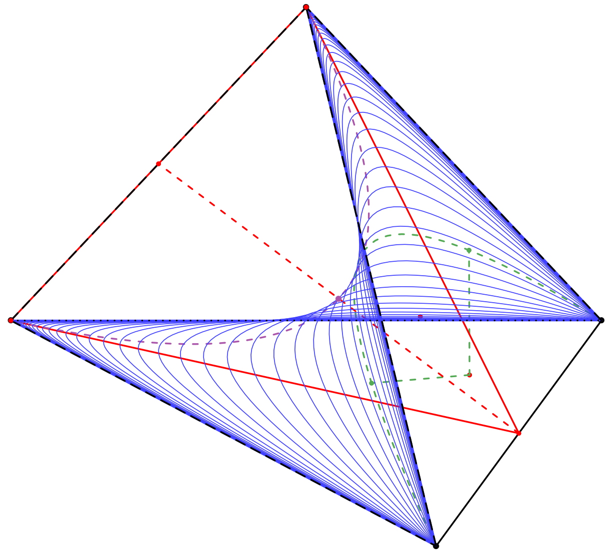


Maximization of Information Divergence from Multinomial Distributions

ROBUST 2010, Králiky, 4^{th} February

Jozef Juríček
Charles University in Prague
Faculty of Mathematics and Physics
Department of Probability and Mathematical Statistics

Supervisor: **Ing. František Matúš, CSc.**
Academy of Sciences of the Czech Republic
Institute of Information Theory and Automation
Department of Decision-Making Theory

Outline

References

[1] Ay, N., Knauf, A. (2006). Maximizing multi-information. *Kybernetika* **45** 517-538.

[2] Csiszár, I., Matúš, F. (2003). Information projections revisited. *IEEE Transactions Information Theory* **49** 1474-1490.

[3] Matúš, F. (2004). Maximization of information divergences from binary i.i.d. sequences. *Proceedings IPMU* (2004) **2** 1303-1306. Perugia, Italy.

- this problem is a generalization of the problem solved in [3]

1 Introduction

- more general problem of maximization of information divergence from exponential family

  emerged in probabilistic models for evolution and learning in neural networks, based on infomax principles

- maximizers admit interpretation as stochastic systems with high complexity w.r.t. exponential family

- weakly speaking, the problem is to *find all (empirical) distributions (data), which lies farthest from the model,*

  when modelling by the exponential family; in the sense of Kullback-Leibler divergence *and* maximum likelihood estimation method

- in other words, in specific situation of multinomial family (N id.indep.trials each with n possible outcomes; $N \geqq 2, n \geqq 2$ fixed)

$$\overline{\mathcal{M}} = \left\{ \left(Q(z) = \binom{N}{z} \prod_{j=1}^{n} p_j^{z_j} \right)_{z\in Z} : p \in [0;1]^n, \sum_{j=1}^{n} p_j = 1 \right\}, \qquad Z = \left\{ z \in \{0,1,\ldots,N\}^n : \sum_{j=1}^{n} z_j = N \right\} \text{ the state space or configurations}, \qquad \binom{N}{z} = \frac{N!}{z_1! z_2! \ldots z_n!}$$

