A Minimax Result for the Kullback Leibler Bayes Risk

Jürgen Krob[*]
Department of Mathematics
University of Kaiserslautern
D-67663 Kaiserslautern[†]
Germany

Holger R. Scholl
Department of Mathematics
University of Kaiserslautern
D-67663 Kaiserslautern
Germany

e–Mail: juergen.krob@bertelsmann.de

e–Mail: holger@mathematik.uni-kl.de

April 23, 1997

Abstract

It is of basic interest to assess the quality of the decisions of a statistician, based on the outcoming data of a statistical experiment, in the context of a given model class $\{P_\theta : \theta \in \Theta\}$ of probability distributions. The statistician picks a particular distribution P, suffering a loss by not picking the 'true' distribution P_{θ^*}. There are several relevant loss functions, one being based on the the relative entropy function or Kullback Leibler information distance. In this paper we prove a general 'minimax risk equals maximin (Bayes) risk' theorem for the Kullback Leibler loss under the hypothesis of a dominated and compact family of distributions over a Polish observation space with suitably integrable densities. We also find that there is always an optimal Bayes strategy (i.e. a suitable prior) achieving the minimax value. Further, we see that every such minimax optimal strategy leads to the same distribution P^* in the convex closure of the model class. Finally, we give some examples to illustrate the results and to indicate, how the minimax result reflects in the structure of least favorable priors.

[*] This paper is mainly based on parts of this author's doctorial thesis.

[†] This author is now at: Bertelsmann Club GmbH, Postbox 1109, D-33339 Rheda–Wiedenbrueck

1 Introduction

There are several means of measuring the divergence of the decision of a statistician from the truth provided by nature. Mathematically, this divergence is usually modelled by a loss function. There are several important examples, maybe the most prominent in statistical estimation theory being the square–distance function and the Kullback Leibler information quotient. The statistician has to choose a particular one fitting into his given context. Then, the expected loss under the true distribution is the risk the statistician is taking.

In many cases, the Kullback Leibler information distance (KL–distance) is an appropriate choice of a risk functional. It was first introduced by S. Kullback and R.A. Leibler in 1951, see [KL51], as a directed measure of the distance of two probability distributions. Let P and Q be distributions over a measurable space X with densities f and g with respect to a σ–finite measure μ. Then

$$\mathbb{K}(P,Q) := \int_X f(x)\log\frac{f(x)}{g(x)}\mu(dx) = \mathbb{E}_P(\frac{f(x)}{g(x)}).$$

With the convention $0\log 0 := 0$ we have that $\mathbb{K} \geq 0$ with $\mathbb{K}(P,Q) = 0$ iff $P = Q$. The KL–distance has several interpretations in information theory and statistics, here it will be used as a risk functional in a parameter estimation context.

In 1956 D.V. Lindley adapted the concept of the information theoretic transmission or information rate to the theory of statistical experiments, see [Lin56]. An experiment, i.e. a family of probability distributions $\{P_\theta : \theta \in \Theta\}$ over an observation space X, is considered as a Shannon information transmitting channel, the true parameter being the unknown character sent, and the data being the characters observed after some trials. A prior's transmission or information rate measures the expected information gain of the experiment given a prior. It is quite natural to choose a prior maximizing this information rate. This choice leads to non subjectivistic priors that contain as little as possible information about the parameter relative to the information provided by the data. In information theory, and we will adapt this here, the maximum information rate of a channel is often called the capacity, and those priors achieving capacity are called optimal. The capacity of a statistical experiment is closely linked to the grade of 'orthogonality' inherent in the given family. We will learn much more about the true parameter if the distributions have largely disjoint areas of support. In this case the capacity will also be large, and vice versa. If the distributions are very similar then we will have a small information gain, regardless which prior we choose. Thus, the capacity will be small. So, the concept of capacity is a useful tool to compare the 'orthogonality' of different experiments, as it doesn't depend on the choice of priors, spaces or dimensionality. But of course, in many practical situations the calculation of the capacity of a given experiment as well as the determination of optimal priors can only be done numerically.

