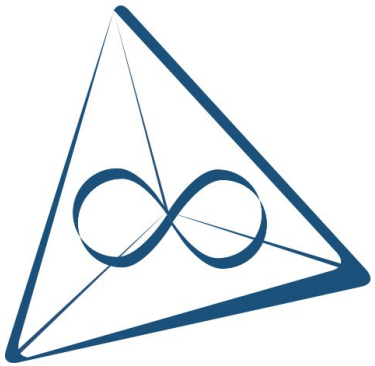
Maximising the Information Divergence from an Exponential Family

Johannes Rauh



Max Planck Institute
Mathematics in the Sciences

April 6, 2011
Universtity of Warwick

Outline

- Maximising the KL divergence
  - Exponential families
  - The projection property
  - Kernel distributions
  - Examples
- Application
  - Approximating exponential families

Contents

Definitions I: Exponential Families

Consider a finite set $\mathcal{X}$.

Definition

Let $a_1, \ldots, a_h \in \mathbb{R}^{\mathcal{X}}$ be real functions on $\mathcal{X}$. The family $\mathcal{E}$ of all probability distributions of the form

$$P_\theta(x) = \frac{1}{Z_\theta} \exp\left(\sum_i \theta_i a_i(x)\right), \qquad x \in \mathcal{X},$$

with θ a vector of h parameters,

Z_θ a normalising constant,

is an *exponential family*. The *normal space* of $\mathcal{E}$ is

$$\mathcal{N} = \left\{ u \in \mathbb{R}^{\mathcal{X}} : \sum_x u(x) a_i(x) = 0, \sum_x u(x) = 0 \right\}.$$

Definitions II: The information divergence

Let P, Q be two probability measures on $\mathcal{X}$.

Definition

The *information divergence* of P and Q is

$$D(P\|Q) = \sum_{x\in\mathcal{X}} P(x)\log\left(\frac{P(x)}{Q(x)}\right).$$

Here we set $0\log 0 = 0$.

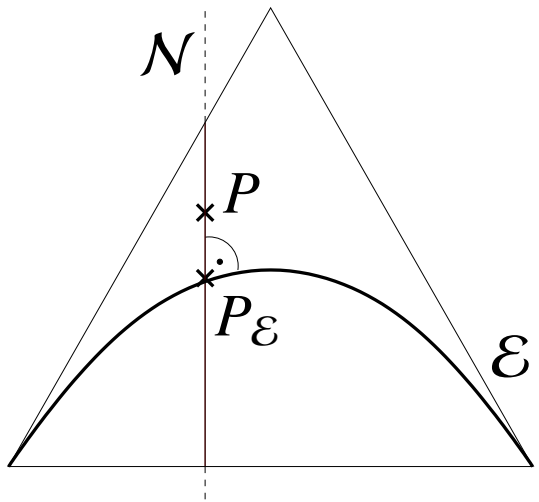
Properties of D:

- $D(P\|Q) \geq 0$ with $D(P\|Q) = 0 \iff P = Q$.
- However, D is not symmetric (and does not satisfy the triangular inequality).

Relations between D and $\mathcal{E}$

Let P be an arbitrary p.m.

- Then $P_{\mathcal{E}} := \operatorname{argmin}_{Q \in \mathcal{E}} D(P\|Q)$ is the *maximum likelihood estimate* (MLE).
- $P_{\mathcal{E}}$ is the *unique* p.m. in $\mathcal{E}$ satisfying $P - P_{\mathcal{E}} \in \mathcal{N}$.
- $P_{\mathcal{E}}$ maximises the entropy among all p.m.s Q such that $P - Q \in \mathcal{N}$.



Note: $P_{\mathcal{E}}$ might not exist. In this case replace $\mathcal{E}$ by its closure $\overline{\mathcal{E}}$.

The main problem

Problem (N. Ay 2002)

- Find the maximum of $D(P||\mathcal{E}) := \inf_{Q \in \mathcal{E}} D(P||Q)$!
- Find the maximising probability measures!

Motivation:

- For independence models: D equals the mutual information / multiinformation MI

  $\Longrightarrow$ Maximise MI as a strategy of learning in neural networks (infomax principle, Linsker '88)
- Interpret D as a complexity measure.

