Finding the Maximizers of the Information Divergence from an Exponential Family

Summary of the thesis submitted by Johannes Rauh

Motivation. This thesis studies the maximization of the information divergence from an exponential family on a finite set, a problem first formulated by Ay in [1]. A special case is the maximization of the mutual information or multiinformation, a mathematical problem arising in the study of information theoretic principles underlying learning of neural networks. Moreover, the information divergence from an exponential family has been proposed as a complexity measure. Upper bounds and properties of the maximizers can be used to characterize these complexity measures. A third motivation coming from machine learning is the search for low-dimensional exponential families that can approximate arbitrary probability distributions well.

Theoretical results. As Matúš showed [4], any maximizer of $D_{\mathcal{E}} : P \mapsto \inf_{Q\in\mathcal{E}} D(P\|Q)$ is a *projection point*. Projection points appear in pairs (P^+, P^-) with disjoint supports, and the difference vector $P^+ - P^-$ is a *facial difference of projection points* (f.d.p.), i.e. $P^+ - P^-$ belongs to the normal space $\mathcal{N}$ of $\mathcal{E}$, and the rI-projection $P_{\mathcal{E}}$ of P^+ and P^- lies in the convex hull of P^+ and P^-. Let

$$\overline{D}_{\mathcal{E}} : u \in \mathcal{N} \mapsto \sum_x u(x) \log \frac{|u(x)|}{\nu_x} = D(u^+\|\nu) - D(u^-\|\nu).$$

Then any f.d.p. $u = P^+ - P^-$ satisfies

$$\exp(D_{\mathcal{E}}(P^+)) = 1 + \exp(D(P^+\|\nu) - D(P^-\|\nu)). \tag{1}$$

In particular, $\max D_{\mathcal{E}} \geq \log(2)$, unless $\overline{\mathcal{E}}$ contains all probability distributions.

Theorem 1. *There is a bijection between projection points and f.d.p.s. that maps the*

$$\left\{\begin{array}{c}\text{local maximizers}\\ \text{global maximizers}\\ \text{critical points}\end{array}\right. \text{of } D_{\mathcal{E}} \text{ onto the } \left\{\begin{array}{c}\text{local maximizers}\\ \text{global maximizers}\\ \text{critical points}\end{array}\right. \text{of } \overline{D}_{\mathcal{E}}.$$

Studying $\overline{D}_{\mathcal{E}}$ instead of directly $D_{\mathcal{E}}$ has two advantages:

1. The dimensionality is reduced.
2. To compute $\overline{D}_{\mathcal{E}}$ it is not necessary to rI-project.

Examples. If $\mathcal{E}$ is algebraic, then the first order conditions of $\overline{D}_{\mathcal{E}}$ are algebraic, if the sign vector of $u \in \partial\mathbf{U}_{\mathcal{N}}$ is known. The process of computing the sign vectors and solving the algebraic equations for each sign vector can be automatized. This strategy is applied to two difficult examples.

In addition, two classes of exponential families are treated: For one-dimensional families restrictions on the number and the possible support sets of maximizers are found. Partition exponential families have interesting properties:

Theorem 2. *If* $\max D_{\mathcal{E}} = \log(2)$*, then* $\dim(\mathcal{E}) \geq \lceil \frac{N}{2} \rceil - 1$*. Assume* $\dim(\mathcal{E}) = \lceil \frac{N}{2} \rceil - 1$*. If* N *is even or if* $\mathcal{E}$ *contains the uniform distribution, then* $\mathcal{E}$ *is a partition exponential family.*

Optimally approximating exponential families. $\mathcal{E}$ is *dimension* D*-optimal* if every exponential family $\mathcal{E}'$ of smaller dimension satisfies $\max D_{\mathcal{E}} \leq D < \max D(\cdot\|\mathcal{E}')$. Let

$$D_{N,k} = \min \left\{ \max D_{\mathcal{E}} : \mathcal{E} \text{ is an exponential family of dimension } k \text{ on } [N] \right\}.$$

Theorem 2 implies $D_{N,k} = \log(2)$ if and only if $\lceil \frac{N}{k+1} \rceil = 2$. A plausible conjecture is:

- $D_{N,k} = \log\lceil \frac{N}{k+1} \rceil$, and the dimension $D_{N,k}$-optimal exponential families containing the uniform distribution are partition exponential family.

The following is a first result in this direction:

Theorem 3. $D_{N,k} \geq \log(N/(k+1))$ *for all* $0 \leq k < N$*. If* $\mathcal{E}$ *has dimension* k *and satisfies* $\max D_{\mathcal{E}} = \log(N/(k+1))$*, then* $\mathcal{E}$ *is a partition model. In particular, if* N *is divisible by* $(k+1)$*, then* $D_{N,k} = \log(N/(k+1))$*, and the dimension* $D_{N,k}$*-optimal models are partition exponential families.*

References

[1] Ay, N., "An information-geometric approach to a theory of pragmatic structuring," *Annals of Probability*, vol. 30, pp. 416–436, 2002.

[2] ——, "Locality of global stochastic interaction in directed acyclic networks," *Neural Computation*, vol. 14, pp. 2959–2980, 2002.

[3] Linsker, R., "Self-organization in a perceptual network," *IEEE Computer*, vol. 21, pp. 105–117, 1988.

[4] Matúš, F., "Optimality conditions for maximizers of the information divergence from an exponential family," *Kybernetika*, vol. 43, no. 5, pp. 731–746, 2007.