We can characterize optimal priors in terms of a Bayes strategy: If we use the Kullback Leibler information quotient as loss function, then the Shannon information rate $\mathcal{I}$ can be seen to be the minimal Bayes risk for the statistician's picked distribution P under a chosen prior:

$$\mathcal{I}(\lambda) = \min_P \int_\Theta I\!K(P_\theta, P)\lambda(d\theta).$$

After rigorously introducing the mentioned quantities, we will show in the next section of this paper, that $P := P_\lambda := \int_\Theta P_\theta \lambda(d\theta)$, the mixture distribution under the prior, minimizes the Bayes risk, so that

$$\mathcal{I}(\lambda) = \int_\Theta I\!K(P_\theta, P_\lambda)\lambda(d\theta).$$

We will also give conditions on the model family $\mathcal{P} := \{P_\theta : \theta \in \Theta\}$, which guarantee that the information rate is continuous. If we finally assume $\mathcal{P}$ as being compact, then we always find a convex set of optimal priors achieving capacity. It turns out that all optimal priors lead to the same mixture distribution. This mixture distribution minimizes the maximal value of the Kullback Leibler risk funtional, thus leading to a minimax equals maximin equals capacity characterization. The mentioned results mainly base on parts of J. Krob's doctorial thesis, which was done in 1992 at the Department of Mathematics of the University of Kaiserslautern, Germany, under the supervision of H. v. Weizsäcker, see [Kro92]. Finally, the last section of this paper gives some numerical examples to illustrate the results.

2 The Capacity of a Statistical Experiment

The first part of this section will introduce the already mentioned quantities and objects on a rigorous basis, in order to switch from a general conceptual framework to a mathematical theory.

With Prob$X$ or ProbΘ we will respectively denote the space of all probability measures over the (measurable) spaces $(X, \mathcal{X})$, and $(\Theta, \mathcal{B})$.

Definition 1: *A* **statistical experiment** *is a triple $((X, \mathcal{X}), \mathcal{P}, \Theta)$ consisting of a measurable space $(X, \mathcal{X})$ which we will additionally assume to be Polish, and a family $\mathcal{P} := (P_\theta)_{\theta\in\Theta}$ of probability distributions over $(X, \mathcal{X})$ with parameter space $\Theta \subseteq I\!R^d, d \in I\!N$.*
If additionally $\mathcal{P}$ is compact in the topology of the variational distance, and Θ is compact in the usual topology on $I\!R^d$, then the experiment is called **compact***.*
If all $P_\theta, \theta \in \Theta$, are dominated by a σ–finite measure $\mu \in$ Prob$X$ then there exist μ–densities f_θ for all $\theta \in \Theta$. In this situation the experiment is called **dominated***. Let $f_\varphi := \int_\Theta f_\theta \varphi(d\theta)$ denote the density of the mixture distribution with respect to the prior φ. If the family*

$$(f_\lambda \log f_\lambda)_{\lambda\in\text{Prob}\Theta}$$

is uniformly μ–integrable then we will say that the experiment is **uniformly integrable***.*

In most cases, just $\mathcal{P}$ will be called experiment, as an abbreviation. For instance, let $\mathcal{P}$ be any d–dimensional standard exponential family, like the Normal family, the Gamma family, the Binomial family or the Poisson family for suitable $d \in I\!\!N$. If Θ is a compact subset of the natural parameter space, then $\mathcal{P}$ is a dominated, compact and uniformly integrable experiment in the sense of definition 1.

Lindley defined the **average amount of information provided by the experiment with prior knowledge** $\varphi \in \mathrm{Prob}\Theta$ to be

$$\mathcal{I}(\varphi) := I\!\!E_x(H(\varphi) - H(\varphi(\cdot|x))), \tag{1}$$

whenever the entropy of the prior and the posterior distribution, i.e. integrals of the form

$$H(\varphi) := \int_\Theta \varphi(\theta) \log \varphi(\theta) d\theta,$$

exist, both distributions φ and $\varphi(\cdot|x)$ having a Lebesgue–density. Thus, the mean loss of entropy by observing the data gives the amount of information we gain about the parameter, being the quantity of interest. We will use another definition, putting all integrability assumptions into the experiment rather than into the prior distributions. This seems to be useful here as the statistician then has full access to all kinds of priors. Following the concept of least favourable or non informative priors this degree of freedom is desirable as examples show that in many typical cases non informative priors are discrete (see the examples at the end of this paper, and [Ber89] and [Zha94]) on one hand, whereas Bernardo's reference priors, e.g. Jeffreys' prior, typically have Lebesgue densities (see [Ber79] and [BS93]).

