Advances in Bayesian analysis and its applications to sciences
–draft contents of thesis–

Ferenc Huszár

July 23, 2012

Part I

Introduction

Part II

Information geometry of probability distributions

Chapter 1

An introduction to scoring rules

In this section I describe scoring rules that can be used to assess the performance of probabilistic forecasting models. The scoring rule framework allows us to define useful generalisations of well-known information quantities, such as entropy, mutual information and divergence. Based on this, scoring rules allow for defining rich geometries of probabilistic models, which can be exploited in a variety of statistical applications, such as parameter estimation, approximate inference and optimal experiment design.

Imagine we want to have build a probabilistic forecaster that predicts the value of a random quantity X. We can describe any such probabilistic forecaster as a probability distribution $P(x)$ over the space of possible outcomes $\mathcal{X}$. After observing the outcome $X = x$ we want to assess how good our predictions were: *scoring rule* is a general term to describe any function that quantifies this: if the outcome is $X = x$, and our prediction was P we incur a score $S(x, P)$. Scoring rules, by convention, are interpreted as losses, so lower values are better. A good example of scoring rules is the logarithmic score, or simply the log score: $S_{\log}(x, P) = -\log P(x)$, which is used in maximum likelihood estimation. It is certainly a very important scoring rule and has several unique features (see section **??**), but it is not the only one. I will give further examples of scoring rules in section **??**. Mathematically, a scoring rule is any measurable function that maps an outcome-probability distribution pair onto real numbers: $S : \mathcal{X} \times \mathcal{M}^1_{\mathcal{X}} \mapsto \mathcal{R} \cup \{\infty\}$.

1.1 Information quantities

A scoring rule allows us to define the following, useful information quantities [**?**, see also]]Blaetal2332.

Definition 1 (Generalised entropy)**.** *Given a scoring rule $S : \mathcal{X} \times \mathcal{M}^1_{\mathcal{X}} \mapsto \mathbb{R}$, let us define the generalised entropy of a distribution $P \in \mathcal{M}^1_{\mathcal{X}}$ as follows:*

$$\mathbb{H}_S[P] = \mathbb{E}_{x\sim P} S(x, P) \tag{1.1}$$

This entropy measures how hard it is to forecast the outcome on average, when true distribution P of outcomes is known and used as the forecasting model. We can often think of this quantity as a measure of uncertainty in the distribution, and as we will see this quantity is also closely related to the Bayes-risk of decision problems (section **??**).

A further quantity of interest is the divergence between two distributions P and Q.

Definition 2 (Generalised divergence)**.** *Given a scoring rule $S : \mathcal{X} \times \mathcal{M}^1_{\mathcal{X}} \mapsto \mathbb{R}$, let us define the divergence between two distributions $P, Q \in \mathcal{M}^1_{\mathcal{X}}$ as follows:*

$$d_S[P\|Q] = \mathbb{E}_{x\sim P} S(x, Q) - \mathbb{E}_{x\sim P} S(x, P). \tag{1.2}$$

The divergence measures how much worse we are at forecasting a quantity X sampled from a distribution P when instead of using the true distribution P, we use an alternative probability

distribution, Q. Ideally, we would like to see that using the true model P should always be better or at least as good as using any alternative model Q, but this is not automatically true for all scoring rules. A scoring rule that has this property is called a *proper scoring rule.*

Definition 3 (Proper scoring rule)**.** *$S : \mathcal{X} \times \mathcal{M}^1_{\mathcal{X}} \mapsto \mathbb{R}$ is a* proper scoring rule *with respect to a class of distributions $\mathcal{Q}$ if $\forall P, Q \in \mathcal{Q}$ the following inequality holds:*

$$\mathbb{E}_{x\sim P} S(x, Q) \geq \mathbb{E}_{x\sim P} S(x, P), \tag{1.3}$$

or equivalently in terms of the divergence $d_S[\cdot\|\cdot]$:

$$d_S[P\|Q] \geq 0. \tag{1.4}$$

The scoring rule s is said to be strictly proper *w. r. t. $\mathcal{Q}$ if equality holds only when $P = Q$.*

The divergence is a measure of the difference between two distributions P and Q. Even if the scoring rule is proper, and therefore $d_s[P\|Q] \geq 0$ always holds, the divergence is normally non-symmetric, that is $d_S[P\|Q] \neq d_S[Q\|P]$. Divergences are often used to match or approximate some *true* or *ideal* distribution with something *approximate*, so that the divergence between the truth and the approximation is minimal. As we can measure divergence in both ways, there is a question of which direction of divergence is to be calculated.

Definition (1.2) suggests that the the first argument, P, should take the role of the true distribution, and Q the approximate. **TODO: elaborate on this.**

So far we have only introduced quantities describing a single random variable, and comparing probability distributions over the same variable. We can extend the scoring rule framework to define information quantities that describe the relationship between multiple variables. A particularly useful quantity is the value of information, that measures the dependence between.

Definition 4 (Generalised value of information)**.** *Let X, Y be random variables with joint distribution $P \in \mathcal{M}^1_{\mathcal{X}\times\mathcal{Y}}$. Let $S : \mathcal{X} \times \mathcal{M}^1_{\mathcal{X}} \mapsto \mathbb{R}$ be a scoring rule over the variable X. We define the value of information in variable Y about variable X with respect to the scoring rule S as*

$$\mathbb{I}_S[X, Y] = \mathbb{E}_{x\sim P_X} S(x, P_X) - \mathbb{E}_{y\sim P_Y} \mathbb{E}_{x\sim P_{X|Y=y}} S(x, P_{X|Y=y}) \tag{1.5}$$

Alternatively, we can write information in terms of the generalised entropy or divergence functions

$$\mathbb{I}_S[X, Y] = \mathbb{H}_S[P_X] - \mathbb{E}_{y\sim P_Y} \mathbb{H}_S\left[P_{X|Y=y}\right] \tag{1.6}$$

$$= \mathbb{E}_{y\sim P_Y} d_S\left[P_X \middle\| P_{X|Y=y}\right] \tag{1.7}$$

This quantity measures the extent to which observing the value of Y is useful in forecasting variable X. Remarkably, this information quantity is non-symmetric. Indeed, the definition only requires a scoring rule over the variable X, but none over variable Y, so defining the value of information in Y about X does not even imply a definition of the value of information in X about Y.

