Jozef Juríček

Maximization of the information divergence from multinomial distributions

Acta Universitatis Carolinae. Mathematica et Physica, Vol. 52 (2011), No. 1, 27--35

Persistent URL: http://dml.cz/dmlcz/143665

Maximization of the Information Divergence from Multinomial Distributions

JOZEF JURÍČEK

Praha

Received April 30, 2010
Revised June 1, 2010

The explicit solution of the problem of maximization of information divergence from the family of multinomial distributions is presented, using result of N. Ay and A. Knauf for the problem of maximization of multi-information [3], which is the special case of maximization of information divergence from hierarchical models [10].
The problem studied in this paper is a generalization of the binomial case, which was solved in [8].
The problem of maximization of information divergence from an exponential family has emerged in probabilistic models for evolution and learning in neural networks that are based on infomax principles [1].
The maximizers admit interpretation as stochastic systems with high complexity w.r.t. exponential family [3].

1. Introduction

Let ν be a nonzero measure on a finite set Z.

Let $\mathscr{F} = \mathscr{E}_{\nu,f} = \{Q_{\nu,f,\vartheta} : \vartheta \in \mathbb{R}^d\}$ be the (full) *exponential family* determined by the *reference measure* ν and the *directional statistic* $f : Z \to \mathbb{R}^d$, $d \in \mathbb{N}$, where $Q_{\nu,f,\vartheta}$ is a probability measure (pm) given by

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

MFF UK, KPMS, Sokolovská 83, 186 75 Praha 8 – Karlín

The work was supported by the grant SVV 261315/2010 and by the grant GAUK 54710/2010.
I would like to express my gratitude to my supervisor Ing. František Matúš, CSc., from Institute of Information Theory and Automation of the Academy of Sciences of the Czech Republic.
Special thanks go to the reviwers for their comments and constructive suggestions.

Key words and phrases. Information divergence, relative entropy, exponential family, information projection, hierarchical models, multi-information, multinomial distribution

E-mail address: jozef.juricek@matfyz.cz

$\langle \cdot, \cdot \rangle$ denotes the scalar product and

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

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

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

$\mathsf{s}(\cdot)$ is the *support* of a measure, i.e. $\mathsf{s}(\nu) = \{z \in Z : \nu(z) > 0\}$. The (information) divergence of a pm P from the family $\mathscr{F}$ is defined by

$$D(P\|\mathscr{F}) = \inf_{Q \in \mathscr{F}} D(P\|Q). \tag{1}$$

The general problem of maximization of information divergence from an arbitrary exponential family has emerged in probabilistic models for evolution and learning in neural networks based on infomax principles [1], [2]. Maximizers of $D(\cdot\|\mathscr{F})$ admit interpretation as stochastic systems with high complexity w.r.t. exponential family $\mathscr{F}$. This maximization problem goes back to [1], for later progress, see [7], [8], [9], [10]. Another equivalent optimization problem is stated in [11, Theorem 3].

From now on, assume $\mathsf{s}(\nu) = Z$, i.e. ν is strictly positive on Z. In this case, by [5, Theorem 1 (2)] and [9, Fact 2.10], the infimum in (1) is finite and uniquely attained in $\overline{\mathscr{F}}$, the (topological) closure of the exponential family $\mathscr{F} = \mathscr{E}_{\nu,f}$. The unique minimizing pm is called the *generalized rI-projection* of P and denoted by $P^{\mathscr{F}}$

$$D(P\|\mathscr{F}) = D(P\|\overline{\mathscr{F}}) = \min_{Q \in \overline{\mathscr{F}}} D(P\|Q) = D(P\|P^{\mathscr{F}}).$$

Function $P \mapsto D(P|\mathscr{F})$ is continuous [9, p. 4] on the simplex of all pm's on Z and therefore has a maximizer, see also [11, p. 7].

This work studies the global maximizers in the special case of multinomial family, hence a generalization of the binomial case [8, Proposition 2].