Definition 2: *For a compact, μ–dominated and uniformly integrable experiment* $((X, \mathcal{X}), \mathcal{P}, \Theta)$ *with* $\mathcal{P} := \{P_\theta : \theta \in \Theta\}$ *having densities* $\{f_\theta : \theta \in \Theta\}$ *the quantity*

$$\begin{aligned} \mathcal{I}(\varphi) &:= -\int_{\mathrm{X}} p_\varphi(x) \log p_\varphi(x) \mu(dx) + \int_\Theta \int_{\mathrm{X}} f_\theta(x) \log f_\theta(x) \mu(dx) \varphi(d\theta) \\ &= H(P_\varphi) - \int_\Theta H(P_\theta) \varphi(d\theta) \end{aligned}$$

defines the **information rate** *or* **average amount of information** *provided by the prior distribution* $\varphi \in \mathrm{Prob}\Theta$.

Again, $P_\varphi := \int_\theta P_\theta \varphi(d\theta)$ is the mixture distribution with respect to φ, having the μ–density f_φ. If the given experiment is uniformly integrable we are guarantueed that $\mathcal{I}$ is well defined and continuous[1]. Since Θ and thus $\mathrm{Prob}\Theta$ are compact there is a set of priors (maybe just containing one element) which achieve the maximum information rate. These priors are called **optimal** and their information rate is called the **capacity** of the experiment:

$$\mathcal{C} := \max_{\varphi \in \mathrm{Prob}\Theta} \mathcal{I}(\varphi).$$

The following lemma gives some identities, whose classical analoga are well known in information theory.

[1] Examples show that the integrability assumption cannot be dropped easily.

Lemma 1: *Let $\mathcal{P}$ be a compact, dominated and uniformly integrable experiment. Then the following holds for any $\varphi \in \mathrm{Prob}\Theta$:*

1. $\mathcal{I}(\varphi) = \int_\Theta \mathbb{K}(P_\theta, P_\varphi)\varphi(d\theta)$.
2. $\mathcal{I}(\varphi) = \int_\Theta \mathbb{K}(P_\theta, P_\lambda)\varphi(d\theta) - \mathbb{K}(P_\varphi, P_\lambda)$ *for any* $\lambda \in \mathrm{Prob}\Theta$.
3. *Let* $\varphi^{(1)}, \ldots, \varphi^{(N)} \in \mathrm{Prob}\Theta$ *with* $N \in \mathbb{N}$ *and let* $s := (s_1, \ldots, s_N) \in S_{N-1}$ *be a probability vector in the* $N-1$*–dimensional unit simplex. Let* $\varphi^{(0)} := \sum_{k=1}^N s_k \varphi^{(k)}$. *Then for every* $\lambda \in \mathrm{Prob}\Theta$*:*

$$\sum_{k=1}^N s_k \mathcal{I}(\varphi^{(k)}) + \sum_{k=1}^N s_k \mathbb{K}(P_{\varphi^{(k)}}, P_\lambda) = \mathcal{I}(\varphi^{(0)}) + \mathbb{K}(P_{\varphi^{(0)}}, P_\lambda).$$

As the Kullback Leibler distance may be interpreted as the information gain when we switch from P_φ to P_θ, the first statement characterizes $\mathcal{I}$ as mean information gain given a prior. From a Bayesian point of view the statement says that $\mathcal{I}$ is the Bayes risk of the prior distribution φ. The second statement tells us then that $\mathcal{I}$ even is the minimal Bayes risk, taking $P_\lambda := P_\varphi$, in this situation.

Proof: By our assumption of uniform integrability we find that

$$\begin{aligned}
\mathcal{I}(\varphi) &= -\int_{\mathrm{X}} p_\varphi(x)\log p_\varphi(x)\mu(dx) + \int_\Theta \int_X f_\theta(x)\log f_\theta(x)\mu(dx)\varphi(d\theta) \\
&= \int_\Theta (\int_X (f_\theta(x)\log f_\theta(x) - f_\theta(x)\log f_\varphi(x))\mu(dx))\varphi(d\theta) \\
&= \int_\Theta \mathbb{K}(P_\theta, P_\varphi)\varphi(d\theta),
\end{aligned}$$

which proves the first statement. The second equality will be proved in two steps. First, we assume that $\mathbb{K}(P_\varphi, P_\lambda) = \infty$ for a prior $\lambda \in \mathrm{Prob}\Theta$. This means that we have for the densities

$$0 = f_\lambda(x) < f_\varphi(x)$$

on a set of positive μ–measure. For these $x \in X$ we also have $0 < f_\theta(x)$, and thus also $\mathbb{K}(P_\theta, P_\lambda) = \infty$ on a set of parameters of positive φ–measure. So