  $\Longrightarrow$ What are the most complex p.m.s?
- Find exponential families which "approximate all p.m.s."

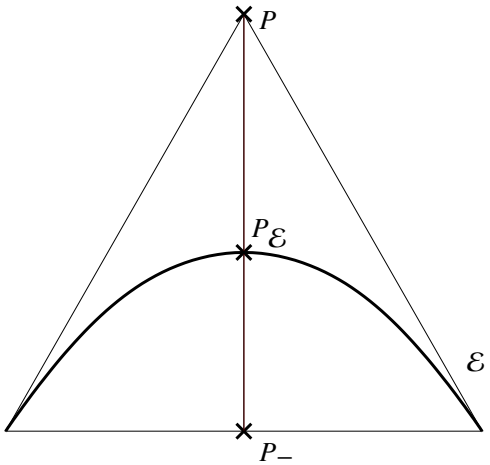
The projection property

Theorem (Matúš 2006)

A local maximiser P of $D(\cdot\|\mathcal{E})$ has the projection property:

$$P(x) = \frac{P_{\mathcal{E}}(x)}{P_{\mathcal{E}}(\mathrm{supp}(P))}, \text{ whenever } x \in \mathrm{supp}(P) = \{x : P(x) > 0\}.$$

Corollary: If $\mathrm{supp}(P) \neq \mathcal{X}$, there exists a p.m. P_- such that:



- $P_{\mathcal{E}}$ is a convex combination of P and P_-.
- $P \equiv P_- \equiv P_{\mathcal{E}} \mod \mathcal{N}$.
- $P_{\mathcal{E}}$ is also the MLE of P_-.
- $\mathrm{supp}(P) \cap \mathrm{supp}(P_-) = \emptyset$.
- P_- has the projection property.

Kernel distributions

Definition

A probability measure P is a *kernel distribution*, if there exists a probability measure Q such that

- $\operatorname{supp}(P) \cap \operatorname{supp}(Q) = \emptyset$,
- and $P - Q \in \mathcal{N}$.

Corollary

If P is a maximiser of $D(\cdot\|\mathcal{E})$, then P is a kernel distribution.

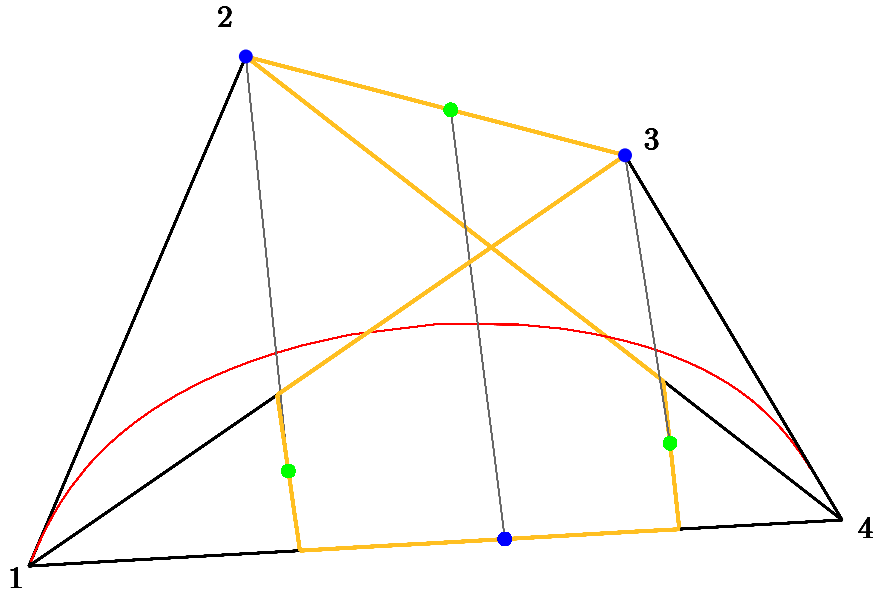
Idea:

Search for the maximisers among the kernel distributions.

Example in 3D

Let $a_1 = (0, 5, 12, 15)$.

- local maxima
- projection points
- kernel distributions
- exponential family



Question:
Is there a nice statistical model containing all kernel distributions in its boundary?