If the scoring rule is proper, the value of information is always non-negative. Furthermore, if the scoring rule is strictly proper, the information is zero, if and only if the two variables are independent.

Theorem 1. *Let $S : \mathcal{X} \times \mathcal{M}^1_{\mathcal{X}} \mapsto \mathbb{R}$ be a strictly proper scoring rule with respect to probability distributions $\mathcal{M}^1_{\mathcal{X}}$, and $P \in \mathcal{M}^1_{\mathcal{X}\times\mathcal{Y}}$ the joint probability of variables X and Y. Then the two statements are equaivalent:*

1. *$\mathbb{I}_S[X, Y] = 0$*
2. *the variables X and Y are independent*

Proof. If X is independent of Y, then $\forall y : P_{X|Y=y} = P_X$, which implies $\forall y : d_S\left[P_X \middle\| P_{X|Y=y}\right] = 0$, and hence $\mathbb{I}_S[X, Y] = 0$.

On the other hand, $\mathbb{I}_S[X, Y] > 0$ implies $\exists y : d_S\left[P_X \middle\| P_{X|Y=y}\right] > 0$, therefore by strict propriety of S, $\exists y : P_X \neq P_{X|Y=y}$, which contradicts independence. □

As a corollary, strictly proper scoring rules are equivalently strong in the sense that if one detects dependence between variables, than any of them will:

Corollary 1. *Let $S_1, S_2 : \mathcal{X} \times \mathcal{M}^1_{\mathcal{X}} \mapsto \mathbb{R}$ be two strictly proper scoring rules over X. X and Y are two random variables. Then $\mathbb{I}_{S_1}[X, Y] > 0$ if and only if $\mathbb{I}_{S_2}[X, Y] > 0$.*

It also follows that the value of information defined by strictly proper scoring rules is weakly symmetric in the following sense:

Corollary 2. *Let $S_X : \mathcal{X} \times \mathcal{M}^1_{\mathcal{X}} \mapsto \mathbb{R}$ be two strictly proper scoring rule over X and $S_Y : \mathcal{Y} \times \mathcal{M}^1_{\mathcal{Y}} \mapsto \mathbb{R}$ be two strictly proper scoring rule over Y. Then $\mathbb{I}_{S_X}[X, Y] > 0$ if and only if $\mathbb{I}_{S_Y}[Y, X] > 0$.*

1.2 Examples of scoring rules

After having discussed general properties of scoring rules and information quantities based on them, let us look at particular examples of scoring rule and the entropies and divergences they define.

1.2.1 The log score

The most straightforward, and most widely used scoring rule is the log-score:

$$S_{log}(x, p) = -\log p(x) \tag{1.8}$$

Maximum likelihood estimation can be interpreted as score-matching with the log score:

$$\theta_{ML} = \underset{\theta}{\operatorname{argmax}} \sum_{n=1}^{N} \log p(x_i|\theta) \tag{1.9}$$

The associated entropy function is Shannon's differential entropy for continuous distributions

$$\mathbb{H}_{log}[p] = -\int p(x) \log p(x) dx \tag{1.10}$$

The resulting divergence function is the Kullback-Leibler (KL) divergence, which is very widely used in approximate Bayesian inference:

$$d_{KL}[p\|q] = \int \frac{\log p(x)}{\log q(x)} p(x) dx \tag{1.11}$$

The KL divergence is only well-defined when the distribution q is absolutely continuous with respect to p. This is one of the most important limitations of the KL divergence for our purposes in later chapters: If p is a continuous density, then q has to be absolutely continuous as well for the KL divergence to be defined. Therefore we cannot compute the KL divergence of, say, an empirical distribution of samples from a continuous distribution. A related problem is that Shannon's entropy of atomic distributions or mixed atomic and continuous distributions is either not well defined, or is trivial and depends only on the relative weight of the atoms but ot on their locations.

These problems all stem from a property of the log-score, known as locality: The value of the scoring rule $s(x, p)$ only depends on the value of the density function evaluated at the point x. This is a unique property of the log score. Any strictly proper and strictly local scoring rule is analogous to the log-score.

The value of information becomes Shannon's mutual information, a crucial quantity in channel coding []. Interestingly, Shannon's mutual information can be rewritten as the KL divergence between the joint distribution and the product of marginals:

$$\mathbb{I}_{Shannon}[X,Y] = \mathbb{H}_{Shannon}[X] - \mathbb{E}_{y\sim P_Y}\mathbb{H}_{Shannon}\left[P_{X|Y=y}\right] \tag{1.12}$$
$$= \mathbb{E}_{y\sim P_Y} d_{KL}\left[P_X \middle\| P_{X|Y=y}\right] \tag{1.13}$$
$$= \mathbb{E}_{y\sim P_Y}\left[\mathbb{E}_{x\sim P_{X|Y=y}} \log \frac{P_{X|Y=y}(x)}{P_X(x)}\right] \tag{1.14}$$
$$= \mathbb{E}_{(x,y)\sim P} \log \frac{P(x,y)}{P_X(x)P_Y(y)} \tag{1.15}$$
$$= d_{KL}\left[P(x,y) \| P_X(x)P_Y(y)\right] \tag{1.16}$$

As a consequence, Shannon's information is actually symmetric. The Shannon information in Y about X is the same as the Shannon information in X about Y. This is a remarkable property of the log-score and ,as we concluded in the previous section, is not generally true for value of information defined based on general scoring rules.

For completeness, we note here that some authors have generalised Shannon's mutual information along the lines of (1.16), by replacing the KL divergence with a more general divergence d:

$$\mathbb{J}_d(X,Y) = \mathrm{d}\left[P(x,y) \| P_X(x)P_Y(y)\right] \tag{1.17}$$

Examples of information functionals defined this way are []. On one hand, an information functional like $\mathbb{J}$ has several nice properties, most notably that it is always symmetric. On the other hand, in the general case we loose the intuitive meaning of information as "the extent to which observing the value of one variable is useful for predicting the value of the other one". Furthermore, if we wanted to use a divergence function corresponding to a scoring rule, the scoring rule should be defined over the joint space $\mathcal{X} \times \mathcal{Y}$, which is often not desired.

