Proceedings of WUPES'03
Hejnice, Czech Republic, Sept. 24-27
University of Economics Prague, pp. 199–204.

February 27, 2004

On maximization of the information divergence from an exponential family

František Matúš and Nihat Ay†*

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 $\mathcal{E}$. For convex exponential families the local maximizers of this function of P are found. General exponential family $\mathcal{E}$ of dimension d is enlarged to an exponential family $\mathcal{E}^*$ of the dimension at most $3d+2$ such that the local maximizers are of zero divergence from $\mathcal{E}^*$.

1. Introduction

Let ν be a measure on a finite set Z, identified with the vector $(\nu(z))_{z\in Z}$ from $\mathbb{R}^Z$, such that the support of ν, $\mathit{supp}(\nu) = \{z \in Z\colon \nu(z) > 0\}$, equals Z. The information divergence $D(P\|\nu)$ (I-divergence, relative entropy, Kullback-Leibler divergence) of a probability measure (pm) P from ν is defined by the sum of $P(z)\,\ln[P(z)/\nu(z)]$ over z in the support of P.

For a vector u from $\mathbb{R}^Z$ let $Q_{\nu,u}$ be the pm proportional to $(\nu(z)\,e^{u(z)})_{z\in Z}$. Given a subspace H of $\mathbb{R}^Z$, the exponential family $\mathcal{E}_{\nu,H}$ based on ν and H is the set of all $Q_{\nu,u}$ with $u \in H$. It is assumed that the space H always contains the constant vector $\mathbb{1} = (1)_{z\in Z}$; this assumption does not restrict generality in the definition of $\mathcal{E}_{\nu,H}$ and reduces technicalities. Thus, the dimension of $\mathcal{E}_{\nu,H}$ is one less than the dimension of H.

The information divergence $D(P\|\mathcal{E}_{\nu,H})$ of a pm P from $\mathcal{E}_{\nu,H}$ is defined as infimum of the information divergences $D(P\|Q_{\nu,u})$ subject to $u \in H$. More general minimizations of this kind have been recently revisited in [4].

Interest in the local maximizers of the function $D(\cdot\,\|\mathcal{E}_{\nu,H})$ have emerged in probabilistic models of neural networks. These models, based on *infomax principles* for a variational characterization of adaptation and learning, see [6, 5, 8], involve optimization of the mutual information and related quantities. Such quantities often correspond to the very

This work was supported by Grant Agency of Academy of Sciences of the Czech Republic under the grant A1075104, by GA ČR under the grant 402/01/0981, and by the MPI MIS in Leipzig.

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

Key words and phrases. Kullback-Leibler divergence, information projection, exponential family, infomax principle.

*F. Matúš is with Institute of Information Theory and Automation, Academy of Sciences of the Czech Republic, Pod vodárenskou věží 4, 182 08 Prague, Czech Republic; matus@utia.cas.cz

†N. Ay is with Institute of Mathematics, Friedrich Alexander University Erlangen-Nürnberg, Bismarckstr. $1\frac{1}{2}$, D-91054 Erlangen, Germany; ay@mi.uni-erlangen.de

I-divergence of a pm from an exponential family. For example, the I-divergence of a pm P from an exponential family generated by first-order marginals is nothing but the mutual information (multi-information) in P. Previous works on this problem include [1, 2, 3].

In this contribution, the local and global maximizers of $D(\,\cdot\,\|\mathcal{E}_{\nu,H})$ are described for any exponential family which is a convex set of pm's, see Proposition 2 of Section 2. These exponential families are characterized in Proposition 1 below as sets of mixtures of singular pm's.

In the general case, by [1, Theorem 3.5], some of the local maximizers of $D(\,\cdot\,\|\mathcal{E})$ belong to the topological closure of an exponential family $\mathcal{E}^*$ that contains $\mathcal{E}$ and has dimension quadratic in the dimension of $\mathcal{E}$. This assertion is improved in Proposition 3 of Section 3 in two directions: our new enlargement $\mathcal{E}^*$ has dimension linear in the dimension of $\mathcal{E}$ and its closure contains all local maximizers of $D(\,\cdot\,\|\mathcal{E})$.