calculate $\sup_{P\in\overline{\mathcal{P}}(Z)} D(P\|\overline{\mathcal{M}})$

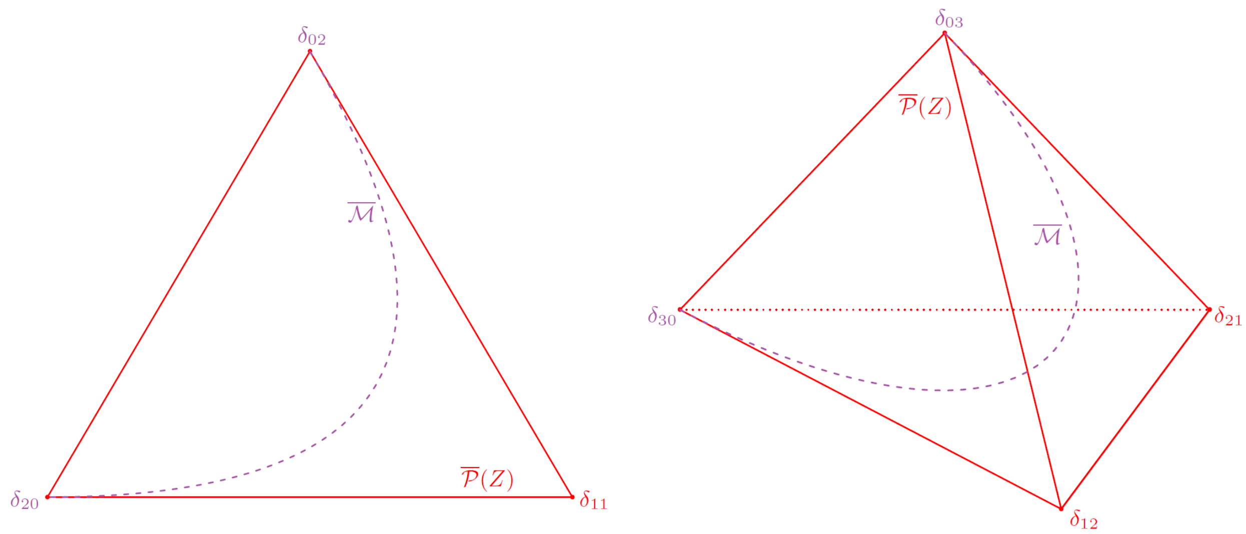
and find all maximizers $\arg\sup_{P\in\overline{\mathcal{P}}(Z)} D(P\|\overline{\mathcal{M}})$

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

$$D(P\|\mathcal{E}) = \inf_{Q\in\mathcal{E}} D(P\|Q) = \min_{Q\in\overline{\mathcal{E}}} D(P\|Q)$$

$$\mathsf{s}(P) = \{z \in Z : P(z) > 0\}$$

$\overline{\mathcal{P}}(Z)$ simplex of all pm's on Z



$N = 2, \quad n = 2$ $\qquad$ $N = 3, \quad n = 2$

Multinomial family $\overline{\mathcal{M}}$ in simplex $\overline{\mathcal{P}}(Z)$

2 Preliminaries

- *Exponential family* $\mathcal{E}_{\mu,f} = \left\{ Q_{\mu,f,\vartheta} \sim \left(e^{\langle\vartheta, f(z)\rangle}\mu(z)\right)_{z\in Z} : \vartheta \in \mathbb{R}^d \right\}$

  μ nonzero *reference measure* on a finite set Z
  $f : Z \to \mathbb{R}^d$ the *directional statistics*

Theorem 1 *Let* $P \in \overline{\mathcal{P}}(Z)$, μ *be a strictly positive measure on* Z *(*$\mathsf{s}(\mu) = Z$*),* $\mathcal{E} = \mathcal{E}_{\mu,f}$ *be the exponential family.*
Then there exists **unique** rl-projection (generalized MLE; GMLE) $P^{\mathcal{E}} \in \overline{\mathcal{E}}$, *s.t.* $P^{\mathcal{E}} = \arg\inf_{Q\in\mathcal{E}} D(P\|Q) = \arg\min_{Q\in\overline{\mathcal{E}}} D(P\|Q)$.
For P *empirical distribution, s.t.* $P^{\mathcal{E}} \in \mathcal{E}$, $P^{\mathcal{E}}$ *is the MLE for data with emprical distribution* P. Details in [2].

- $\overline{\mathcal{M}} = \overline{\mathcal{E}}_{\mu,f}$ with $f(z) = z$, $\mu(z) = \binom{N}{z}$, $z \in Z = \left\{ z \in \{0, 1, \ldots, N\}^n : \sum_{j=1}^n z_j = N \right\}$

- denote $X = \{1, \ldots, n\}^N$, for permutation $\pi \in \mathcal{S}_N, \pi : \{1, \ldots, N\} \overset{1-1}{\leftrightarrow} \{1, \ldots, N\}$, $x \in X$ and $P \in \overline{\mathcal{P}}(X)$ dentote $x^\pi = (x_{\pi(1)}, \ldots, x_{\pi(N)})^\top$, $P^\pi(x) = P(x^\pi)$
- *Exchangable distributions* $\overline{\mathcal{E}} := \{P \in \overline{\mathcal{P}}(X) : P(x) = P(x^\pi), x \in X; \forall \pi \in \mathcal{S}_N\}$
- *1-factorizable distributions* $\overline{\mathcal{F}} := \{Q \in \overline{\mathcal{P}}(X) : Q(x) = \prod_{i=1}^N Q_i(x_i), x \in X\}$, $Q_i(x_i) = \sum_{\substack{x' \in X: \\ x'_i = x_i}} Q(x')$, $i = 1, \ldots, N$ are the marginals