1.2.2 The pseudolikelihood

The idea of maximum pseudolikelihood estimation was introduced originally by [**?**] to estimate parameters of Gaussian random fields. Later it was popularised in the context if parameter estimation in Boltzmann machines [**?**] and Markov random fields. The pseudolikelihood is particularly useful for estimating parameters of statistical models with intractable normalisation constants.

$$S_{\text{pseudo}}(x,P) = -\sum_{d=1}^{D} \log P(x_d|x_{\neg d}), \tag{1.18}$$

Where $x_{\neg d}$ denotes the vector composed of all components of x other than the d^{th} component x_d.

In the pseudo-likelihood each of the terms is the conditional probability over one variable conditioned on all the remaining variables. Such quantities can be computed by marginalising a single variable at a time, therefore by computing a one dimensional integral or sum

$$p(x_d|x_{\neg d}) = \frac{p(x)}{\int p(x_d = y, x_{\neg d})dy} = \frac{C \cdot p(x)}{\int C \cdot p(x_d = y, x_{\neg d})dy} \tag{1.19}$$

This can be computed even if the joint probability of all variables P is known only up to a multiplicative constant C.

Take the Boltzmann distribution with parameters W and b as an example.

$$P(x) = \frac{1}{Z}\exp(x^T W x + b^T x), x \in \{0,1\}^D, \tag{1.20}$$

where $X = \sum_{x\in\{0,1\}^D} \exp(x^T W x + b^T x)$ is the partition function or normalisation constant that is analytically intractable to compute in the general case. On the other hand, the conditional distribution of a single component of x conditioned on the rest is easy to compute as follows:

$$P(x_d|x_{\neg d}, W, b) = \frac{p(x)}{\int p(x_d = y, x_{\neg d})dy} \tag{1.21}$$

$$= \frac{\frac{1}{Z}\exp(x^T W x + b^T x)}{\sum_{x_d \in \{0,1\}} \frac{1}{Z}\exp(x^T W x + b^T x)} \tag{1.22}$$

$$= \frac{\exp(x^T W x + b^T x)}{\sum_{x_d \in \{0,1\}} \exp(x^T W x + b^T x)} \tag{1.23}$$

$$= \frac{\exp\left(x_d\left(W_{d,d} + 2W^T_{d,\neg d}x_{\neg d} + b_d\right)\right)}{\exp(W_{d,d} + 2W^T_{d,\neg d}x_{\neg d} + b_d) + 1} \tag{1.24}$$

$$\tag{1.25}$$

The pseudo-likelihood thus becomes a sum of easy-to-compute sigmoidal terms. These sigmoidal terms, and their derivatives with respect to parameters W and b can be computed in polynomial time, allowing for fast estimation algorithms. [**?**] showed that pseudolikelihood estimation is consistent for fully visible Boltzmann machines.

The difference between the pseudolikelihood score and the log score becomes more apparent when rewriting the log score by the chain rule of joint probabilities:

$$S_{\log}(x, p) = -\log P(x) = -\sum_{d=1}^{D} \log P(x_d|x_{1:d-1}) \tag{1.26}$$

Here the d^{th} term is a probability conditioned on $d-1$ variables, and computing the d^{th} term therefore would require $D - d$ dimensional integral. The pseudo-likelihood makes computations more efficient by conditioning on more variables than needed by the chain rule. The two scoring rules are equivalent if and only if the joint distribution P conforms to a directed acyclic graphical model, i. e. there is a *natural causal ordering* of variables $\pi : \{1 \dots D\} \mapsto \{1 \dots D\}$ such that $X_{\pi_d} \perp\!\!\!\perp X_{\pi_{d+1}}, \dots, X_{\pi_D} | X_{\pi_1}, \dots, X_{\pi_{d-1}}$.

[] showed that pseudolikelihood estimation strictly proper for strictly postitive distributions. Moreover, for always positive distributions the following generalisation of the pseudolikelihood is also strictly proper scoring rule:

$$S_{\text{DLP12}}(x, P) = -\sum_{d=1}^{D} S_d\left(x_d, P(x_d|x_{\neg d})\right), \tag{1.27}$$

Where S_d are strictly proper scoring rules for each dimension

1.2.3 The kernel scoring rule

To my knowledge, the kernel scoring rule first appeared in the statistics literature in [**?**], who referred to it by the name *kernel scoring rule*. Recently, essentially the same concept, but derived from different first principles, has become known in the machine learning community as *maximum mean discrepancy* (MMD, []), and has been adopted in a variety of applications in machine learning and statistics, including two sample tests [], kernel moment matching [] embedding of probability distributions [] and the kernel-based message passing [].

Here I am going to define the kernel scoring rule by first introducing the divergence it gives rise to, maximum mean discrepancy, following the definitions in [**?**].

MMD measures the divergence between two distributions, p and q. It belongs to a rich class of divergences called integral probablity metrics[**?**], which define the distance between p and q, with respect to a class of integrand functions $\mathcal{F}$ as follows:

$$d_{\mathcal{F}}\left[p \| q\right] = \sup_{f \in \mathcal{F}} \left| \int f(x)p(x)dx - \int f(x)q(x)dx \right| \tag{1.28}$$

Intuitively, if two distributions are close in the integral probability metric sense, then no matter which function f we choose from function class $\mathcal{F}$, the difference in its integral over p or q should be small. This class of divergences include Wasserstein distance [], Dudley metric [] and MMD, which differ in their choice of the function class $\mathcal{F}$.