$$\int_\Theta \mathbb{K}(P_\theta, P_\lambda)\varphi(d\theta) = \infty.$$

For the second step let $\mathbb{K}(P_\varphi, P_\lambda) < \infty$. We have

$$\begin{aligned}
\mathcal{I}(\varphi) + \mathbb{K}(P_\varphi, P_\lambda) &= \int_\Theta \mathbb{K}(P_\theta, P_\varphi)\varphi(d\theta) + \int_\Theta \mathbb{K}(P_\varphi, P_\lambda)\varphi(d_\lambda) \\
&= \int_\Theta (\int_X f_\theta(x)(\log\frac{f_\theta(x)}{f_\varphi(x)} + \log\frac{f_\varphi(x)}{f_\lambda(x)})\mu(dx))\varphi(d\theta) \\
&= \int_\Theta \mathbb{K}(P_\theta, P_\lambda)\varphi(d\theta),
\end{aligned}$$

prooving the second assertion. The third statement is just 'a convex combination' of statements 1 and 2, thus the lemma is proved. ■

The next lemma's classical analogon is a core result for computing the channel capacity. It is due to Shannon (1948) and to Eisenberg and Gallager (1962), see [Gal68]. The early proofs of the classical version involved Kuhn–Tucker criteria which are hardly to use here. So we adapt an elementary proof given by F. Topsøe in [Top74].

Lemma 2: *Let $\mathcal{P}$ be a compact, dominated and uniformly integrable experiment with capacity $\mathcal{C}$. Then the following holds:*
If $\varphi \in \mathrm{Prob}\Theta$ is a prior with corresponding mixture distribution $P_\varphi \in \mathrm{Prob}X$, then the following condition is sufficient and necessary for φ being optimal:
There is a constant $C < \infty$ so that

1. *$\mathbb{K}(P_\theta, P_\varphi) = C$ for φ-almost all $\theta \in \Theta$,*
2. *$\mathbb{K}(P_\theta, P_\varphi) \leq C$ for all $\theta \in \Theta$.*

If the condition holds then $C = \mathcal{I}(\varphi) = \mathcal{C}$.

Proof: The 'sufficient' part follows directly by the previous lemma 1. For the proof of the 'necessary' part, we first note, that if condition (2) holds for an optimal prior $\varphi \in \mathrm{Prob}\Theta$, then also condition (1) is valid. Next, we choose $C := \mathcal{I}(\varphi) = \mathcal{C}$. If there was a $\theta_0 \in \Theta$ with $\mathbb{K}(P_{\theta_0}, P_\varphi) > C$, then it would be possible to increase the information rate by sending θ_0 with a slightly higher rate. Modifying φ by

$$\varphi_t := t \cdot \delta_{\theta_0} + (1-t) \cdot \varphi,$$

with $t \in (0,1)$ and applying lemma 1 leads to

$$\mathcal{I}(\varphi_t) = t\mathcal{I}(\delta_{\theta_0}) + (1-t)\mathcal{C} + t\mathbb{K}(P_{\theta_0}, P_{\varphi_t}) + (1-t)\mathbb{K}(P_\varphi, P_{\varphi_t}).$$

Thus we have the inequality

$$\mathcal{I}(\varphi_t) \geq (1-t)\mathcal{C} + t\mathbb{K}(P_{\theta_0}, P_{\varphi_t}).$$

As $t \mapsto \mathbb{K}(P_{\theta_0}, P_{\varphi_t})$ is a continuous function, we find a $0 < t^* < 1$ such that $\mathbb{K}(P_\varphi, P_{\varphi_{t^*}}) > \mathcal{C}$, and thus

$$\begin{aligned} \mathcal{I}(\varphi_{t^*}) &\geq (1-t^*)\mathcal{C} + t^*\mathbb{K}(P_{\theta_0}, P_{\varphi_{t^*}}) \\ &> \mathcal{C}. \end{aligned}$$

But this is impossible as $\mathcal{I}(\varphi) = C = \mathcal{C} = \max_{\lambda \in \mathrm{Prob}\Theta} \mathcal{I}(\lambda)$ is maximal. Hence, it must be $\mathbb{K}(P_{\theta_0}, P_\varphi) \leq C$ for all $\theta_0 \in \Theta$. Therefore the conditions hold with $C = \mathcal{C}$. ■

Interpreting the lemma geometrically, we can say the following: $\mathcal{C}$ is the radius of the smallest Kullback Leibler 'circle' around the mixture distribution induced by optimal priors, such that all elements P_θ of the experiment are inside or on the circle. In Bayesian terms we can say that if we choose an optimal prior φ and take P_φ as our guess of nature's behaviour, we will never suffer a risk higher than $\mathcal{C}$, not depending on the true θ_0, with $\mathcal{C}$ being the smallest value with this attribute. In the next section we will provide some examples illustrating the lemma. These interpretations motivate the following theorem, which is the main result of this paper.