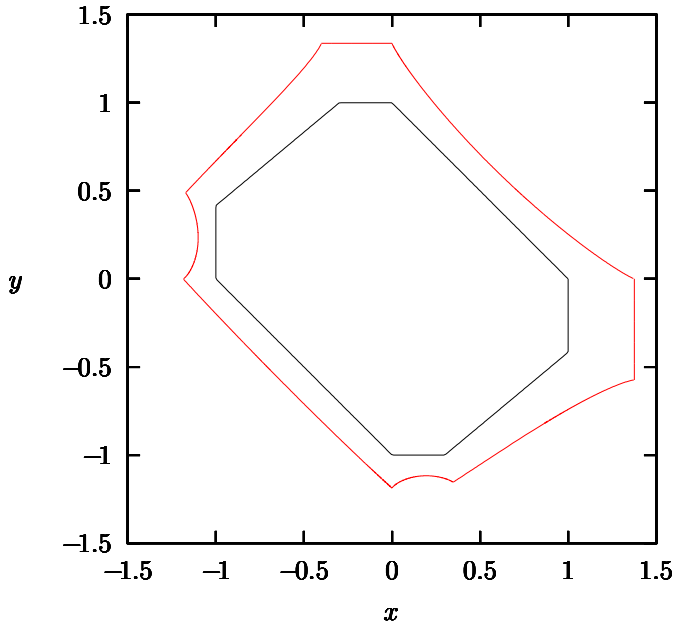
Kernel distributions in the kernel

How to find kernel distributions:

- Take $u \in \mathcal{N} \setminus \{0\}$.
- Decompose $u = u^+ - u^-$ into positive and negative part. Then $d_u := \sum_x u^+(x) = \sum_x u^-(x)$.
- $\frac{1}{d_u} u^+$ and $\frac{1}{d_u} u^-$ are kernel distributions.

The set $\{u \in \mathcal{N} : d_u = 1\}$ is the boundary of a polytope $\mathbf{U}_{\mathcal{N}} = \{u \in \mathcal{N} : d_u \leq 1\}$.

– indicates $D(P_+ \| \mathcal{E})$.



The second optimisation problem

Get rid of $P_{\mathcal{E}}$: If P_+ is a local maximiser, then

$$D(P_+||P_{\mathcal{E}}) = \log(1 + \exp(H(P_-) - H(P_+))).$$

Define a function $\overline{D} : \partial \mathbf{U}_{\mathcal{N}} \to \mathbb{R}$ via

$$\overline{D} : P^+ - P^- \mapsto H(P^-) - H(P^+)$$
$$u \mapsto \sum_{x \in \mathcal{X}} u(x) \log |u(x)|.$$

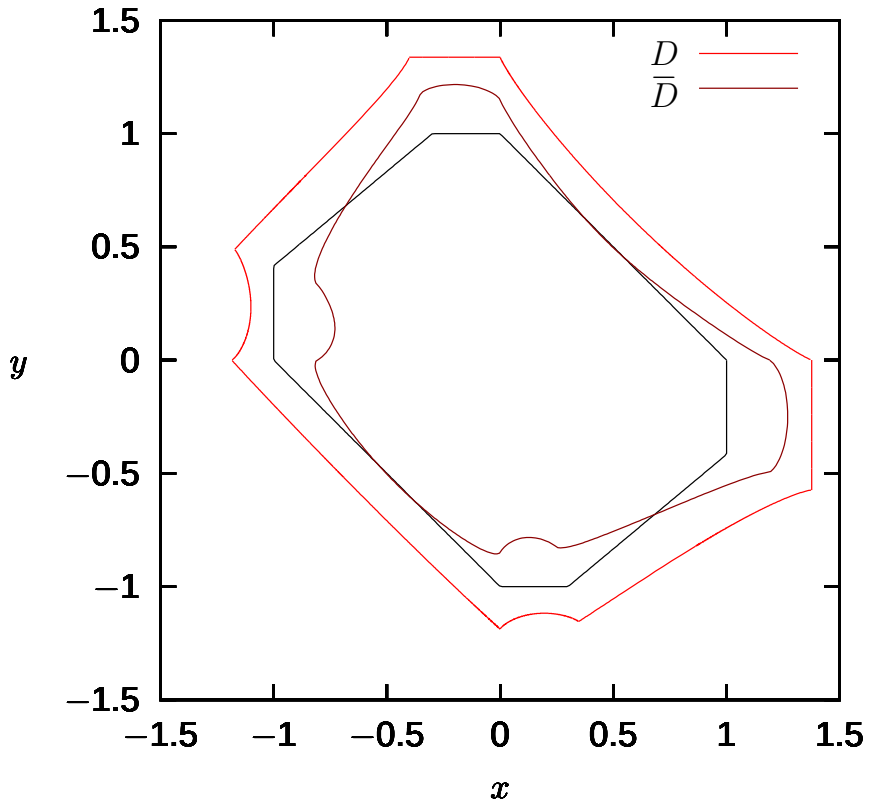
Theorem (Matúš, Rauh 2010)

