E}_{\nu,f})$ on $\mathcal{U}_{\nu,f}$ for the binomial family with $n = 3$. Black dots and two highlighted segments depict critical values.

Fig. 2 illustrates the problem (A). The set $\mathcal{U}^+_{\nu,f}$ is a union of six highlighted segments because the mapping $u \mapsto u^+$ sends two edges of the octagon to two maximizers in (A), marked by 1 and 2. Since $D(\cdot\|\mathcal{E}_{\nu,f})$ is a function on the tetrahedron Fig. 3 presents instead the function $u \mapsto D(u^+\|\mathcal{E}_{\nu,f})$ on $\mathcal{U}_{\nu,f}$. It is constant on the two edges whence it has more maximizers.

ACKNOWLEDGEMENT

This work was supported by the Grant Agency of the Czech Republic under Grant P202/10/0618 and the Research Academy Leipzig.

REFERENCES

[1] N. Ay (2002) An information-geometric approach to a theory of pragmatic structuring. *Ann. Probab.* **30** 416–436.
[2] N. Ay and A. Knauf (2006) Maximizing multi-information. *Kybernetika* **45** 517–538.
[3] O. Barndorff-Nielsen (1978) *Information and Exponential Families in Statistical Theory.* Wiley, New York.
[4] L. D. Brown (1986) *Fundamentals of Statistical Exponential Families.* (Lecture Notes – Monograph Series 9.) Institute of Mathematical Statistics, Hayward, CA.
[5] N. N. Chentsov (1982) *Statistical Decision Rules and Optimal Inference.* (Translations of Mathematical Monographs) American Mathematical Society, Providence, R.I. (Russian original: Nauka, Moscow 1972.)
[6] I. Csiszár and F. Matúš (2003) Information projections revisited. *IEEE Trans. Inform. Theory* **49** 1474–1490.
[7] I. Csiszár and F. Matúš (2005) Closures of exponential families. *Ann. Probab.* **33** 582–600.
[8] I. Csiszár and F. Matúš (2008) Generalized maximum likelihood estimates for exponential families. *Probab. Theory Rel. Fields* **141** 213–246.
[9] G. Letac (1992) *Lectures on Natural Exponential Families and their Variance Functions.* (Monografias de Matemática 50.) Instituto de Matemática Pura e Aplicada, Rio de Janeiro.
[10] F. Matúš (2004) Maximization of information divergences from binary i.i.d. sequences. In: *Proc. IPMU 2004*, Perugia, Vol. 2, 1303–1306.
[11] F. Matúš (2007) Optimality conditions for maximizers of the information divergence from an exponential family. *Kybernetika* **43** 731–746.
[12] F. Matúš (2009) Divergence from factorizable distributions and matroid representations by partitions. *IEEE Trans. Inform. Theory* **55** 5375–5381.
[13] F. Matúš and N. Ay (2003) On maximization of the information divergence from an exponential family. In: *Proc. WUPES'03* (J. Vejnarová, ed.), University of Economics, Prague, 199–204.
[14] J. Rauh (2011) Finding the maximizers of the information divergence from an exponential family. (to appear in *IEEE Trans. Inform. Theory* **57**, No. 6, available at arXiv:0912.4660)