2. Convex exponential families

For a partition ϱ of Z and a block A of ϱ, let R_A be a pm on Z with the support equal to A. The set $\mathcal{F}_{\{R_A\,:\,A\in\varrho\}}$ of all pm's $\sum_{A\in\varrho} t_A\, R_A$ with $t_A > 0$ summing to one coincides with the exponential family based on the measure $\sum_{A\in\varrho} R_A$ and the space spanned by the vectors $\mathbb{1}_A$, $A \in \varrho$, where $\mathbb{1}_A(z)$ equals 1 if $z \in A$ and 0 otherwise. This exponential family is obviously convex.

Proposition 1. *Every convex exponential family $\mathcal{E}_{\nu,H}$ based on a measure ν with the support equal to Z coincides with $\mathcal{F}_{\{R_A\,:\,A\in\varrho\}}$ for a partition ϱ of Z and pm's R_A.*

A proof of Proposition 1 is based on the following lemma.

Lemma 1. *The smallest convex exponential family containing two pm's P and Q with the supports equal to Z coincides with $\mathcal{F}_{\{R_A\,:\,A\in\varrho_{P,Q}\}}$ where $\varrho_{P,Q}$ is the partition of Z having $x, y \in Z$ in the same block if and only if $P(x)Q(y) = P(y)Q(x)$ and R_A equals the conditioning $P(\cdot|A)$ of P to A.*

Proof. Let $\varrho_{P,Q}$ have n blocks and an element z_A of A be fixed for each $A \in \varrho_{P,Q}$. The numbers $P(z_A)^k Q(z_A)^{-k}$, $A \in \varrho_{P,Q}$, $0 \leqslant k < n$, are elements of a Vandermonde matrix which has nonzero determinant because $P(z_A)/Q(z_A)$, $A \in \varrho_{P,Q}$, are pairwise different. Therefore, for $0 \leqslant k < n$ the vectors $(P(z_A)^k Q(z_A)^{-k})_{A\in\varrho_{P,Q}}$ are linearly independent, and so are the vectors $(P(z)^k Q(z)^{-k})_{z\in Z}$. Then the pm's proportional to $(P(z)^{k+1} Q(z)^{-k})_{z\in Z}$ are independent. These pm's belong to any exponential family containing P and Q and, in turn, their convex hull is contained in any convex exponential family containing P and Q. In particular, it is contained in $\mathcal{F} = \mathcal{F}_{\{R_A\,:\,A\in\varrho_{P,Q}\}}$ because P and Q, equal to $\sum_{A\in\varrho} Q(A)\, R_A$, belong to $\mathcal{F}$ by construction. Since the convex hull has the same dimension as $\mathcal{F}$ any convex exponential family containing P and Q includes $\mathcal{F}$. $\square$

Proof of Proposition 1. Let ϱ be a partition of Z with the maximal number of blocks such that $\mathcal{E} = \mathcal{E}_{\nu,H}$ contains $\mathcal{F} = \mathcal{F}_{\{R_A\,:\,A\in\varrho\}}$ for some pm's R_A. For any pm P with the support equal to Z and $x \in A$, $y \in B$ belonging to different blocks A, B of ϱ, denote by $H_{P,x,y}$ the hyperplane of vectors $(t_C)_{C\in\varrho}$ satisfying

$$t_A \cdot P(x) R_A(y) - t_B \cdot P(y) R_B(x) = 0\,.$$

Since no such $H_{P,x,y}$ contains the hyperplane given by $\sum_{A\in\varrho} t_A = 1$ a pm $Q = \sum_{A\in\varrho} t_A\, R_A$ in $\mathcal{F}$ exists such that all equations $P(x)Q(y) = P(y)Q(x)$ with x, y in different blocks of ϱ are simultaneously violated. This implies that each block of ϱ is union of blocks of $\varrho_{P,Q}$. If, additionally, $P \in \mathcal{E}$ then $\mathcal{F}_{\{P(\cdot|A)\,:\,A\in\varrho_{P,Q}\}}$ is contained in $\mathcal{E}$ on account of Lemma 1. By maximality of the number of blocks, $\varrho_{P,Q} = \varrho$. Hence, $P = \sum_{A\in\varrho} P(A)\, P(\cdot|A)$ belongs to $\mathcal{F}$, and thus $\mathcal{E} = \mathcal{F}$. $\square$