- *If u is a local or global maximiser of $\overline{D}$, then P_+ is a local or global maximiser of $D(\cdot||\mathcal{E})$.*
- *If P is a local or global maximiser of $D(\cdot||\mathcal{E})$, then $P - P_-$ is a local or global maximiser of $\overline{D}$.*

D and $\overline{D}$ in the kernel

The set $\{u \in \mathcal{N} : d_u = 1\}$ is the boundary of a polytope $\mathbf{U}_{\mathcal{N}} = \{u \in \mathcal{N} : d_u \leq 1\}$.

– indicates $D(P_+ \| \mathcal{E})$.
– indicates $\overline{D}$.



Advantages of the $\overline{D}$-maximisation

The maximisation of $\overline{D}$ is easier than the maximisation of D in some aspects:

- **reduced dimensionality:**

The dimension of $\mathcal{N}$ is less than the dimension of the set of all p.m.s (which is $|\mathcal{X}| - 1$).

- **computation of the MLE is not necessary:**

$\overline{D}$ has a nice "almost analytic" formula. In contrast, to compute $D(P||\mathcal{E})$ we need to find $P_{\mathcal{E}}$.

Example (Codimension one)

Suppose $\mathcal{N}$ is spanned by one element $u = P^+ - P^-$. There are only two possibilities for the global maximiser. If $H(P^+) \leq H(P^-)$, then P^+ is the (or a) global maximiser.

$\Longrightarrow$ This example can be used to construct (counter)examples.

The general strategy

The function $\overline{D} : u \in \partial \mathbf{U}_{\mathcal{N}} \mapsto \sum_{x \in \mathcal{X}} u(x) \log |u(x)|$ is continuous, but not differentiable. However, $\overline{D}$ is smooth when the sign vector $\sigma_x := \operatorname{sgn}(u(x))$ is fixed.

Strategy:
Solve the critical equations of $\overline{D}$ for each sign vector occurring in $\mathcal{N}$.

Usually, there are many sign vectors (→*Oriented Matroids*).

Lemma

- $\sum_{x:\sigma=0} v(x) = 0$ *for all* $v \in \mathcal{N}$.
- $|\operatorname{supp}(P)| \leq \dim \mathcal{E} + 1$ *for all maximisers* P *of* $D(\cdot \| \mathcal{E})$.

Important for Automation:
If all a_i have only integer values, then the critical equations of $\overline{D}$ are algebraic (→*Computer Algebra Systems*)

Example: The 4-2 model

The *"pair interaction model of four binary random variables"* is an exponential family of dimension 10 with $\dim \mathcal{N} = 5$.

- There are 73 sign vectors up to symmetry.
- Only 20 of these satisfy the condition $\sum_{x:\sigma=0} v(x) = 0$.
- The critical equations can be solved with the help of SINGULAR.

Result

Up to symmetry, there is only one global maximiser. This confirms a conjecture by Thomas Kahle.

Example: The independence model 3-3-2

Let $\mathcal{E}$ be the independence model of two ternary and one binary random variables. The dimension is 5 and $\dim \mathcal{N} = 12$.