1.1 Formulation of the main result

Let $n, d \in \mathbb{N}$, $p_1, \ldots, p_d \geq 0$ s.t. $\sum_{j=1}^d p_j = 1$ and denote $[d] = \{1, \ldots, d\}$. The *multinomial distribution* with paramaters $n, d, p = (p_1, \ldots, p_d)$ is the frequency distribution of the i.i.d. sequence of n d-ary random variables taking values in $[d]$ with probabilities $p_1, \ldots, p_d$ respectively, i.e. is determined by the pm q_p on Z s.t.

$$q_p(z) = \binom{n}{z} \prod_{j=1}^d p_j^{z_j}, \; z \in Z = \left\{(z_1, \ldots, z_d) \in \{0, \ldots, n\}^d : \textstyle\sum_{j=1}^d z_j = n\right\}, \tag{2}$$

denoting by $\binom{n}{z} = \frac{n!}{\prod_{j=1}^d z_j!}$ the multinomial coefficient.

To simplify the notation let measures on an arbitrary finite set A be identified with the points in $\mathbb{R}^A$.

The *multinomial family* $\overline{\mathscr{M}}_{n,d} = \{q_p : p \text{ is a pm}\}$ is the closure of the exponential family $\mathscr{M}_{n,d} = \{q_p : p > 0 \text{ is a pm}\} = \mathscr{E}_{\nu,f} = \{Q_{\nu,f,\vartheta} : \vartheta \in \mathbb{R}^d\}$ with $\nu(z) = \binom{n}{z}$, $f(z) = z$.

Parametrization by p corresponds to the *mean value parametrization* (dp is the q_p mean) while $\vartheta \in \mathbb{R}^d$ is the *natural* (*canonical*) parameter [4] and $q_p = Q_{\nu,f,\vartheta}$ for $p_j = \frac{e^{\vartheta_j}}{\sum_{j' \in [d]} e^{\vartheta_{j'}}}$.

The main result of this paper is the explicit description of the set of maximizers of $D(\|\overline{\mathscr{M}}_{n,d})$ over all pm's on Z. This result is formulated in Theorem 1.1 and its proof is given in Section 3.

The remarkable fact of Theorem 1.1 is that the form of maximizers splits up into two cases, $n = 2$ and $n > 2$. For $n = 2$, the form is rather non-trivial.

Denote by $e^j = e^{j,d} = (0, \dots, 0, \underset{j}{1}, 0, \dots, 0)$ the j-th standard basis vector in $\mathbb{R}^d$, by δ_z the Dirac measure concentrated at the point $z \in Z$ and by S_A the symmetric group on an arbitrary finite set A; $\mathsf{S}_{[d]}$ shortens to S_d. To avoid the trivial cases let $n, d \geq 2$.

Theorem 1.1 (Maximizers of divergence from multinomial family)

(i) *The maximum value of* $D(\cdot\|\overline{\mathscr{M}}_{n,d})$ *is equal to* $(n-1)\ln d$.

(ii) *The maximizers project to the unique pm* μ *s.t.* $\mu(z) = \frac{\binom{n}{z}}{d^n}$, $z \in Z$.

(iii) *If* $n > 2$, *the unique maximizer is the pm uniformly concentrated on the set* $\{(n, 0, \dots, 0), \dots, (0, \dots, 0, n)\}$, *i.e. it has the form* $\frac{1}{d}\sum_{j=1}^{d} \delta_{ne^j}$.

(iv) *If* $n = 2$, *the maximizers have the form* $\frac{1}{d}\sum_{j=1}^{d} \delta_{e^j + e^{\pi(j)}}$, $\pi \in \mathsf{S}_d : \pi = \pi^{-1}$, *i.e.* π *is such a permutation on* $[d]$ *that it is a composition of independent cycles of lengths at most two, hence, only of transpositions (and identities).*

Remark 1.2 The rI-projection μ of any maximizer is proportional to ν, the reference measure of naturally parametrized multinomial family, i.e. $\mu \propto \nu$.