A particularly interesting case is when the function class $\mathcal{F}$ is functions of unit norm from a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$. In this case, the MMD between two distributions can be conveniently expressed using expectations of the associated kernel $k(x, x')$ only [**?**]:

TODO: do we need the square? Yes we do, and we have to explain why

$$
\begin{aligned}
d_{MMD,\mathcal{H}}\left[p\|q\right]^2 &= \sup_{\substack{f\in\mathcal{H}\\ \|f\|_{\mathcal{H}}=1}} \left| \int f(x)p(x)dx - \int f(x)q(x)dx \right|^2 && (1.29)\\
&= \sup_{\substack{f\in\mathcal{H}\\ \|f\|_{\mathcal{H}}=1}} \left| \int \langle f, k(\cdot,x)\rangle\, p(x)dx - \int \langle f, k(\cdot,x)\rangle\, q(x)dx \right|^2 && (1.30)\\
&= \sup_{\substack{f\in\mathcal{H}\\ \|f\|_{\mathcal{H}}=1}} \left| \left\langle f, \int k(\cdot,x)p(x)dx - \int k(\cdot,x)q(x)dx \right\rangle \right|^2 && (1.31)\\
&= \sup_{\substack{f\in\mathcal{H}\\ \|f\|_{\mathcal{H}}=1}} \langle f, \mu_p - \mu_q\rangle^2 && (1.32)\\
&= \|\mu_p - \mu_q\|_{\mathcal{H}}^2 && (1.33)\\
&= \iint k(x,y)p(x)p(y)dxdy - 2\iint k(x,y)p(x)q(y)dxdy + \iint k(x,y)q(x)q(y)dxdy, && (1.34)
\end{aligned}
$$

In the derivation above $\mu_p(\cdot) = \int k(\cdot, x)p(x)dx$ is the so called mean element or RKHS embedding of the probability distributions p. The most interesting kernels for the purposes of Hilbert-space embedding of distributions are those called *characteristic* []. If the kernel k is characteristic, the mapping from Borel probability measures to mean elements in a characteristic RKHS is injective, that is $\mu_p = \mu_q \iff p = q$. This also means that for characteristic Hilbert spaces $d_{MMD,\mathcal{H}}\left[p\|q\right] = 0 \iff p = q$ holds.

The mean embedding μ_p can be thought of as a generalisation of characteristic functions []. The characteristic function of a probability distribution p over the real line is defined as follows:

$$\phi_p(t) = \mathbb{E}_{x\sim p}\left[e^{itx}\right] = \int e^{itx}p(x)dx, \tag{1.35}$$

where i is the imaginary number $i = \sqrt{-1}$. The characteristic function is known to uniquely characterise any Borel probability measure on the real line. Indeed, it corresponds to an RKHS-embedding with the fourier kernel $k_{Fourier}(x, y) = \exp(ixy)$, which is an example of characteristic kernels. Note, that the final formula 1.34 assumed a real valued kernel function, therefore it is not valid for the Fourier kernel. Other, practically more relevant examples of characteristic kernels include the squared exponential, and the Laplacian kernels (see chapter **??**). As a counterexample, polynomial kernels, and in general kernels corresponding to finite dimensional Hilbert spaces are not characteristic.

The maximum mean discrepancy with characteristic kernels has been applied in various contexts in machine learning. One of the first of these recent application were two-sample tests. In two-sample testing we are provided i. i. d. samples from two distributions, and we have to determine whether the two distributions are the same or not. [] developed and analysed empirical estimators of MMD for this problem. Herding [], a method for generating pseudosamples has been shown to minimise MMD between a target distribution and the empirical distribution of pseudo-samples. Lastly, in kernel moment matching [**?**] MMD is used for density estimation: parameters

of a parametric density model are set by minimising MMD from the empirical distribution of data. This is a special case of score matching, as we will see shortly.

MMD in fact is a Bregman divergence of the form, that corresponds to the following scoring rule:

$$s_{MMD,k}(x,q) = k(x,x) - 2\int k(x,y)q(y)dy + \iint k(y,z)q(y)q(z)dydz \tag{1.36}$$

$$= \sup_{\substack{f\in\mathcal{H}\\ \|f\|_{\mathcal{H}}=1}} \left| f(x) - \int f(y)q(y)dy \right|^2 \tag{1.37}$$

The generalised entropy defined by this scoring rule becomes:

$$S_{MMD,k}[q] = \int k(x,x)q(x)dx - \iint k(x,y)q(x)q(y)dxdy \tag{1.38}$$

This entropy function is very general, and has several favourable properties in comparison to Shannon's entropy. Firstly, if we assume that the kernel k is non-negative and bounded, then the entropy functional is also non-negative and bounded. Secondly, The only requirement for the distribution q is that we can compute expectations with respect to it. This means that every probability distribution, and indeed every Borel measure, has a well-defined entropy of this form. This is not true for the Shannon's differential entropy, where the entropy of atomic distributions or mixtures of atomic and continuous distributions is not well defined. Thirdly, the entropy function has the kernel as free parameter, which is a curse and a blessing at the same time **TODO: check idiom**. On the one hand this gives us extra flexibility: even if we commit to the square exponential kernel, we can fine-tune the entropy function to our needs by adjusting the length-scale parameter [**?**]. On the other hand there is no general principled way of choosing the kernel or it's parameters if we are unsure what it should be.

The divergence between two distributions p and q under the kernel scoring rule becomes the maximum mean discrepancy.

$$d_{MMD,k}[p\|q] = \mathbb{E}_{x\sim p}[s(x,q)] - \mathbb{E}_{x\sim p}[s(x,p)] \tag{1.39}$$

$$= \int k(x,x)p(x)dx - 2\int k(x,y)p(x)q(y)dydx + \iint k(y,z)q(y)q(z)dydz \tag{1.40}$$

$$- \left(\int k(x,x)p(x)dx - \iint k(x,y)p(x)p(y)dxdy \right) \tag{1.41}$$

$$= \iint k(x,y)p(x)p(y)dxdy - 2\iint k(x,y)p(x)q(y)dxdy + \iint k(x,y)q(x)q(y)dxdy \tag{1.42}$$

kernel value of information

$$\mathbb{I}_k[X,Y] = \mathbb{E}_{y\sim P_Y} d_k\left[P_X \middle\| P_{X|Y=y}\right] \tag{1.43}$$

$$= \mathbb{E}_{y\sim P_Y} \|\mu_{X|Y=y} - \mu_X\|^2_{\mathcal{H}} \tag{1.44}$$

$$= k(P_X,P_X) - 2 * \mathbb{E}_{y\sim P_Y} k(P_X, P_{X|Y=y}) + \mathbb{E}_{y\sim P_Y} k(P_{X|Y=y}, P_{X|Y=y}) \tag{1.45}$$

$$= \mathbb{E}_{y\sim P_Y} \mathbb{E}_{P_{x_1,x_2\sim P_{X|Y=y}}} k(x_1,x_2) - \mathbb{E}_{x_1,x_2\sim P_X} k(x_1,x_2) \tag{1.46}$$