Theorem 1: Minimax–Theorem
Let $\mathcal{P}$ be a compact, dominated and uniformly integrable experiment.

1. *Each Shannon optimal prior $\varphi \in \mathrm{Prob}\Theta$ maximizes the minimal Bayes risk. The set of all optimal $\varphi \in \mathrm{Prob}\Theta$ is a non empty convex subset of $\mathrm{Prob}\Theta$.*
2. *There is a unique distribution $P \in \mathrm{convexhull}(\mathcal{P}) \subseteq \mathrm{Prob}X$, which minimizes the maximal value of the Bayes risk function. It is $P = P_\varphi$ for any optimal prior $\varphi \in \mathrm{Prob}\Theta$.*
3. *If $\mathcal{C}$ is the capacity of the experiment, then*

$$\begin{aligned}\mathcal{C} &= \min_{\varphi\in\mathrm{Prob}\Theta} \max_{\theta\in\Theta} I\!K(P_\theta, P_\varphi) \\ &= \max_{\varphi\in\mathrm{Prob}\Theta} \min_{\lambda\in\mathrm{Prob}\Theta} \int_\Theta I\!K(P_\theta, P_\lambda)\ \varphi(d\theta).\end{aligned}$$

Proof: By definition, a Shannon optimal prior φ maximizes the information rate $\mathcal{I}(\varphi)$ of an experiment, which is, by lemma 1, the corresponding minimal Bayes risk. The minimax value equals the capacity of the experiment by lemma 2.

Item (3) of lemma 1 implies the strict concavity of $\mathcal{I}$: Using the same notation as above we see, that the Kullback Leibler terms vanish if and only if $P_{\varphi^{(0)}} = P_\lambda = P_{\varphi^{(k)}}$ for all k with $s_k > 0$. Thus, the mixtures $P_{\varphi^{(1)}}$ and $P_{\varphi^{(2)}}$ of two optimal priors $\varphi^{(1)}$ and $\varphi^{(2)}$ must be the same. The concavity of the information rate/minimal Bayes risk also implies that the set of all priors maximizing it, is convex. Note that the continuity of $I\!K$ in its first variable is implied by our condition of uniform integrability. Then, lemma 2 implies that the maximum value of the Bayes risk is minimized by the (unique) mixture P_φ of optimal priors. ■

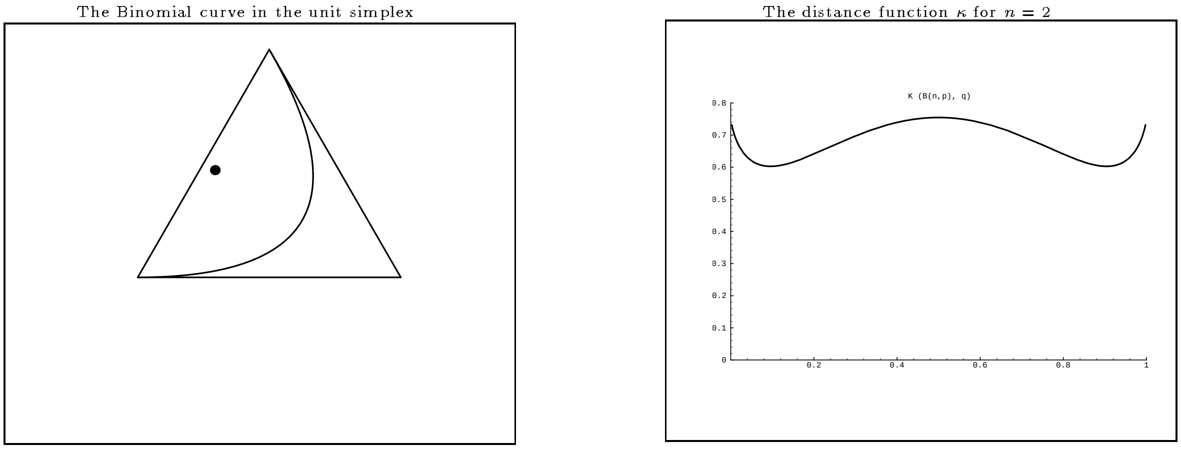
This means that $\mathcal{C}$ is both minimax and maximin risk value for a Bayes strategy with $I\!K$ as risk function. In other words: Optimal priors are least favorable under Kullback Leibler risk. For a related minimax result, see also D. Haussler's [Hau95]. It is interesting to relate our statement with an asymptotic analogon: In 1994 B. Clarke and A. Barron proved that Jeffreys' prior, which is Bernardo's reference prior in this situation, is asymptotically least favorable under Kullback Leibler risk, see [CB94]. Both results are linked together by a theorem proved in [Sch97], which shows that any sequence of Shannon optimal priors converges weakly to Jeffreys' prior. Thus, for large sample sizes, Shannon optimal priors have a very similar 'mass–distribution' like Jeffreys' prior, approximately revealing the same invariance properties. On the other hand, Jeffreys' prior can be used as a ready at hand approximation of finite sample priors, which would be hard to derive exactly.