For π the identity on $[d]$, the two forms coincide.

For $n = 2$ and any fixed π s.t. $\pi = \pi^{-1}$, the maximizing pm sits on

$$(0, \dots, 0, \underset{j=\pi(j)}{2}, 0, \dots, 0), \qquad j \in [d] : j = \pi(j) \quad \text{with mass } \tfrac{1}{d}$$

and $$(0, \dots, 0, \underset{k}{1}, 0, \dots, 0, \underset{\pi(k)}{1}, 0, \dots, 0), \quad k \in [d] : k < \pi(k) \quad \text{with mass } \tfrac{2}{d}.$$

Let $\gamma : [d]^n \to Z$ be a transformation of outcomes of the original i.i.d. sequence to frequencies, i.e.

$$\gamma : (x_1, \dots, x_n) \mapsto (z_1, \dots, z_d), \quad z_j = |\{i \in [n] : x_i = j\}|, \ j \in [d]. \tag{3}$$

Remark 1.3 In statistical terms, γ defined by (3) is the *sufficient statistics* for the parameter p in the model $\overline{\mathscr{M}}_{n,d}$.

It is going to be shown, that the maximizers of $D(\cdot\|\overline{\mathscr{M}}_{n,d})$ over all pm's on Z, described in Theorem 1.1, are the γ-images of the subset of maximizers of $D(\cdot\|\mathscr{F})$, the divergence from the family $\mathscr{F}$ of all positive product pm's over all pm's on $[d]^n$, described in [3, Corollary 4.10] and stated here as Theorem 2.7. Moreover, projection of maximizers of $D(\cdot\|\overline{\mathscr{M}}_{n,d})$ is the γ-image of the pm uniformly concentrated on $[d]^n$, the projection of maximizers of $D(\cdot\|\mathscr{F})$ and the maximal value of $D(\cdot\|\overline{\mathscr{M}}_{n,d})$ is equal to the maximal value of $D(\cdot\|\mathscr{F})$.

2. Preliminaries

The previous idea of transformation of the problem by sufficient statistics is introduced in a more general setting.

Denote by $\overline{\mathscr{P}}(X)$ the simplex of all pm's on a finite set X and by $\mathscr{P}(X)$ its interior, i.e. the family of all strictly positive pm's on X.

Let Z be a nonempty finite set and γ be a surjection $\gamma : X \overset{\text{onto}}{\to} Z$. For $P \in \overline{\mathscr{P}}(X)$ let $\gamma(P) \in \overline{\mathscr{P}}(Z)$ be its γ-image, i.e. $\gamma(P)(z) = P(\gamma^{-1}(z))$, $z \in Z$.

For arbitrary families $\mathscr{E} \subseteq \overline{\mathscr{P}}(X)$ and $\mathscr{G} \subseteq \overline{\mathscr{P}}(Z)$ let

$$\gamma(\mathscr{E}) = \{\gamma(P) : P \in \mathscr{E}\} \text{ and } \gamma^{-1}(\mathscr{G}) = \{P \in \overline{\mathscr{P}}(X) : \ \gamma(P) \in \mathscr{G}\}.$$

2.1 Exchangeable and symmetrical families

Recall that S_X denotes the symmetric group (of all permutations) on X. For a permutation $\pi \in \mathsf{S}_X$ and a pm $P \in \overline{\mathscr{P}}(X)$ denote by πP the π^{-1}-image of P, i.e. $\pi P(x) = P(\pi(x))$, $x \in X$.

Definition 2.1 Let $\mathsf{G} \leq \mathsf{S}_X$ be a subgroup of S_X. A *family* $\mathscr{F} \subseteq \overline{\mathscr{P}}(X)$ is said to be G-*symmetrical* iff $\forall P \in \mathscr{F}\ \forall \pi \in \mathsf{G} : \ \pi P \in \mathscr{F}$.