When $Q = \sum_{A\in\varrho} t_A\, R_A$ belongs to $\mathcal{F} = \mathcal{F}_{\{R_A\colon A\in\varrho\}}$ then

$$D(P\|Q) \;=\; \sum_{A\in\varrho_P} P(A)\, D(P(\cdot|A)\|R_A) \;\;+\;\; P(A)\, \ln \frac{P(A)}{t_A}$$

where ϱ_P is the set of blocks from ϱ with $P(A) > 0$, and therefore

$$D(P\|\mathcal{F}) \;=\; \sum_{A\in\varrho_P} P(A)\, D(P(\cdot|A)\|R_A) \;=\; D(P\|\textstyle\sum_{A\in\varrho_P} P(A)\, R_A)\,. \tag{1}$$

By convexity of the information divergence in both coordinates, the function $D(\,\cdot\,\|\mathcal{F})$ is convex. Hence, the set of its local maximizers is union of simplices. They can be described explicitly as follows.

Proposition 2. *For a convex exponential family $\mathcal{F} = \mathcal{F}_{\{R_A\colon A\in\varrho\}}$, a pm P is a local maximizer of $D(\cdot\|\mathcal{F})$ if and only if $\mathsf{supp}(P)$ equals $\{z_A\colon A \in \varrho_P\}$ where each z_A is an element of A such that for some $p > 0$*

$$R_A(z_A) = p \text{ when } A \in \varrho_P, \quad \text{and} \quad R_B(y) \geqslant p \text{ when } y \in B \text{ and } B \in \varrho \setminus \varrho_P\,.$$

Proof. Using (1) and convexity of the information divergence, any local maximizer P can have $P(z)$ positive for a unique element, denoted by z_A, of A in ϱ_P. Then

$$D(P\|\mathcal{F}) = -\sum_{A\in\varrho_P} P(z_A)\, \ln R_A(z_A)$$

implies that $R_A(z_A) = p$ for some $p > 0$ and all $A \in \varrho_P$. For $y \in B$ and $B \in \varrho \setminus \varrho_P$

$$D(P\|\mathcal{F}) \geqslant D((1-\varepsilon)\, P + \varepsilon\, \mathbb{1}_{\{y\}}\|\mathcal{F})$$

rewrites to

$$-\ln p \geqslant -\sum_{A\in\varrho_P} (1-\varepsilon)\, P(z_A)\, \ln p \;-\; \varepsilon\, \ln R_B(y)$$

and the inequality $R_B(y) \geqslant p$ follows when ε is close to 0.

On the other hand, let P satisfy the condition of Proposition 2, and thus $D(P\|\mathcal{F})$ equals $-\ln p$. Since $P(\cdot|A) = \mathbb{1}_{\{z_A\}}$ and the information divergence is strictly convex in the first argument the inequalities

$$D(Q(\cdot|A)\|R_A) \;\leqslant\; D(P(\cdot|A)\|R_A) \;=\; -\ln p, \qquad A \in \varrho_P,$$

hold for each pm Q in a neighbourhood of P. On account of $\varrho_Q \supseteq \varrho_P$ and

$$D(Q(\cdot|A)\|R_B) \;\leqslant\; -\ln \min_{y\in B} R_B(y) \;\leqslant\; -\ln p, \qquad B \in \varrho_Q \setminus \varrho_P,$$

the identity (1) with P replaced by Q implies that $D(Q\|\mathcal{F})$ cannot exceed $-\ln p$. Thus, P is a local maximizer of $D(\cdot\|\mathcal{F})$. $\square$

Corollary 1. *A pm P is a global maximizer of $D(\cdot\|\mathcal{F})$ if and only if the support of P is contained in the set of minimizers of $\sum_{A\in\varrho} R_A(z)$ over $z\in Z$ and intersects each block of ϱ in at most one element. Furthermore, the following statements are equivalent:*