Lemma 2 Denote $X^z := \left\{x \in X : \forall j \in \{1, \ldots, n\} : |\{i \in \{1, \ldots, N\} : x_i = j\}| = z_j\right\}, z \in Z$ (remind $Z = \left\{z \in \{0, 1, \ldots, N\}^n : \sum_{j=1}^n z_j = N\right\}$). It holds:

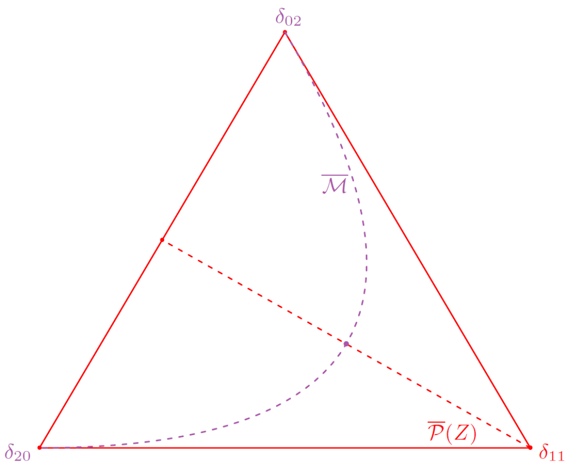
(i) The mapping $h : \overline{\mathcal{P}}(Z) \overset{1-1}{\leftrightarrow} \overline{\mathcal{E}}$ such that $h(P) = P'$, $P'(x) = \frac{P(z)}{\binom{N}{z}}$ for $z \in Z$ s.t. $x \in X^z$ is a bijection,
$h\restriction_{\overline{\mathcal{M}}} : \overline{\mathcal{M}} \overset{1-1}{\leftrightarrow} \overline{\mathcal{E}} \cap \overline{\mathcal{F}} = \overline{\mathcal{E} \cap \mathcal{F}}$ and for $h^{-1} : \overline{\mathcal{E}} \overset{1-1}{\leftrightarrow} \overline{\mathcal{P}}(Z)$, the inverse of h, $h^{-1}(P') = P$ and $P(z) = \binom{N}{z} P'(x)$ for any $x \in X^z$.

(ii) For any $P \in \overline{\mathcal{P}}(Z)$, $Q \in \overline{\mathcal{M}}$, it holds $D(P\|Q) = D\big(h(P)\|h(Q)\big)$.

(iii) For any $P \in \overline{\mathcal{E}}$, $Q \in \overline{\mathcal{F}} \setminus \overline{\mathcal{E} \cap \mathcal{F}}$, there exists $\pi \in \mathcal{S}_N$, such that for Q^π, $Q^\pi(x) = Q(x^\pi)$, it holds $Q^\pi \not\equiv Q$ and $D(P\|Q) = D(P\|Q^\pi)$.

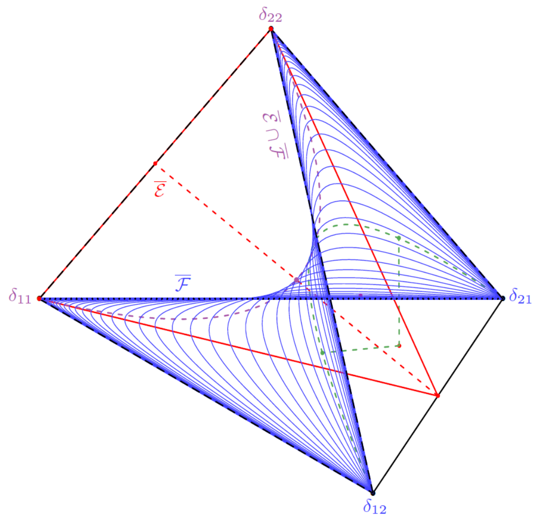
(iv) For any $P \in \overline{\mathcal{E}}$: $D(P\|\mathcal{F}) = \inf_{Q \in \mathcal{E} \cap \mathcal{F}} D(P\|Q)$ and $\arg\inf_{Q \in \mathcal{E} \cap \mathcal{F}} D(P\|Q) = P^{\mathcal{F}} \in \overline{\mathcal{E} \cap \mathcal{F}}$.