Remarks 2.2

(a) The closure of a G-symmetrical family is again G-symmetrical.

(b) The family of maximizers of divergence from a G-symmetrical exponential family over G-symmetrical family is again G-symmetrical, as well as the set of projections of maximizers.

Definition 2.3 Let $\mathsf{G} \leq \mathsf{S}_X$. A *pm* $P \in \overline{\mathscr{P}}(X)$ is G-*exchangeable* iff

$$\forall \pi \in \mathsf{G} : \ \pi P = P.$$

Let $\overline{\mathscr{E}} = \overline{\mathscr{E}}_{\mathsf{G}}(X)$ be the family of all G-exchangeable pm's on X.

Remarks 2.4

(a) For $x \in X$ let $\mathsf{G}x = \{\pi(x) : \pi \in \mathsf{G}\}$ be the *x-orbit* of G and denote by $\mathsf{G}X = \{\mathsf{G}x : x \in X\}$ the set of all orbits (the *coinvariants*). As G is a group, $\mathsf{G}X$ forms a partition of X.

(b) $\overline{\mathscr{E}}$ is the closure of the exponential family $\mathscr{E} = \mathscr{E}_{\nu,f}$ with $f = (f_g)_{g\in \mathsf{G}X}$ s.t. $f_g(x) = \mathsf{I}(x \in g)$ and $\nu(x) = 1$, where $\mathsf{I}(\cdot)$ is the identificator

(c) P is G-exchangeable on X iff it is constant on every orbit $\mathsf{G}x$, $x \in X$.

Theorem 2.5 *Let* $\mathsf{G} \le \mathsf{S}_X$, $\mathscr{F} \subset \overline{\mathscr{P}}(X)$ *be* G*-symmetrical exponential family and* $\overline{\mathscr{E}} = \overline{\mathscr{E}}_{\mathsf{G}}(X)$ *be the the family of all* G*-exchangeable pm's on* X.

For any G*-exchangeable pm* $P \in \overline{\mathscr{E}}:\ P^{\mathscr{F}} = P^{\overline{\mathscr{E}}\cap\mathscr{F}}$.

Proof. It is assumed by contradiction, that $P^{\mathscr{F}} \in \overline{\mathscr{F}} \setminus \overline{\mathscr{E}}$. Thus, there exists $\pi \in \mathsf{G}$ s.t. $\pi P^{\mathscr{F}} \not\equiv P^{\mathscr{F}}$ and $\pi P^{\mathscr{F}} \in \overline{\mathscr{F}}$ by symmetry.

By exchangeability of $P \in \overline{\mathscr{E}}:\ D(P\|P^{\mathscr{F}}) = D(\pi P\|\pi P^{\mathscr{F}}) = D(P\|\pi P^{\mathscr{F}})$ and this is in contradiction with the uniqueness of the rI-projection $P^{\mathscr{F}}$.

In conclusion, the symbol $P^{\overline{\mathscr{E}}\cap\mathscr{F}}$ is well defined because $\overline{\mathscr{E}} \cap \mathscr{F} = \mathscr{E} \cap \mathscr{F}$ is intersection of two exponential families and it is nonempty due to the fact that $P^{\mathscr{F}} \in \in \overline{\mathscr{E} \cap \mathscr{F}}$, thus $\mathscr{E} \cap \mathscr{F}$ is again an exponential family. □

Corollary 2.6 *Let* $\mathsf{G} \le \mathsf{S}_X$, $\mathscr{F} \subset \overline{\mathscr{P}}(X)$ *be* G*-symmetrical exponential family and* $\overline{\mathscr{E}} = \overline{\mathscr{E}}_{\mathsf{G}}(X)$ *be the the family of all* G*-exchangeable pm's on* X.

Define $\gamma_{\mathsf{G}} : X \to \mathsf{G}X$ *s.t.* $\gamma_{\mathsf{G}}:\ x \mapsto \mathsf{G}x,\ x \in X$.