1. *There exists an isolated global maximizer of $D(\cdot\|\mathcal{F})$.*
2. *All global maximizers of $D(\cdot\|\mathcal{F})$ are isolated.*
3. *The set of minimizers is contained in a single block of ϱ.*

Example 1. Let $Z=\{1,\ldots,9\}$ be partitioned into $A_1=\{1,2,3\}$, $A_2=\{4,5,6,7\}$ and $A_3=\{8,9\}$, and R_{A_1} take the values $\frac{1}{6},\frac{1}{3},\frac{1}{2}$, R_{A_2} take the values $\frac{1}{6},\frac{1}{6},\frac{1}{3},\frac{1}{3}$, and R_{A_3} take the values $\frac{1}{3},\frac{2}{3}$ on elements of the blocks. By Corollary 1, a pm P is a global maximizer of $D(\cdot\|\mathcal{F})$ if and only if $\mathit{supp}(P)\subseteq\{1,4\}$ or $\mathit{supp}(P)\subseteq\{1,5\}$, and thus no isolated global maximizer of $D(\cdot\|\mathcal{F})$ exists. There are also local maximizers P that are not global, in the cases $\mathit{supp}(P)\subseteq\{2,6,8\}$ or $\mathit{supp}(P)\subseteq\{2,7,8\}$ with $p=\frac{1}{3}$.

3. Enlarging exponential families

The probability measures from an exponential family $\mathcal{E}_{\nu,H}$ can be written as

$$Q_{\nu,u}(z)=\nu(z)\,e^{u(z)-\Lambda_{\nu,H}(u)}\,,\qquad z\in Z$$

where

$$\Lambda_{\nu,H}(u)=\ln\sum_{z\in Z}\nu(z)\,e^{u(z)}\,,\qquad u\in H\,.$$

Then for any pm P

$$D(P\|Q_{\nu,u})=D(P\|\nu)-\langle u,P\rangle+\Lambda_{\nu,H}(u)\,,\qquad u\in H$$

where $\langle\cdot\,,\cdot\rangle$ is the scalar roduct on $\mathbb{R}^Z$,

$$D(P\|\mathcal{E}_{\nu,H})=D(P\|\nu)-\sup_{u\in H}\,\big[\langle u,P\rangle-\Lambda_{\nu,H}(u)\big]\,, \tag{2}$$

and thus $D(\,\cdot\,\|\mathcal{E}_{\nu,H})$ is difference of two convex functions.

Lemma 2. *If P is a local maximizer of $D(\,\cdot\,\|\mathcal{E}_{\nu,H})$ then the restriction of the coordinate projection of $\mathbb{R}^Z$ onto $\mathbb{R}^{\mathit{supp}(P)}$ to H is surjective.*

Proof. Let w be a vector in $\mathbb{R}^{\mathit{supp}(P)}$ orthogonal to the projection of H to $\mathbb{R}^{\mathit{supp}(P)}$. The vector $v\in\mathbb{R}^Z$ equal to w on $\mathit{supp}(P)$ and 0 otherwise is obviously orthogonal to H. Now, $P+tv$ is a pm for t close to 0. Using (2) for P and $P+tv$,

$$D(P\|\mathcal{E}_{\nu,H})-D(P+tv\|\mathcal{E}_{\nu,H})=D(P\|\nu)-D(P+tv\|\nu)\,.$$

Since P is a local maximizer of $D(\,\cdot\,\|\mathcal{E}_{\nu,H})$ this implies $D(P\|\nu)\geqslant D(P+tv\|\nu)$. By convexity of the information divergence and [7, Theorem 32.1], $D(P+tv\|\nu)$ is constant for t in a neighbourhood of 0. The strict convexity of the information divergence in the first coordinate implies $v=0$. Hence, $w=0$ and the assertion follows. ∎

Corollary 2. *Any local maximizer of $D(\,\cdot\,\|\mathcal{E}_{\nu,H})$ has cardinality of its support bounded from above by $d+1$, the dimension of H.*

The preceding assertion was proved in [1, Proposition 3.2] for the local maximizers P that have $D(P\|\mathcal{E}_{\nu,H})$ equal to $D(P\|Q)$ for some $Q\in\mathcal{E}_{\nu,H}$.

