Proceedings IPMU 2004, July 4–9 2004, Perugia, Italy, Vol. 2, 1303–1306

Maximization of Information Divergences Related to Binary i.i.d. Sequences*

František Matúš†
Prague, Czech Republic
matus@utia.cas.cz

Abstract

Two instances of maximization of the information divergence from an exponential family are studied: infimum of the information divergences of a probability measure P from the distributions of $\{0,1\}$-valued i.i.d. sequences, and analogous infimum over the distributions of sums of the variables are maximized over P. The maxima and global maximizers are found explicitly.

Keywords: Infomax principle, Information divergence, Exponential family, Binary sequences, Binomial family.

1 Introduction

The information divergence of a probability measure (pm) P on a finite set X from an exponential family $\mathcal{E}$ of pm's on X is

$$D(P\|\mathcal{E}) \triangleq \inf_{Q\in\mathcal{E}}\ D(P\|Q)$$

where the information divergence $D(P\|Q)$ of P from Q equals

$$\textstyle\sum_{x\in X:\,P(x)>0} P(x)\ln\frac{P(x)}{Q(x)}$$

if P is dominated by Q and $+\infty$ otherwise.

*This work was supported by Grant Agency of Academy of Sciences of the Czech Republic under the grant A1075104 and by GA CR under the grant 201/04/0393.

†Institute of Information Theory and Automation, Pod vod. věží 4, 182 08 Prague, Czech Republic.

An interest in local maximizers of the function $D(\cdot\|\mathcal{E})$ emerged in [1, 2], and stems from probabilistic models of neural networks based on *infomax* principles for adaptation and learning [6, 8]. Though some general results on the maximizers are available [1, 7], to cope with an individual family an additional effort may be needed. The local maximizers for an exponential family $\mathcal{E}$ which is a convex set of pm's are described in [7]. The case when $\mathcal{E}$ consists of positive product measures, thus $D(P\|\mathcal{E})$ coincides with the mutual information in P, is attacked in [3]. This note presents the maxima and global maximizers of $D(\cdot\|\mathcal{E})$ for the family $\mathcal{E}$ of positive power measures on $\{0,1\}^n$ and for Binomial family.

2 Main results

The i.i.d. sequence of n binary random variables taking the values 1 and 0 with probabilities p and $1-p$ is distributed as

$$Q_p(x) = p^{\Sigma x_i}\cdot(1-p)^{n-\Sigma x_i}$$

where $x=(x_1,\dots,x_n)\in\{0,1\}^n$. For a pm P the infimum of $D(P\|Q_p)$ subject to $0<p<1$ is, by simple calculus,

$$\rho(P) \triangleq n\,h\big(\tfrac{1}{n}\,\Sigma_x\,P(x)\,\Sigma_i x_i\big) - H(P)$$

where $H(P)$ is the Shannon entropy of P and $h(p)$ denotes the Shannon entropy of a binary variable taking one value with probability p. The entropies involve natural logarithms. Obviously, $\rho=0$ for $n=1$, and

$$\rho(P) \leqslant n\ln 2$$

with the equality if and only if $P(x) = 1$ for some x with $\Sigma\, x_i = \frac{n}{2}$. This takes place only for n even. For n odd the maximization of ρ seems to be nontrivial.

Proposition 1. *The maximum of ρ over the pm's P on $\{0,1\}^n$ equals $n\, h(\frac{1}{n} \lfloor \frac{n}{2} \rfloor)$. When $n \geqslant 2$ a pm P attains the maximum if and only if $P(x) = 1$ for some x having $\lfloor \frac{n}{2} \rfloor$ zeros or ones.*

Proof. To avoid trivialities let $n \geqslant 3$ odd.

By the properties of Shannon entropy, the function ρ restricted to the polytope of all pm's P with $\Sigma_x P(x)\, \Sigma_i x_i$ equal to a fixed number is strictly convex. Then any (global) maximizer P must be an extreme point of such a polytope[‡]. This implies that $P(x) = 1$ for some x with $\Sigma x_i = j$, or P is positive only for such an x and y with $\Sigma y_i = k$, say $j < k$.