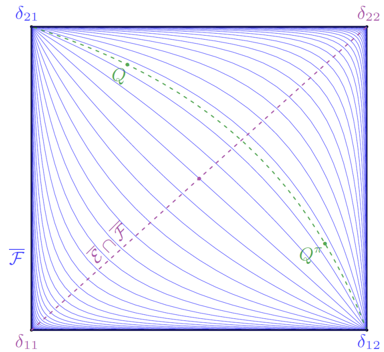
(v) $\sup_{P \in \mathcal{P}(Z)} D(P\|\mathcal{M}) = \sup_{P \in \mathcal{E}} D(P\|\mathcal{E} \cap \mathcal{F}) = \sup_{P \in \mathcal{E}} D(P\|\mathcal{F}) \leq \sup_{P \in \mathcal{P}(X)} D(P\|\mathcal{F})$ and $\arg\sup_{P \in \mathcal{P}(Z)} D(P\|\mathcal{M}) = h^{-1}(\arg\sup_{P \in \mathcal{E}} D(P\|\mathcal{E} \cap \mathcal{F})) = h^{-1}(\arg\sup_{P \in \mathcal{E}} D(P\|\mathcal{F}))$.



$$\overline{\mathcal{P}}(Z) \underset{h^{-1}}{\overset{h}{\rightleftarrows}} \overline{\mathcal{E}} \subset \overline{\mathcal{P}}(X)$$

$$\overline{\mathcal{M}} \underset{h^{-1}\restriction_{\overline{\mathcal{M}}}}{\overset{h\restriction_{\overline{\mathcal{M}}}}{\rightleftarrows}} \overline{\mathcal{E}} \cap \overline{\mathcal{F}}$$





- $P \in \mathcal{P}(X)$: $D(P\|\mathcal{F}) = M(P)$, the *multi-information*

Theorem 3 (Maximizing the multi-information) $\arg\sup\limits_{P\in\mathcal{P}(X)} D(P\|\mathcal{F}) = \left\{ P_\Pi = \frac{1}{n}\sum_{j=1}^{n} \delta_{(j,\pi_2(j),\ldots,\pi_N(j))^\top} : \Pi = (\pi_2,\ldots,\pi_N) \in \mathcal{S}_n^{N-1} \right\}$, $D(P_\Pi\|\mathcal{F}) = (N-1)\ln n$ *and* $P_\Pi^{\mathcal{F}} = U^X = \frac{1}{n^N}\sum_{x\in X}\delta_x$, $\forall \Pi \in \mathcal{S}_n^{N-1}$.

Main result of [1].

3 Result

Notation: $e^{j,j} = (0,\ldots,0,1_j,0,\ldots,0_n)^\top$ $\quad \epsilon^{j,j} = \delta_{2e^{j,j}}$

$e^{k,l} = (0,\ldots,0,1_k,0,\ldots,0,1_l,0,\ldots,0_n)^\top$ $\quad \epsilon^{k,l} = 2\delta_{e^{k,l}}, k < l$

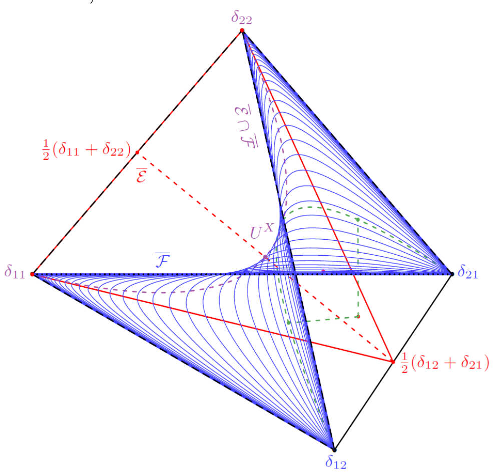
Corollary 4 (Maximizing $D(\cdot\|\overline{\mathcal{M}})$) $\arg\sup_{P\in\mathcal{P}(Z)} D(P\|\mathcal{M}) = h^{-1}\left(\overline{\mathcal{E}} \cap \arg\sup_{P\in\mathcal{P}(X)} D(P\|\mathcal{F})\right)$.