3 Examples

We will illustrate our result in this section by the following examples: Given an experiment, we calculate (numerically) an optimal prior φ by the algorithm of Arimoto and Blahut. This algorithm was introduced independently by the two authors in 1972 (see [Ari72] and [Bla72]). It also leads us to the (approximative) optimal mixture distribution P_φ. To get a numerically treatable family, we take a discretization of the parameter interval with $t = 33$ points. Then, we plot the distance function

$$\kappa : \Theta \to [0, \infty], \theta \mapsto I\!\!K(P_\theta, P_\varphi).$$

This function tells us which loss we suffer by using P_φ instead of the 'true' P_θ. The maxima are (modulo numerical noise) all of the same value, which is the capacity of the experiment. The maximum points of this function correspond to the points of support of optimal priors by lemma 2. Any optimal prior is supported on the set of the maximum points of κ.


The Binomial curve in the unit simplex

The distance function κ for $n = 2$

Figure 1: The Binomial family for 2 outcomes.

The first diagram displays the Binomial family $(\mathcal{B}(n,p))_{p\in[0,1]}$ as a curve in the two–dimensional unit simplex. The big blop in the convex hull of the curve represents the calculated mixture distribution P_φ. This distribution is the centre of the minimax Kullback Leibler 'circle' with radius $\mathcal{C}$. All family distributions are inside this circle. Further, it is apparent in the present example that an optimal prior must be of discrete support, as κ has only isolated maximum points.

The next two diagrams show the analogue situation for a truncated Poisson family. In the example, the Poisson sequence with parameter λ is truncated after the third entry, which is replaced by $1 - \sum_{k=0}^{1} \mathcal{P}(\lambda)(k)$, to get a three–dimensional probability vector. In the first picture the 'inner' curve displays the resulting probability family for $\lambda \in [0, 10]$. The other curve is the Binomial curve. Again, the maximum points give the capacity and the points of support of optimal priors.

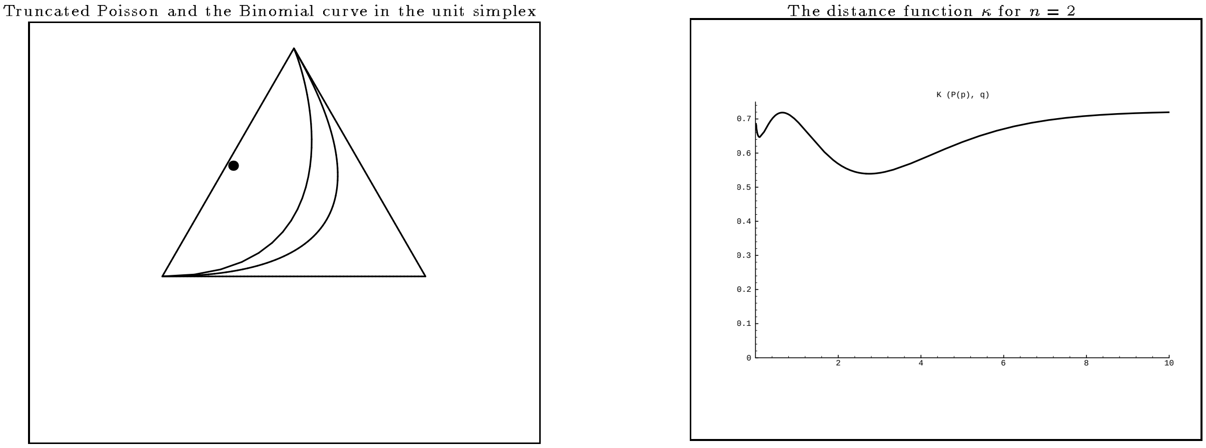


Figure 2: Truncated Poisson family and the Binomial family.