In the former case, $\rho(P) = n\, h(\frac{j}{n})$ which maximizes obviously when $j = \frac{n\pm 1}{2}$. In the latter case with $P(y) = t$ strictly between 0 and 1, $\rho(P)$ equals

$$\phi_{j,k}(t) \triangleq n\, h\big(\tfrac{j}{n}(1-t) + \tfrac{k}{n}t\big) - h(t).$$

Thus, it suffices to prove

$$\phi_{j,k}(t) < n\, h(\tfrac{n-1}{2n}) \tag{1}$$

for $0 \leqslant j < k \leqslant n$ and $0 < t < 1$. Since

$$\phi_{j,k}(t) = \phi_{n-k,n-j}(1-t) \tag{2}$$

$t \geqslant \frac{1}{2}$ can be assumed.

The inequality

$$n \ln 2 - \tfrac{1}{n} < n\, h(\tfrac{n-1}{2n})$$

follows from $\ln(1+t) < t$ for $t = \pm\frac{1}{n}$. Then (1) holds whenever $h(t) \geqslant \frac{1}{n}$. Otherwise,

$$h(t) \geqslant 2(1-t) \ln 2\,, \quad t \geqslant \tfrac{1}{2}, \tag{3}$$

and $h(t) < \frac{1}{n}$ imply $t > 1 - \frac{1}{n}$ and, in turn,

$$\tfrac{j}{n}(1-t) + \tfrac{k}{n}t > \tfrac{k-1}{n}.$$

Since this sum is obviously less than $\frac{k}{n}$ the validity of (1) follows unless $k = \frac{n+1}{2}$.

[‡]An analogous argument works more generally in exponential families, [7, Corollary 2].

If $j > 0$ then

$$\tfrac{j}{n}(1-t) + \tfrac{n+1}{2n}t = \tfrac{j-1}{n}(1-r) + \tfrac{n+1}{2n}r$$

holds for some r satisfying $t < r < 1$. This implies that $\phi_{j,k}(t)$ is less than $\phi_{j-1,k}(r)$. Now, it suffices to prove (1) for $j = 0$ and $k = \frac{n+1}{2}$.

Take the derivatives $\phi'_{j,k}$

$$\ln \tfrac{t}{1-t} + (k-j) \ln \tfrac{n-j(1-t)-kt}{j(1-t)+kt}$$

and $\phi''_{j,k}$

$$\tfrac{1}{t} + \tfrac{1}{1-t} - (k-j)^2 \Big[\tfrac{1}{j(1-t)+kt} + \tfrac{1}{n-j(1-t)-kt}\Big].$$

For $j = 0$ the ratios collect together and

$$\phi''_{0,k}(t) = \tfrac{tk(n-1)-n(k-1)}{t(1-t)(n-kt)} \tag{4}$$

which rewrites for k equal to $\ell \triangleq \frac{n+1}{2}$ as

$$\phi''_{0,\ell}(t) = \tfrac{(n-1)[t(n+1)-n]}{t(1-t)[2n-(n+1)t]}.$$

Therefore, $\phi_{0,\ell}$ is strictly concave for $t < \frac{n}{n+1}$ and strictly convex otherwise. Since

$$\phi'_{0,\ell}(\tfrac{n}{n+1}) = \ln n > 0$$

$\phi_{0,\ell}$ is less than its limit $n\, h(\frac{n-1}{2n})$ when $t \nearrow 1$, and thus (1) takes place. $\square$

Let R_p be the image of Q_p under $x \mapsto \Sigma\, x_i$. For a pm P on $\{0,1,\dots,n\}$ the infimum of $D(P \| R_p)$ subject to $0 < p < 1$ is, by simple calculus,

$$\tau(P) \triangleq n\, h\big(\tfrac{1}{n}\Sigma_k\, P(k)\, k\big) - H(P) - \Sigma_k\, P(k) \ln \tbinom{n}{k}.$$

Obviously, $\tau = 0$ for $n = 1$.

Proposition 2. *The maximum of τ over the pm's P on $\{0,1,\dots,n\}$ equals $(n-1)\ln 2$. When $n \geqslant 2$ a pm P attains the maximum if and only if $P(0) = P(n) = \frac{1}{2}$ or $P(1) = 1$, $n = 2$.*

Proof. As in the previous proof, the function τ restricted to the polytope of all pm's P with $\Sigma_k P(k)\, k$ equal to a fixed number is strictly convex. Then any maximizer P must be an extreme point of such a polytope which implies that $P(j)=1$ for some j with $0\leqslant j\leqslant n$ or P is positive only for some j and k with $0\leqslant j<k\leqslant n$.