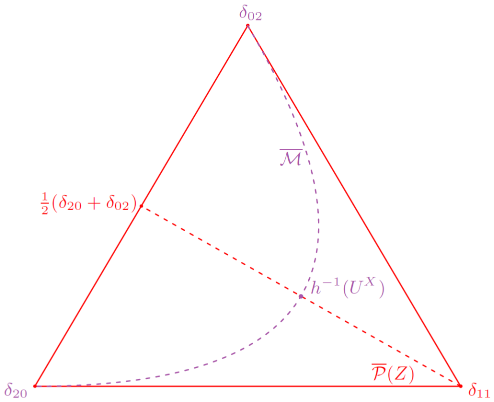
For $N = 2$, $\arg\sup\limits_{P\in\mathcal{P}(Z)} D(P\|\mathcal{M}) = \left\{ P_\pi = \frac{1}{n}\left(\sum_{\substack{j\in\{1,\ldots,n\}:\\ j\le\pi(j)}} \epsilon^{j,\pi(j)} \right), \pi \in \mathcal{S}_n : \begin{array}{l} \forall j,k \in \{1,\ldots,n\}: \\ [\pi(j)=k] \Rightarrow [\pi(k)=j] \end{array} \right\}$. *For* $N > 2$, *the only maximizer* $P_{\mathrm{Id}} = h^{-1}(U^X) = \frac{1}{n}\sum_{j=1}^{n}\delta_{Ne^{j,j}}$.

$\sup\limits_{P\in\mathcal{P}(Z)} D(P\|\mathcal{M}) = (N-1)\ln(n)$ *and for every maximizer* $P_{\sup}$, *it holds* $P_{\sup}^{\mathcal{M}}(z) = \frac{\binom{N}{z}}{n^N}, z \in Z$.

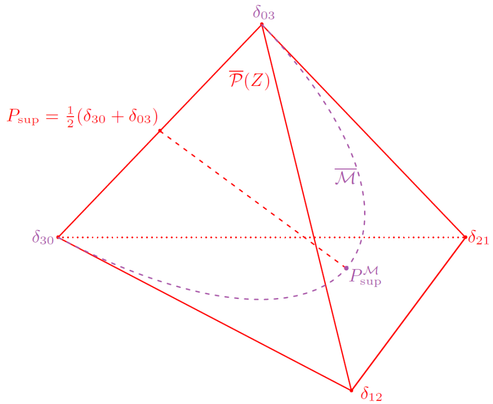
- application of Theorem 1, Lemma 2 and Theorem 3 essentially simplified proof (and in more general situation) than proof given in [3]

$N = 2, n = 2$





$N = 3, n = 2$



$N = 2, n = 3$

$\arg\sup_{\mathcal{P}(X)} D(P\|\mathcal{F}) =$
$\{ \frac{1}{3}(\delta_{11}+\delta_{22}+\delta_{33}), \frac{1}{3}(\delta_{11}+\delta_{23}+\delta_{32}),$
$\frac{1}{3}(\delta_{13}+\delta_{22}+\delta_{31}), \frac{1}{3}(\delta_{12}+\delta_{21}+\delta_{33}),$
$\frac{1}{3}(\delta_{12}+\delta_{23}+\delta_{31}), \frac{1}{3}(\delta_{13}+\delta_{21}+\delta_{32}) \}$

$\arg\sup_{\mathcal{E}} D(P\|\mathcal{E}\cap\mathcal{F}) =$
$\{ \frac{1}{3}(\delta_{11}+\delta_{22}+\delta_{33}), \frac{1}{3}(\delta_{11}+\delta_{23}+\delta_{32}),$
$\frac{1}{3}(\delta_{13}+\delta_{22}+\delta_{31}), \frac{1}{3}(\delta_{12}+\delta_{21}+\delta_{33}) \}$

$\arg\sup_{\mathcal{P}(Z)} D(P\|\mathcal{M}) =$
$\{ \frac{1}{3}(\delta_{200}+\delta_{020}+\delta_{002}), \frac{1}{3}\delta_{200}+\frac{2}{3}\delta_{011},$
$\frac{1}{3}\delta_{020}+\frac{2}{3}\delta_{101}, \frac{1}{3}\delta_{002}+\frac{2}{3}\delta_{110} \}$

$\sup_{\mathcal{P}(Z)} D(P\|\mathcal{M}) = \ln 3$

4 Further research

(?) for $\overline{\mathcal{F}_k}$, the k-*factorizable*, study of $D\left(\cdot\,\|h^{-1}(\overline{\mathcal{E}}\cap\overline{\mathcal{F}_k})\right)$