Note: MMD can be derived from a loss-calibrated viewpoint Let's say your task is to estimate value of a functions $f \in \mathcal{F}$ evaluated at θ. The action can be interpreted as a functional $a : \mathcal{F} \mapsto \mathbb{R}$, that gives the estimated value of $f(\theta)$ for any function $f \in \mathcal{F}$. The loss ℓ you incur is equal to the maximal squared error you incur on any of these functions.

$$\ell(\theta, a) \sup_{f \in \mathcal{F}} \left(f(\theta) - a(f)\right)^2 \tag{1.47}$$

Given a probabilistic forecast p over θ, the Bayes optimal decision $a(f)$ simply computes the mean of f under the distribution p:

$$a_p^* = \int f(\theta) p(\theta)\, d\theta \tag{1.48}$$

Thus, we can define the following scoring rule S:

$$S(\theta, p) = \sup_{f \in \mathcal{F}} \left(f(\theta) - \int f(\theta) p(\theta)\, d\theta \right)^2 \tag{1.49}$$

When $\mathcal{F}$ is chosen to be the unit ball in a reproducing kernel Hilbert space $\mathcal{H}$ defined by a positive definite kernel k, this scoring rule will be equivalent to the kernel scoring rule for probability distributions:

1.2.4 Scoring rules based on general decision problems

The scoring rule framework is very flexible, in fact for every Bayesian decision problem it is possible to derive a corresponding scoring rule as we will show in this section.

Let us assume we are faced with a decision problem of the following form: We have to decide to take one of several possible actions $a \in \mathcal{A}$. The loss/utility of our action will depend on the action we have chosen and on the state of the environment X, the value of which is unknown to us. If the environment is in state X, and we choose action *action*, we incur a loss $\ell(X, a)$. Let us assume we have a probabilistic forecast or belief $P(x)$ about the state of the environment. Given this we can choose an action that minimises our expected loss:

$$a_P^* = \operatorname*{argmax}_{a \in \mathcal{A}} \mathbb{E}_{x \sim P} \ell(x, a) \tag{1.50}$$

When we observe the value of X we can score the probabilistic forecast, by evaluating the loss incurred by using this optimal action a_P^* in state X.

$$S(x, P) = \ell(x, a_P^*) \tag{1.51}$$

Chapter 2

Visualising information geometries

In this section I aim to develop an understanding of the differences between various scoring rules and corresponding divergence metrics by visualising the manifold structure they give rise to. For the purposes of illustration here I concentrate only on two-parameter families of distributions such as Gaussian distributions, as these define two-dimensional manifolds which can then be shown in a two dimensional pages of a thesis.

Our goal is to create a two-dimensional map of particular manifolds in such a way that distances measured between points on the map correspond to geodesic distances measured along manifold as precisely as possible. First, it is important to understand that a perfect embedding of this sort does not always exists.

Take the surface of a three-dimensional ball as an example. The sphere is a two-dimensional manifold, which can be parametrised by longitude and latitude. Still, it is impossible to stretch this surface out and represent it faithfully in two dimensional Cartesian coordinate system. This problem – representing the surface of a three-dimensional object as part of a two-dimensional plane – is in fact at the core of cartography, and is called *map projection*. When drawing a full map of the surface of the Earth, usually the manifold has to be cut at certain places, but even then, the embedding is only approximate. There are various map projections used in cartography, and the purpose for which the map is used dictates what kind of distortions are tolerable, and what is not.

Understanding that a perfect map of two-dimensional statistical manifolds cannot necessarily be produced, we will resort to approximate embedding techniques developed in the machine learning community. These approximate embedding procedures numerically find a *stretched* manifold in two dimensions that best represents distances on the statistical manifold defined by a particular scoring rule and divergence.

Using numerical isometric embedding techniques developed in the machine learning community, it is possible to visualise the differences between geometries induced by various scoring rules. To do this here we use a techniques called ISOMAP [**?**].

1. take a set of probability distributions, preferably so that they relatively densely cover an interesting part of the manifold
2. compute pairwise symmetrised divergence matrix between the selected distributions
3. compute approximate geodesics
4. use metric multidimensional scaling technique embed distributions as points a two-dimensional space

2.0.5 Visualising the Shannon-information geometry

The most widely used divergence in statistics and machine learning is without doubt the Kullback-Leibler divergence. Here I show the geometry it induces on various parametric families of distri-

butions.

Let us start with a very simple, single parameter distribution, the Bernoulli. A Bernoulli distribution describes a binary random variable, where the parameter controls the probability of taking value 1. In Figure **??** I illustrate the differences between the KL divergence, and the Brier divergence, which corresponds simply to the Euclidean distance between parameter values. As we can see the KL divergence is more sensitive to differences in small (close to 0) and large (close to 1) probabilities, but puts less emphasis on.

When using the KL divergence or the log-score in practical situations, such as in classification, we should therefore expect that much of the statistical power is going to be spent on faithfully matching small probabilities. This is not always desirable: Imagine we were to model the probability that users click on certain news articles on an on-line news website. In this application, most potential clicks have negligible probability, but some user-article combinations may have probabilities closer to 0.5. If we are to build a recommender system based on this analysis, it is these large probabilities that will be of importance. In this case we are better off using the Brier-score, rather than the log-score which spends serious effort in modelling how small are the small probabilities exactly.

Gaussian distributions are probably the most important family of distributions due to their convenient analytical properties. **TODO: further blah blah about this** The KL divergence between two univariate Gaussian distributions is available in a closed form and is given by the following formula:

$$d_{KL}\left[\mathcal{N}_{\mu_1,\sigma_1}\|\mathcal{N}_{\mu_2,\sigma_2}\right] = \frac{\left(\mu_1-\mu_2\right)^2}{2\sigma_2^2} + \frac{1}{2}\left(\frac{\sigma_1^2}{\sigma_2^2} - 1 - \log\frac{\sigma_1^2}{\sigma_2^2}\right) \tag{2.1}$$

Figure **??** illustrates the manifold structure of normal distributions induced by the KL divergence. We can observe that assuming p and q have the same mean, the larger their variance, the easier it becomes to distinguish between them.