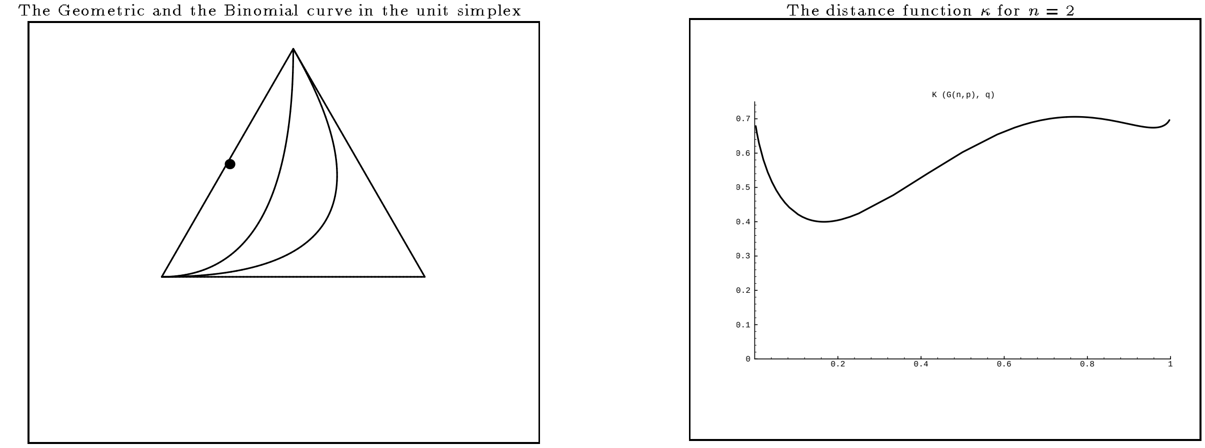
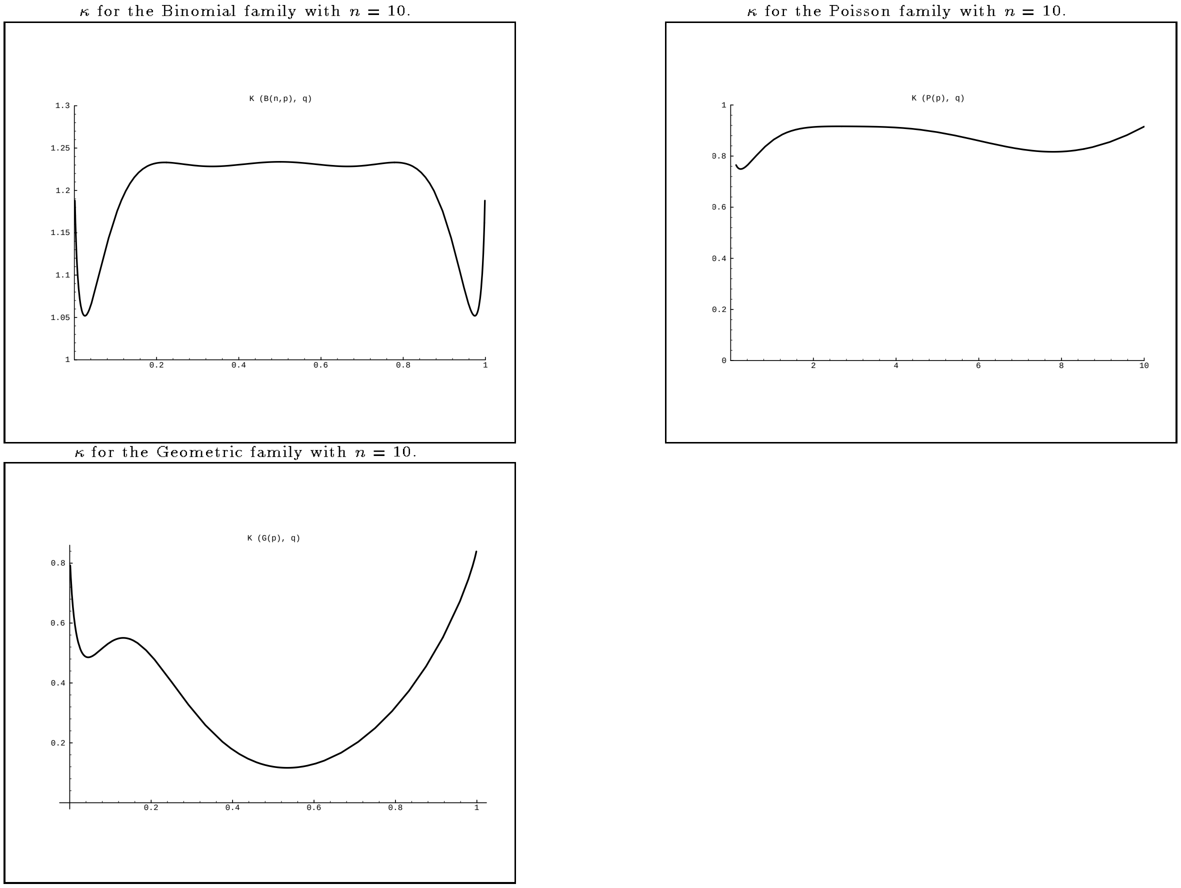