The following statements are equivalent

(i) *P is a maximizer of* $D(\cdot\|\overline{\mathscr{F}})$ *over* $\overline{\mathscr{E}}$*;*

(ii) *P is a maximizer of* $D(\cdot\|\overline{\mathscr{E}} \cap \mathscr{F})$ *over* $\overline{\mathscr{E}}$*;*

(iii) $\gamma_{\mathsf{G}}(P)$ *is a maximizer of* $D(\,\cdot\,\|\gamma_{\mathsf{G}}(\overline{\mathscr{E}} \cap \mathscr{F}))$ *over* $\overline{\mathscr{P}}(\mathsf{G}X)$*,*
$\gamma_{\mathsf{G}}(P)$ *projects to* $\gamma_{\mathsf{G}}(P^{\mathscr{F}})$ *and* $D(\gamma_{\mathsf{G}}(P)\|\gamma_{\mathsf{G}}(P^{\mathscr{F}})) = D(P\|P^{\mathscr{F}})$.

If there exists a maximizer of $D(\cdot\|\mathscr{F})$ *over* $\overline{\mathscr{P}}(X)$ *that is* G*-exchangeable, i.e.* $\overline{\mathscr{E}} \cap$ $\cap \arg\max_{\overline{\mathscr{P}}(X)} D(\cdot\|\mathscr{F}) \neq \emptyset$*, then*

$$\begin{aligned} &\arg\max_{\overline{\mathscr{P}}(\mathsf{G}X)} D(\,\cdot\,\|\gamma_{\mathsf{G}}(\overline{\mathscr{E}} \cap \mathscr{F})) &&= \gamma_{\mathsf{G}}(\overline{\mathscr{E}} \cap \arg\max_{\overline{\mathscr{P}}(X)} D(\cdot\|\mathscr{F})) \\ \textit{and}\qquad &\max_{\overline{\mathscr{P}}(\mathsf{G}X)} D(\,\cdot\,\|\gamma_{\mathsf{G}}(\overline{\mathscr{E}} \cap \mathscr{F})) &&= \max_{\overline{\mathscr{P}}(X)} D(\cdot\|\mathscr{F}). \end{aligned}$$

Proof. Propositions (i) and (ii) are equivalent by Theorem 2.5.

The remaining part follows from the following facts:

(a) γ_{G} maps measures of $\overline{\mathscr{E}}$ bijectively onto $\overline{\mathscr{P}}_{\mathsf{G}}$(and $\overline{\mathscr{E}} \cap \mathscr{F}$ onto $\gamma(\overline{\mathscr{E}} \cap \mathscr{F})$):

$$Q \in \overline{\mathscr{P}}_{\mathsf{G}}:\ \gamma_{\mathsf{G}}^{-1}(Q) \cap \overline{\mathscr{E}} = P,\ \text{s.t. } P(x) = \frac{Q(\mathsf{G}x)}{|\mathsf{G}x|}, x \in X;$$

(b) γ_{G} preserves the divergence on $\overline{\mathscr{E}}$, i.e. $\forall\, P, R \in \overline{\mathscr{E}}$ s.t. $D(P\|R) < \infty$:

$$\begin{aligned} D(P\|R) &= \sum_{x\in\mathfrak{s}(P)} P(x)\ln\frac{P(x)}{R(x)} = \sum_{g\in\mathsf{G}X}\sum_{x\in\gamma_{\mathsf{G}}^{-1}(g)\cap\mathfrak{s}(P)} P(x)\ln\frac{P(x)}{R(x)} = \\ &\overset{(*)}{=} \sum_{g\in\mathfrak{s}\left(\gamma_{\mathsf{G}}(P)\right)} \gamma_{\mathsf{G}}(P)(g) \sum_{x\in\gamma_{\mathsf{G}}^{-1}(g)\cap\mathfrak{s}(P)} \ln\frac{P(x)}{R(x)} = \\ &\overset{(**)}{=} \sum_{g\in\mathfrak{s}\left(\gamma_{\mathsf{G}}(P)\right)} \gamma_{\mathsf{G}}(P)(g)\ln\frac{\gamma_{\mathsf{G}}(P)(g)}{\gamma_{\mathsf{G}}(R)(g)} = D(\gamma_{\mathsf{G}}(P)\|\gamma_{\mathsf{G}}(R)). \end{aligned}$$