We can look at the geometry Shannon's entropy induces within another two-parameter family of continuous distributions, Gamma distributions. Gamma distributions are strictly positive, their probability density function of Gamma distributions is as follows:

$$(x) = \beta^\alpha \frac{1}{\Gamma(\alpha)} x^{\alpha-1}\exp(-\beta x) \tag{2.2}$$

where $\alpha, \beta > 0$ are called shape and rate parameters respectively. Special cases of Gamma distributions are exponential distributions when $\alpha = 1$.

The KL divergence between Gamma distributions can be computed in closed form and is given by the following formula:

$$d_{KL}\left[\Gamma_{\alpha_1,\beta_1}\|\Gamma_{\alpha_2,\beta_2}\right] = \left(\alpha_1-\alpha_2\right)\psi\left(\alpha_1\right) - \log\frac{\Gamma(\alpha_1)}{\Gamma(\alpha_2)} + \alpha_1\log\frac{\beta_1}{\beta_2} + \alpha_1\frac{\beta_2-\beta_1}{\beta_1} \tag{2.3}$$

Figure **??** shows the manifold of Gamma distributions for parameters $a \leq \alpha \leq b, c \leq \beta \leq d$. As we can see this manifold is less symmetric than that of the Gaussians.

For large values of α the standard deviation of the distribution shrinks, and by the central limit theorem, the distribution converges to a Gaussian. We can illustrate this convergence in the manifold structure. For this we first reparametrise the Gamma distribution in terms of its mean and standard deviation. The mean and standard deviation of a Gamma distribution with parameters α and β are given by the following formulae:

$$\mu = \frac{\alpha}{\beta} \tag{2.4}$$

$$\sigma^2 = \frac{\alpha}{\beta^2} \tag{2.5}$$

Solving for α and β in these equations we get

$$\alpha = \frac{\mu^2}{\sigma^2} \tag{2.6}$$

$$\beta = \frac{\mu}{\sigma^2} \tag{2.7}$$

Plugging these into Eqn. (2.3) we can now map Gamma distributions with particular mean and variance. Figure 1 compares Normal and Gamma distributions with mean $\mu \in [0.5, 1.5]$ and standard deviation $\sigma \in [0.1, 1]$. We can observe that as the variance increases, the manifold of Gamma distributions shows a fan-like structure very similarly that of Normal distributions. However, for larger variance, the distributions look less Gaussian, and the manifold becomes more asymmetric. The effect of the central limit theorem would perhaps be even more prominent for smaller values of σ, but for those cases that case Eqn. (2.3) becomes numerically imprecise, as it relies on look-up-table implementation of the Gamma (Γ) and bigamma (ψ) functions.

2.0.6 Visualising geometries induced by divergences other than KL

The main purpose of this section is to visualise differences between the geometries induced by various divergence measures over the same set of distributions. Here we will mainly focus on Gaussian distributions, as it is analytically convenient to compute various divergences between Gaussians in closed form.

A particularly interesting divergence that we will use in subsequent chapters is that based on the kernel scoring rule, called the MMD (section **??**). The kernel scoring rule itself is very flexible, and its properties are dictated by the choice of kernel function.

For several well-known kernels the MMD between two univariate Gaussians can be computed in closed form. For the squared exponential kernel $k(x, y) = \exp(-\frac{(x-y)^2}{\ell^2})$ the divergence is given by the following formula:

$$d_{MMD}\left[\mathcal{N}_{\mu_1,\sigma_1} \| \mathcal{N}_{\mu_2,\sigma_2}\right] = \ell \left(\frac{1}{\ell + 2\sigma_1} + \frac{1}{\ell + 2\sigma_2} - \frac{2}{\ell + \sigma_1 + \sigma_2} \exp\left(-\frac{(\mu_1 - \mu_2)^2}{(\ell + \sigma_1 + \sigma_2)^2} \right) \right) \tag{2.8}$$

Figure **??** illustrates the map according to the MMD divergence choosing various values for the lengthscale ℓ. **TODO: conclusions** We observe that the structure of the manifold induced by this divergence is qualitatively very similar to that induced by the KL divergence. However, using MMD with the squared exponential kernel allows us the extra flexibility of choosing a characteristic lengthscale, thereby modulating the sensitivity to small differences in variance and mean.

Another widely used kernel is the so-called Laplacian: $k(x, y) = \exp\left(-\frac{|x-y|}{\ell}\right)$, for which the MMD between Gaussian distributions can still be computed in closed form:

TODO: find out what the formula is

Not all scoring rules give rise to smooth manifolds. As an extreme example, consider the following decision problem:

You are uncertain about the temperature of the reactor in a power plant. If the temperature is too high, above a critical temperature T_{crit}, the reactor may melt down causing you a loss of \$10 billion. You may choose to shut down the reactor, which costs you \$1 million of lost revenue, irrespective of whether the reactor was indeed overheated or not. You make a probabilistic forecast about the reactor's temperature, and want to make a decision based on that.

This decision rule segments probabilistic forecasts into only two subsets: those which would result in a "shut down" decision, and those that result in a "keep on going".

$$d_{reactor}\left[p \| q\right] = \begin{cases} 0 & p(\{t \geq T_{crit}\}) \geq \ell \text{ and } q(\{t \geq T_{crit}\}) \geq \ell \\ \ell_1 & p(\{t \geq T_{crit}\}) \geq \ell \text{ and } q(\{t \geq T_{crit}\}) \leq \ell \\ \ell_2 & p(\{t \geq T_{crit}\}) \leq \ell \text{ and } q(\{t \geq T_{crit}\}) \geq \ell \end{cases} \tag{2.9}$$

This divergence therefore does not give rise to a smooth manifold. Figure **??** shows a map of Gaussian distributions with respect to the KL divergence. The way $d_{reactor}[\cdot\|\cdot]$ segments distributions into "shut down" or "keep on going" types is also shown. We can make the KL divergence more sensitive to the decision problem at hand by considering a convex combination between $d_{KL}[\cdot\|\cdot]$ and $d_{reactor}[\cdot\|\cdot]$.