Figure 3: Truncated Geometric and the Binomial family.

The third figure shows the Geometric distribution family, analogously truncated, together with the Binomial curve, and the distance function κ. Again, all distributions of the family have a Kullback Leibler distance to the optimal mixture distribution equal to or smaller than the capacity of the experiment.

The last figure shows κ for the three example families, now for $n = 10$. Increasing the sample size (and thus the dimensionality) just increases the number of isolated maximum points.

The examples show that, due to the minimax characterization of the capacity, the structure of finite sample least favorable priors can be studied, at least numerically. In the above examples, all least favorable priors have a discrete structure, which is apparent from the fact, that the distance function κ has only isolated maxima. The methods based on information theory presented in this paper also allow to investigate the 'orthogonality' structure of a statistical experiment. As already indicated in the introduction, capacity, and of course also the structure of least favorable priors, are closely connected to the diversity of the



κ for the Binomial family with $n = 10$.

κ for the Poisson family with $n = 10$.

κ for the Geometric family with $n = 10$.

Figure 4: The example families' κ for $n = 10$.

model family. Optimal priors favour those parameters $\theta_1, \theta_2, \ldots$ with P_{θ_i} being as orthogonal as possible to all other $P_{\theta_j}, i \neq j$. Further, the induced optimal mixture distribution P_φ represents the 'information theoretic centre' of the family's convex hull. Families with distributions near to the extreme points of the simplex have a convex hull which is large, measured by the Kullback Leibler distance from this 'centre'. Thus, a statistically more informative structure reflects in larger value of the experiment's capacity. This can be observed in the presented examples: the Binomial family has the largest capacity, followed by the (truncated) Poisson and the Geometric family.

References

[Ari72] S. Arimoto. An Algorithm for Computing the Capacity of Arbitrary Discrete Memoryless Channels. *IEEE Trans. Inform. Theory.*, 18:14–20, 1972.

[Ber79] J.M. Bernardo. Reference Posterior Distributions for Bayesian Inference. *J. R. Statist. Soc., Ser. B*, 41:113–147, 1979.

[Ber89] Berger, J.O., Bernardo, J.M. and Mendoza, M. On priors that maximize expected information. *In: Recent Developments in Statistics and their Applications*, 1989. J.P. Klein and J.C. Lee, eds.

[Bla72] R.E. Blahut. Computation of Channel Capacity and Rate-Distortion Functions. *IEEE Trans. Inform. Theory*, 18:460–473, 1972.

[BS93] J.M. Bernardo and A.F.M. Smith. *Bayesian Theory.* Wiley Series in Probability and Mathematical Statistics, Chichester, 1993.

[CB94] B.S. Clarke and A.R. Barron. Jeffreys prior is asymptotically least favorable under entropy risk. *J. of Stat. Planning and Inference*, 41:37–60, 1994.

[Gal68] R.G. Gallager. *Information theory and reliable communication.* Wiley, New York, 1968.

[Hau95] D. Haussler. A general minimax result for relative entropy. *Preprint, University of California at Santa Cruz*, 1995.

[KL51] S. Kullback and R.A. Leibler. On information and sufficiency. *Ann. Math. Statist.*, 22:79–86, 1951.

[Kro92] J. Krob. *Kapazität statistischer Experimente.* PhD thesis, Univ. Kaiserslautern, 1992.

[Lin56] D.V. Lindley. On a Measure of the Information provided by an Experiment. *Ann. Math. Stat.*, 27:986–1005, 1956.

[Sch97] H. R. Scholl. Shannon optimal priors on iid statistical priors converge weakly to Jeffreys' prior . *Preprint*, April 1997.

[Top74] F. Topsøe. *Informationstheorie.* B.G. Teubner, Stuttgart, 1974.

[Zha94] Z. Zhang. *Discrete noninformative priors.* PhD thesis, Department of Statistics, Yale University, New Haven, 1994.