In the former case, $\tau(P)$ equals

$$\delta_j \triangleq n\,h(\tfrac{j}{n})-\ln\tbinom{n}{j}=\ln\tfrac{n^n}{n!}\tfrac{j!}{j^j}\tfrac{(n-j)!}{(n-j)^{n-j}}.$$

Then $\delta_{j+1}>\delta_j$ for $0<j<\lfloor\frac{n}{2}\rfloor$ because

$$e^{\delta_{j+1}-\delta_j}=\Big(1+\tfrac{1}{n-j-1}\Big)^{n-j-1}\Big(1+\tfrac{1}{j}\Big)^{-j}>1$$

due to $(1+\frac{1}{j})^j\nearrow e$. Since obviously $\delta_j=\delta_{n-j}$ and $\delta_0=\delta_n=0$ it follows that $\delta_j\leqslant\delta_\ell$ where $\ell\triangleq\lfloor\frac{n}{2}\rfloor$. The equality takes place here if and only if $j=\ell$. In addition, $\delta_\ell\leqslant(n-1)\ln 2$ and this inequality is tight if and only if $n=2$.

In the latter case with $P(k)=t$, $\tau(P)$ equals

$$\psi_{j,k}(t)\triangleq\phi_{j,k}(t)-(1-t)\ln\tbinom{n}{j}-t\ln\tbinom{n}{k}.$$

Thus, it suffices to prove

$$\psi_{j,k}(t)\leqslant(n-1)\ln 2 \qquad (5)$$

for $0\leqslant j<k\leqslant n$ and $0<t<1$, and the tightness here to be equivalent to $j=0$, $k=n$ and $t=\frac{1}{2}$.

If $0<j<k<n$ then

$$(1-t)\ln\tbinom{n}{j}+t\ln\tbinom{n}{k}\geqslant\ln n>\ln 2$$

and (5) is valid with a strict inequality. Then, by the symmetry (2),

$$\psi_{j,k}(t)=\psi_{n-k,n-j}(1-t),$$

and it suffices to consider $j=0$.

If $t<\frac{n}{k+n}$ then

$$\big(\tfrac{1}{2}-\tfrac{k}{n}t\big)^2>\big(\tfrac{1}{2}-t\big)^2$$

and $h(\frac{k}{n}t)<h(t)$, thus

$$\psi_{0,k}(t)<(n-1)h(t)-t\ln\tbinom{n}{k}.$$

Hence, (5) with $<$ is valid unless $k=n$.

Otherwise, $t\geqslant\frac{n}{k+n}$ implies $t\geqslant\frac{1}{2}$, and

$$\psi_{0,k}(t)\leqslant nh(\tfrac{k}{n}t)-2(1-t)\ln 2-t\ln\tbinom{n}{k}$$

on account of (3). If $k<n$ then $\binom{n}{k}\geqslant 2$ and

$$\psi_{0,k}(t)\leqslant n\ln 2-2\ln 2+t\ln 2<(n-1)\ln 2.$$

Finally, $j=0$ and $k=n$. Strict concavity of $\psi_{0,n}$ is a consequence of

$$\psi''_{0,n}(t)=\phi''_{0,n}(t)=-\tfrac{n-1}{t(1-t)}<0,$$

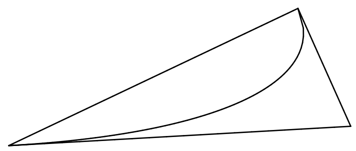
obtained from (4). Since $\psi'_{0,n}(\frac{1}{2})$ vanishes $\psi_{0,n}$ is bounded above by $\psi_{0,n}(\frac{1}{2})=(n-1)\ln 2$ and (5) holds. $\square$