Chapter 3

Approximate Bayesian inference

Chapter 4

The role of information geometry in approximate Bayesian analysis

4.1 Introduction

In practically interesting Bayesian models, the posterior is often computationally intractable to obtain and therefore one has to resort to approximate inference techniques. The most popular approximation methods are variational inference and Markov chain Monte Carlo.

Variational methods operate by minimising an information theoretic divergence between a simple, often exponential family, distribution and the true posterior. The divergence is often chosen to be a form of Kullback-Leibler divergence, as it allows easy rearrangement of terms and makes local message-passing style computations possible. In section **??** argue that when Bayesian inference is performed to solve a particular decision problem, these algorithms are sub-optimal as they are ignorant of the structure of losses. We devised a framework we termed loss-calibrated approximate inference [], which generalises traditional variational approaches by minimising generalised divergences based on scoring rules. I will demonstrate this framework on a loss-critical toy problem and on a well-known nonparametric Bayesian model, Gaussian process regression.

Monte Carlo methods produce random samples (approximately) drawn from the posterior, which then allows for approximating relevant integrals over the posterior. Monte Carlo techniques are applicable to a wide variety of interesting Bayesian models, and allow for an intuitive trade-off between computation time and accuracy. However, just as most variational approaches, Monte Carlo techniques are also ignorant of the decisions and losses involved in a decision problem. In section **??** I introduce a new class of approximate inference algorithms that I call loss-calibrated quasi-Monte Carlo methods. These algorithms produce a deterministic sequence of pseudo-samples in such a way, that the divergence between the empirical distribution of pseudosamples is minimised from the target distribution. I show how kernel herding, a recent algorithm proposed by [**?**] can be seen as a special case of loss-calibrated quasi-Monte Carlo, and point out the connection between this method and Bayesian Quadrature.

We can also argue, that when we cannot perform inference exactly, the usual practice of performing approximate inference and then using the approximate posterior to calculate a decision is weakly motivated. One may want to instead directly approximate the optimal decision, without producing a direct estimate of the posterior. Following our work published in [**?**], I introduce approximate Bayesian decision theory, and derive an Expectation-Maximisation style variational algorithm for solving it. We illustrate the framework on Gaussian process classification, and present experimental comparisons to standard approaches based on approximate inference.

The work presented in this chapter on loss-calibrated approximate inference and approximate decision theory is joint work with Simon Lacoste-Julien and Zoubin Ghahramani, and most of the results presented here have been published in [**?**]. The work presented on the equivalence between

optimally weighted kernel herding and Bayesian Quadrature is joint work with David Duvenaud, and has been published [**?**].

4.2 Loss-calibrated approximate inference

Although often overlooked, the main theoretical motivations for the Bayesian paradigm are rooted in Bayesian decision theory [**?**], which provides a well-defined theoretical framework for rational decision making under uncertainty about a hidden parameter θ. The ingredients of Bayesian decision theory are (see Ch. 2 of [**?**] or Ch. 1 of [**?**] for example):

- a (statistical) loss $\ell(\theta, a)$ which gives the cost of taking action $a \in \mathcal{A}$ when the world state is $\theta \in \Theta$;
- an observation model $p(\mathcal{D}|\theta)$ which gives the probability of observing $\mathcal{D} \in \mathcal{O}$ assuming that the world state is θ;
- a prior belief $p(\theta)$ over world states.

The loss ℓ describes the decision task that we are interested in, whereas the observation model and the prior represent our beliefs about the world. Given these components, the ultimate objective for evaluating a possible action a after observing $\mathcal{D}$ is the *expected posterior loss* (also called the *posterior risk* [**?**])

$$\mathcal{R}_{p_\mathcal{D}}(a) \doteq \int_\Theta \ell(\theta, a)\, p(\theta|\mathcal{D}) d\theta \tag{4.1}$$

In the Bayesian framework, the optimal action $a_{p_\mathcal{D}}$ is the one that minimizes $\mathcal{R}_{p_\mathcal{D}}$.

In this framework it is therefore easy to see that Bayesian decision making decomposes into two separate computation. First, a posterior $p_\mathcal{D}$ is inferred from observed data $\mathcal{D}$, then the optimal action is selected by minimising risk under this posterior.

In many practically relevant cases computing the posterior is not analytically tractable. There are two reasons. Either the marginal likelihood cannot be computed analytically in closed form, or there is a closed form expression for the posterior, but its complexity increases exponentially with the amount of observed data, as in the case of for example switching state space models. Either way, it is usual practice to approximate the intractable posterior by something simpler, an approximate distribution q. The approximate distribution is often chosen from an exponential family of distributions $\mathcal{Q}$, and it is also often common practice to choose q such that it factorises over multivariate quantities.

Overview of KL minimisation in one way Variational methods

Overview of EP and minimizing KL in other way

But none of these takes into account the structure of the decision problem

Toy example

Framework

Example: Gaussian process regression In this case we do not actually need to perform approximate inference, as the posterior is Gaussian and available in closed form. However it allows us to express the quantities relevant for loss-calibrated approximate inference.

Gaussian process regression.

4.3 Loss-calibrated quasi-Monte-Carlo

Monte Carlo, powerful but

Chapter 5

Bayesian experiment design

Part III

Bayesian analysis in experimental sciences

Chapter 6

Cognitive tomography: Bayesian analysis of choice probabilities

Optimal experiment design v.s. MCMC with People

An important comparison we always wanted to make.

Chapter 7

Quantum Tomography

7.1 Introduction

Quantum computing and quantum communication are rapidly exploding areas of modern computer science.