Equalities $(*)$ and $(**)$ follow from the fact that P and R are constant on each orbit of G by Remark 2.4(c). □

2.2 Maximization of the multi-information

Let $\mathscr{F}$ be the family of all positive product pm's on $X = [d]^n$, i.e.

$$\mathscr{F} = \{Q = Q_1 \times \ldots \times Q_n : Q_i \text{ is a strictly positive pm on } [d], i \in [n]\}. \tag{4}$$

It is easy to see, that $\mathscr{F}$ is an exponential family: $\mathscr{F} = \mathscr{E}_{\nu,f}$ with $\nu(x) = 1$ and $f(x) = x$. The divergence of a pm $P \in \overline{\mathscr{P}}(X)$ from $\mathscr{F}$ is in other words the *multi-information* of P and is denoted by $M(P) = D(P\|\mathscr{F})$, see [3].

Theorem 2.7 (Maximizers of multi-information)
(i) *The maximum value of* $M(\cdot)$ *over all pm's on* $[d]^n$ *is equal to* $(n-1)\ln d$.
(ii) *The maximizers project to the unique pm, uniformly concetrated on* $[d]^n$.
(iii) *The maximizers have the form* $\frac{1}{d}\sum_{j=1}^{d}\delta_{(j,\pi_2(j),\ldots,\pi_n(j))}$, $\pi_2, \ldots, \pi_n \in \mathsf{S}_d$.

Proof. In [3, Corollary 4.10]. □

Remark 2.8 For a fixed $\pi_2, \ldots, \pi_n \in \mathsf{S}_d$, the maximizer is uniformly concentrated on the set $\{(1, \pi_2(1), \ldots, \pi_n(1)), \ldots, (d, \pi_2(d), \ldots, \pi_n(d))\}$.

In fact, a more general form of Theorem 2.7 holds true [3, Theorem 4.3] (for a more general form of X).

3. Proof of Theorem 1.1

Let $X = [d]^n$, define (the permutation group) $\mathsf{G} \le \mathsf{S}_X$ as

$$\mathsf{G} = \{\pi_\rho \in \mathsf{S}_X : \pi_\rho(x) = (x_{\rho(1)}, \ldots, x_{\rho(n)})^\top, x \in X, \rho \in \mathsf{S}_n\} \tag{5}$$

and $\overline{\mathscr{E}} = \overline{\mathscr{E}}_{\mathsf{G}}(X)$ be the family of all G-exchangeable pm's on X.

Recall that the multinomial family $\overline{\mathscr{M}}_{n,d}$ and its state space Z are defined by (2) and the sufficient statistics γ is defined by (3).

Remarks 3.1

(a) For $z, z' \in Z$ and $x, x' \in X$ s.t. $z = \gamma(x)$ and $z' = \gamma(x')$:

$$z = z' \;\Leftrightarrow\; \gamma(x) = \gamma(x') \;\Leftrightarrow\; x' \in \mathsf{G}x \;\Leftrightarrow\; \mathsf{G}x = \mathsf{G}x' \;\Leftrightarrow\; \gamma_{\mathsf{G}}(x) = \gamma_{\mathsf{G}}(x');$$

(b) $\mathscr{F}$ defined by (4) is G-*symmetrical*;

(c) $\overline{\mathscr{E}} \cap \mathscr{F} = \{\underbrace{p \times \ldots \times p}_{n} : p \text{ is a strictly positive pm on } [d]\}$;

(d) $\mathscr{M}_{n,d} = \overline{\mathscr{M}}_{n,d} \cap \mathscr{P}(Z) = \gamma(\mathscr{E} \cap \mathscr{F})$.