3 Example

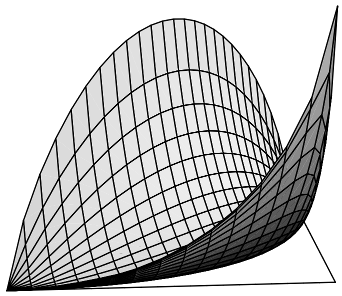
If the pm's on $\{0,1,2\}$ are identified with the points of a triangle then Binomial family

$$\big\{\,((1-p)^2,2p(1-p),p^2)\colon 0\leqslant p\leqslant 1\big\}$$

with the parameter $n=2$ corresponds to the curve



connecting two vertices of the triangle. If τ, which is the information divergence from Binomial family, is viewed as a function on the triangle then the graph of τ takes the form



drawing values in the vertical direction. Here, the two global maxima of τ are easily recognizable, cf. Proposition 2. Obviously, $\tau\geqslant 0$ and the equality takes place along the curve. The curves on the graph were chosen to represent the values of τ on one-dimensional exponential and linear families.

4 Discussion

The information divergence from a full exponential family $\mathcal{E}$ of pm's on a finite set X can be expressed also as

$$D(P\|\mathcal{E}) \triangleq \min_{Q\in \mathrm{cl}(\mathcal{E})} D(P\|Q)$$

where $\mathrm{cl}(\mathcal{E})$ is the closure of $\mathcal{E}$ considered for a subset of a Euclidean space. This minimum is always finite when the pm's from $\mathcal{E}$ are positive. Then, for each pm P on X it is uniquely attained by some pm $\Pi_{P\to \mathrm{cl}(\mathcal{E})} \in \mathrm{cl}(\mathcal{E})$ that can be viewed as an information projection of P onto $\mathrm{cl}(\mathcal{E})$. If $x^1, \dots, x^N$ is an i.i.d. sample from an unknown pm from $\mathcal{E}$ and P is the empirical distribution, thus $P(x)$ for $x \in X$ is the proportion of k's with $x^k = x$, $1 \leqslant k \leqslant N$, then $\Pi_{P\to \mathrm{cl}(\mathcal{E})}$ can be interpreted as a maximum likelihood (ML) estimate in $\mathrm{cl}(\mathcal{E})$ or a generalized ML estimate to $\mathcal{E}$, from the sample. In particular, if the projection $\Pi_{P\to \mathrm{cl}(\mathcal{E})}$ belongs to $\mathcal{E}$ then, under any parametrization of $\mathcal{E}$, any ML estimate from the sample parameterizes $\Pi_{P\to \mathrm{cl}(\mathcal{E})}$. These facts are special cases of results of [5] on general exponential families; results on ML estimation in $\mathrm{cl}(\mathcal{E})$ go back to [4, pp. 154-5].

If P is the empirical distribution from a sample then the quantity $D(P\|\mathcal{E})$, equal to $D(P\|\Pi_{P\to \mathrm{cl}(\mathcal{E})})$, measures the fit between the sample and the statistical model $\mathcal{E}$. If $D(P\|\mathcal{E})$ is small the sample is believed to follow the distribution $\Pi_{P\to \mathrm{cl}(\mathcal{E})}$ from the model $\mathcal{E}$ or from its closure. Propositions 1 and 2 identify samples that are the most poorly modeled by two families. For example, Proposition 2 implies that the worst samples $x^1, \dots, x^N$ of even length N for Binomial family with the parameter $n \geqslant 3$ consist of 0's and n's in the same proportion.

References

[1] N. Ay (2002) An information-geometric approach to a theory of pragmatic structuring. *The Annals of Probab.* **30** 416–36.

[2] N. Ay (2002) Locality of Global Stochastic Interaction in Directed Acyclic Networks. *Neural Computation* **14** 2959–80.

[3] N. Ay and A. Knauf (2003) Maximizing Multi-Information. Preprint Nr. 42/2003, M. Planck Inst. for Mathematics in the Sciences, Leipzig, Germany (submitted).

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

[5] I. Csiszár and F. Matúš (2003) Information projections revisited. *IEEE Transactions IT* **49** 1474–90.

[6] R. Linsker (1988) Self-organization in a perceptual network. *IEEE Computer* **21** 105–17.

[7] F. Matúš and N. Ay (2003) On maximization of the information divergence from an exponential family. In: *Proceedings of WUPES'2003*, 199–204, Prague, Czech Republic, Sept. 2003.

[8] G. Tononi, O. Sporns and G.M. Edelman (1994) A measure for brain complexity: relating functional segregation and integration in the nervous system. *Proc. Natl. Acad. Sci. USA* **91** 5033–37.