Corollary 3. *If P is a local maximizer of $D(\cdot\|\mathcal{E}_{\nu,H})$ then there exists $u \in H$ such that P equals the conditioning of $Q_{\nu,u}$ to $\mathsf{supp}(P)$.*

Proof. By Lemma 2, there exists $u \in H$ such that $u(z) = \ln[P(z)/\nu(z)]$ for $z \in \mathsf{supp}(P)$. Then, obviously, the pm $Q_{\nu,u}$ conditions to P given $\mathsf{supp}(P)$. $\square$

Proposition 3. *To an exponential family $\mathcal{E}_{\nu,H}$ based on a space $H \subseteq \mathbb{R}^Z$ of dimension $d+1$ there exists a subspace H^* of $\mathbb{R}^Z$ of dimension at most $3d+3$ such that $H \subseteq H^*$ and $D(P\|\mathcal{E}_{\nu,H^*}) = 0$ for every local maximizer P of $D(\cdot\|\mathcal{E}_{\nu,H})$.*

Proof. There is no loss of generality in assuming $Z = \{1, 2, \ldots, n\}$ for some $n \geqslant d+1$. Let H^* be the subspace of $\mathbb{R}^Z$ generated by H and the vectors $v^\ell = (1^\ell, 2^\ell, \ldots, n^\ell) \in \mathbb{R}^Z$, $0 \leqslant \ell \leqslant 2d+2$. Obviously, $H \subseteq H^*$ and, since $\mathbb{1} \in H$, the dimension of H^* is at most $3d+3$.

For a local maximizer P of $D(\cdot\|\mathcal{E}_{\nu,H})$, Corollary 2 implies that $P(y) > 0$ for y in a set Y of cardinality at most $d+1$. Expanding the nonnegative polynomial

$$g(t) = \prod_{y\in Y} (t-y)^2 = \sum_{\ell=0}^{2d+2} a_\ell\, t^\ell$$

one deduces that the vector $v = \sum_{\ell=0}^{2d+2} a_\ell\, v^\ell$ from H^* has all its coordinates nonnegative and $v(y) = 0$ if and only if $y \in Y$. By Corollary 3, there exists $u \in H$ such that $P = Q_{\nu,u}(\cdot|Y)$. Since also $P = Q_{\nu,u+tv}(\cdot|Y)$, a straightforward calculation gives

$$D(P\|Q_{\nu,u+tv}) = -\ln\, Q_{\nu,u+tv}(Y) = -\ln \frac{\sum_{y\in Y} \nu(y)\, e^{u(y)}}{\sum_{y\in Y} \nu(y)\, e^{u(y)} + \sum_{z\in Z\setminus Y} \nu(z)\, e^{u(z)+t\,v(z)}}\,.$$

For $Z \setminus Y$ nonempty, this information divergence converges to 0 with t decreasing to $-\infty$. This implies $D(P\|\mathcal{E}_{\nu,H^*}) = 0$. For $Z = Y$ the set Z has the cardinality $n = d+1$ and $\mathcal{E}_{\nu,H}$ consists of all positive pm's on Z. Obviously, $D(P\|\mathcal{E}_{\nu,H^*}) = 0$ because $P \in \mathcal{E}_{\nu,H}$. $\square$

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. (2003) Maximizing Multi-Information. (submitted)

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

[5] Bialek W., Rieke F., de Ruyter van Steveninck R., Warland D. (1991) Reading a neural code. *Science* **252** 1854–1857.

[6] Linsker R. (1988) Self-organization in a perceptual network. *IEEE Computer* **21** 105–117.

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

[8] Tononi G., Sporns O. and Edelman G.M. (1994) A measure for brain complexity: relating functional segregation and integration in the nervous system. *Proc. Natl. Acad. Sci. USA* **91**, 5033–5037.