Denote by $P_{\pi_2,\ldots,\pi_n}$ the maximizer of $D(\cdot\|\mathscr{F})$ from Theorem 2.7 corresponding to $\pi_2, \ldots, \pi_n \in \mathsf{S}_d$ and let Id be the identity on $[d]$.

$P_{\mathsf{Id},\ldots,\mathsf{Id}}$ is maximizer of $D(\|\mathscr{F})$ over $\overline{\mathscr{P}}(X)$ and even $P_{\mathsf{Id},\ldots,\mathsf{Id}}$ is G-exchangeable, i.e. $P_{\mathsf{Id},\ldots,\mathsf{Id}} \in \overline{\mathscr{E}}$. Therefore, by Corollary 2.6 and Remarks 3.1, the assertions (i) and (ii) follow.

Assume that $P = P_{\pi_2,\ldots,\pi_n} \in \overline{\mathscr{E}}$ is such a maximizer that $P \neq P_{\mathsf{Id},\ldots,\mathsf{Id}}$, i.e. there exists $i \in \{2, \ldots, n\}$ s.t. $\pi_i \not\equiv \mathsf{Id}$, w.l.o.g. let $\pi_2 \not\equiv \mathsf{Id}$. Thus, there exists $j \in [d]$ s.t. $\pi_2(j) \neq j$. By the fact that $P \in \overline{\mathscr{E}}$, there exist $x_3, \ldots, x_n \in [d]$ s.t. $(j, \pi_2(j), x_3, \ldots, x_n) \in \mathsf{s}(P)$ and also $(\pi_2(j), j, x_3, \ldots, x_n) \in \mathsf{s}(P)$.

If $n = 2$, then the assertion (iv) simply follows (with $\pi = \pi_2$).

If $n > 2$, then $x_k = \pi_k(j) = \pi_k(\pi_2(j))$, $k = 3, \ldots, n$ and $j \neq \pi_2(j)$, hence $\pi_3, \ldots, \pi_n$ are not injective and therefore are not permutations and the assertion (iii) follows by contradiction. □

4. Examples

The theoretical results are illustrated by examples summarized in Table 1.

Maximizers of divergence from multinomial family $\overline{\mathscr{M}}_{n,d}$ are computed and the results of the corresponding problem of maximization of multi-information are listed.

5. Discussion

Application of Corollary 2.6 of Theorem 2.5 and the result of N. Ay and A. Knauf, Theorem 2.7 [3, Corollary 4.10], substantially simplified the proof of [8, Proposition 2], and in a more general setting.

Considering Corollary 2.6, a natural problem arises, i.e. to represent forms of the space X, permutation groups G (i.e. the families $\overline{\mathscr{E}} = \overline{\mathscr{E}}_{\mathsf{G}}(X)$ of all G-exchangeable pm's on X) and G-symmetrical exponential families $\mathscr{F}$ for which there exists a G-exchangeable maximizer of $D(\cdot\|\mathscr{F})$ over all pm's on X. In this situation, it is possible to reduce the dimensionality of the problem by sufficient statistics γ_{G}.