Even though large classes of algorithms can be implemented efficiently using quantum computers, there is an important limitation that is a barrier to progress towards studying large quantum computers: state reconstruction. The hearth of this problem lies the fact that the end result of a quantum computation is a quantum state, and quantum states cannot be directly observed. In order to figure out what state a quantum computer produced as the result of computation one has to make a *measurement* on it. A measurement in quantum physics has two characteristics: Firstly, even if the state of the system on which the measurement is made and the measurement itself are fully known, the outcome of a measurement is generally non-deterministic. It is also true therefore that, in most cases, a single measurement doesn't provide full information about the state of the system, so repeated measurements are needed. Secondly, a measurement destroys, or at the very least modifies the quantum state itself. This means that there is only a limited amount of information one can observe about the quantum state in any experiment. To overcome these problems physicists studying quantum systems usually produce several independent copies of the same system (equivalent to "running" a quantum computer several times), and make measurements on each of the independent copies. Reconstructing the state on the basis of this batch of non-deterministic measurement outcomes is a statistical inference problem known generally as *state reconstruction* or *quantum state tomography*.

Technological and implementational constraints aside, a barrier in studying large, multipartite quantum systems today is that the number of independent copies required to accurately reconstruct the state via quantum tomography grows at least exponentially with the size (number of qubits) of the system. So even though a classically NP-complete algorithm can be implemented using polynomial number of quantum operations, reading out the result can still take exponentially long. Fortunately, in future practical applications of quantum computers, such as finding prime factors, rich prior information is available about the structure of the results, which can be exploited to speed up the tomography process.

However, in current experimental quantum physics, when researchers invent, for example, a novel physical implementation of a quantum gate, they have to demonstrate that in multiple situations their equipment produces a state that resembles the theoretically predicted state with high fidelity. Often these implementations are imperfect and the produced state isn't quite exactly the desired state. To be able to measure the success of their implementations, experimenters often have to perform full quantum tomography, or quantum hypothesis testing [**?**], which is equally resource-intensive. Therefore any method that speeds these processes up may be of great practical importance.

In this part of the thesis I will formally introduce the problem of quantum state tomography,

provide Bayesian analysis of the problem and then propose a

7.2 Overview of quantum statistics

quantum states

An example of a simple, two-dimensional quantum state is the polarisation state of a single photon. A photon's polarisation is described by two components: its linear polarisation, that is whether it's polarised horizontally (denoted as $|H\rangle$), vertically ($|V\rangle$) or at an angle in between. Light can also have circular polarisation. The two extremes are left ($|R\rangle$) and right ($|R\rangle$) circular polarisation. A combination of linear and circular polarisation can be represented by a unit-length complex number $|\phi\rangle = a + bi$

The polarisation state of a photon is indeed one of the most widely used physical model system used to demonstrate quantum phenomena on, and throughout this section I will use photons as an example to illustrate physical analogues of mathematical formalism. Other examples of quantum systems include [] For recent reviews on the current state of experimental quantum physics see [**?**].

The quantum state of a system cannot be directly observed, only via measurements performed on the system. Measurements in quantum physics have two distinctive features: the outcome is non-deterministic and performing a measurement alters the state of the system on which the measurement was performed.

An example of a measurement in case of a photon would be letting it pass through a linear polarising filter. Depending on the state of the photon $|\phi\rangle$ and the measurement describing the filter M_0, M_1, the photon either 'bounces back' from the filter or with a certain probability passes trough. By placing a photodetector after the polar filter one can record which one of these two outcomes happened. The probability of the two outcomes is governed by the state of the photon and the measurement itself.

Crucially, measuring a quantum system alters the state. This phenomenon is sometimes

For our purposes of quantum tomography we assume, that after one measurement has been made on a system, it's state is destroyed and we cannot use it anymore. Therefore after each measurement, once the outcome is recorded, the measured system is discarded, and a new, independent copy of the system is generated.

There are alternative approaches that use a sequence of measurements that only partially destroy the state; these approaches are referred to as weak or continuous measurement[**?**], and quantum control[**?**]. Weak measurements are of high importance in quantum cryptanalysis[**?**].

In the previous paragraphs we have seen that quantum measurements are inherently non-deterministic in nature. But in some cases there is another source of uncertainty effecting the outcome of out measurements. We will call this other source classical uncertainty, and when both kinds of uncertainties are present, we will say that the quantum system is in a *mixed state*. As an example, a quantum system in a mixed state can be a noisy source of photons that 50% of the time produces a horizontally polarised photon, 50% of the time a vertically polarised one, randomly.

Let us now assume that we are given two such noisy sources. One produces state $|H\rangle$ with probability $\frac{1}{2}$ and $|V\rangle$ with probability $\frac{1}{2}$. The second experiment produces state $\frac{1}{\sqrt{2}}|H\rangle + \frac{1}{\sqrt{2}}|V\rangle$ or $\frac{1}{\sqrt{2}}|H\rangle - \frac{1}{\sqrt{2}}|V\rangle$ randomly. Let's see what happens if we perform a measurement $\langle\phi|_0, \langle\phi|_0$ on the two noisy systems.

$$e = mc^2 \tag{7.1}$$

In both cases the probability of observing 0 and 1 is the same for both sources, and is a function of the measurement. We can therefore conclude that no matter what measurements we perform, there is no way to tell apart the two sources on the basis of observations. We therefore may call these two sources *observationally equivalent*. In more general terms, classical and quantum uncertainty cannot be disambiguated by observing a system. We can therefore define a equivalence classes of systems, and parametrise them via the so called density matrix ρ.

As we have seen, the two noisy systems in the previous example were equivalent, and indeed they had we can describe them by the same density matrix $\rho = \frac{1}{2}I$. In the context of photon sources, such light source is called *unpolarised*. There are several 'different' unpolarised light sources, but these are all equivalent observationally.

Born rule

Bloch sphere representation The centre of the Bloch sphere is the perfectly mixed state, whose density operator is proportional to identity $\rho = \frac{1}{D}I$. The surface of the sphere contains pure states. Of particular significance are

7.2.1 Inference in quantum tomography

In theprevious section I described how the outcome of a measurement depends on the measurement and the state of the system. In quantum tomography are given a sequence of copies of an unknown state ρ, perform a known measurements on each of these copies and observe their outcomes. Determining the state from observations is a classical statistical inference problem.

The first approaches to solving this inference problem tried directly 'inverting' the Born rule.

7.2.2 optimal experiment design and active tomography

7.3 Adaptive Bayesian Quantum Tomography

7.4 Results