- There are 369 593 sign vectors up to symmetry.
- Only 975 sign vectors satisfy $\sum_{x:M=0} v(x) = 0$.
- Only 240 sign vectors satisfy $|\operatorname{supp}(\sigma^+)| \leq \dim \mathcal{E} + 1$.
- These 240 sign vectors can be treated with SINGULAR.

Result

Up to symmetry there is a unique global maximiser.

Contents

Approximating exponential families

The following problem is due to Yaroslav Bulatov:

Problem
Find a "small" exponential family $\mathcal{E}$ such that $\max D(\cdot\|\mathcal{E})$ is "small"!

Here, *small* can refer to

- the *dimension*,
- the *inclusion ordering*.

Motivation: Such families are good to approximate arbitrary distributions ($\Rightarrow$ learning).

More precise questions:

- Given $c > 0$, which minimal $\mathcal{E}$ satisfy $\max D(\cdot\|\mathcal{E}) \leq c$?
- Given $\mathcal{E}$, does there exist a smaller $\mathcal{E}'$ such that $\max D(\cdot\|\mathcal{E}') \leq \max D(\cdot\|\mathcal{E})$?

Partition models

Let $\mathcal{X}' = \{\mathcal{X}^1, \ldots, \mathcal{X}^l\}$ be a partition of $\mathcal{X}$.

Definition

The *partition model* is the exponential family / linear model

$$\mathcal{E}_{\mathcal{X}'} := \left\{P : P(x) = P(y) \text{ whenever } \{x, y\} \subseteq \mathcal{X}^i \text{ for some } i\right\}.$$

Lemma: $\max D(\cdot\|\mathcal{E}_{\mathcal{X}'}) = \log \max_i(|\mathcal{X}^i|)$

Proposition

If $|\mathcal{X}^i| = c$ for all i, then any $\mathcal{E} \subset \mathcal{E}_{\mathcal{X}'}$ satisfies $\max D(\cdot\|\mathcal{E}_{\mathcal{X}'}) > \log(c)$.

Proof: Let $P \in \mathcal{E}_{\mathcal{X}'} \setminus \mathcal{E}$. Find Q such that $\log(c) = D(Q\|P) = D(Q\|\mathcal{E}_{\mathcal{X}'})$, then $D(Q\|\mathcal{E}) > \log(c)$. □

First results

Proposition

If $\mathcal{E}$ has dimension k, then $\max D(\cdot\|\mathcal{E}) \geq \log(N/(k+1))$. If $\max D(\cdot\|\mathcal{E}) = \log(N/(k+1))$, then $k+1$ divides N, and $\mathcal{E}$ is a partition model of a homogeneous partition of coarseness $N/(k+1)$.

Conjecture

The above results generalise as follows: If $k+1$ does not divide N, then $\max D(\cdot\|\mathcal{E}) \geq \log\lceil N/(k+1)\rceil$, and if equality holds, then $\mathcal{E}$ is a partition model.

The conjecture holds true if $\lceil N/(k+1)\rceil = 2$.

Conclusion

Let $\mathcal{E}$ be an exponential family with normal space $\mathcal{N}$.

- The maximisers of $D(\cdot\|\mathcal{E})$ have interesting properties:
  - They have the *projection property*.
  - They are *kernel distributions*.
- Maximising $D(\cdot\|\mathcal{E})$ is "equivalent" to the maximisation of

$$\overline{D}(u) = \sum_x u(x) \log |u(x)| \qquad (u \in \partial \mathbf{U}_{\mathcal{N}}).$$

  - Reduction of dimensionality.
  - Nice function to maximise.
- Application:
  - Finding "small" $\mathcal{E}$ such that $\max D(\cdot\|\mathcal{E})$ is "small".

Acknowledgements/References

Many thanks to:

- Nihat Ay, Thomas Kahle, Bastian Steudel (MPI MIS Leipzig),
- Fero Matúš (ÚTIA Prague)

References:


N. Ay
The Annals of Probability 30 (2002), 416–436.
Neural Computation 14 (2002), 2959–2980


F. Matúš
Kybernetika 43 (2007), 731-746.
IEEE TIT 55, No. 12 (2009), 5375–5381
...


J. Rauh
Finding the maximizers of the information divergence...
accepted to IEEE TIT, available at arXiv:0912.4660