TABLE 1. Examples

n	$\arg\max_{\overline{\mathscr{P}}(X)} M(\cdot) \cap \overline{\mathscr{E}}$	$\arg\max_{\overline{\mathscr{P}}(Z)} D(\cdot\|\overline{\mathscr{M}}_{n,d})$
d	$\arg\max_{\overline{\mathscr{P}}(X)} M(\cdot) \setminus \overline{\mathscr{E}}$	$\max_{\overline{\mathscr{P}}(Z)} D(\cdot\|\overline{\mathscr{M}}_{n,d}) = \max_{\overline{\mathscr{P}}(X)} M(\cdot)$
		μ
$n=2$	$\frac{1}{2}(\delta_{11}+\delta_{22}), \frac{1}{2}(\delta_{12}+\delta_{21})$	$\frac{1}{2}(\delta_{20}+\delta_{02}), \delta_{11}$
$d=2$		$\ln 2$
		$\frac{1}{4}(\delta_{20}+\frac{1}{4}\delta_{02})+\frac{1}{2}\delta_{11}$
$n=3$	$\frac{1}{2}(\delta_{111}+\delta_{222})$	$\frac{1}{2}(\delta_{30}+\delta_{03})$
$d=2$	$\frac{1}{2}(\delta_{112}+\delta_{221})$	$2\ln 2$
	$\frac{1}{2}(\delta_{121}+\delta_{212})$	$\frac{1}{8}(\delta_{30}+\delta_{03})+\frac{3}{8}(\delta_{12}+\delta_{21})$
	$\frac{1}{2}(\delta_{211}+\delta_{122})$	
$n=2$	$\frac{1}{3}(\delta_{11}+\delta_{22}+\delta_{33})$	$\frac{1}{3}(\delta_{200}+\delta_{020}+\delta_{002})$
$d=3$	$\frac{1}{3}(\delta_{11}+\delta_{23}+\delta_{32})$	$\frac{1}{3}\delta_{200}+\frac{2}{3}\delta_{011}$
	$\frac{1}{3}(\delta_{13}+\delta_{22}+\delta_{31})$	$\frac{1}{3}\delta_{020}+\frac{2}{3}\delta_{101}$
	$\frac{1}{3}(\delta_{12}+\delta_{21}+\delta_{33})$	$\frac{1}{3}\delta_{002}+\frac{2}{3}\delta_{110}$
	$\frac{1}{3}(\delta_{12}+\delta_{23}+\delta_{31})$	$\ln 3$
	$\frac{1}{3}(\delta_{13}+\delta_{21}+\delta_{32})$	$\frac{1}{9}(\delta_{200}+\delta_{020}+\delta_{002})+$
		$+\frac{2}{9}(\delta_{011}+\delta_{101}+\delta_{110})$

One interesting special case arises when the state space X is a cartesian product of finite spaces, $\mathscr{F}$ is a family of pm's factorizable w.r.t. a hypergraph (hierarchical model) and G is defined s.t. $\mathscr{F}$ is G-symmetrical.

References

[1] AY, N.: *An information-geometric approach to a theory of pragmatic structuring*. Ann. Prob. **30** (1) (2002), 416–436.

[2] AY, N., WENNEKERS, T.: *Dynamical properties of strongly interacting Markov chains*. Neural Networks **16** (2003), 1483–1497.

[3] AY, N., KNAUF, A.: *Maximizing multi-information*. Kybernetika **45** (2006), 517–538.

[4] BROWN, L. D.: Fundamentals of Statistical Exponential Families. IMS Lecture Notes Monograph Series Vol. 9, Hayward, CA (1986).

[5] CSISZÁR, I., MATÚŠ, F.: *Information projections revisited*. IEEE Trans. Inform. Theory **49** (2003), 1474–1490.

[6] CSISZÁR, I., MATÚŠ, F.: *Generalized maximum likelihood estimates for exponential families*. Probab. Theory Related Fields **141** (2008), 213–246.

[7] Matúš, F., Ay, N.: *On maximization of the information divergence from an exponential family*. Proceedings of WUPES'03 (ed. J. Vejnarová), University of Economics, Prague (2003), 199–204.
[8] Matúš, F.: *Maximization of information divergences from binary i.i.d. sequences*. Proceedings of IPMU 2004, Perugia, **2** (2004), 1303–1306.
[9] Matúš, F.: *Optimality conditions formaximizers of the information divergence from an exponential family*. Kybernetika **41** (2007), 731–746.
[10] Matúš, F.: *Divergence from factorizable distributions and matroid representations by partitions*. IEEE Trans. Inform. Theory **55** (12) (2009), 5375–5381.
[11] Rauh, J.: *Finding the maximizers of the information divergence from an exponential family*. arXiv:0912.4660 (2009).