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

Ferenc Huszár

August 7, 2012

Contents

Part I

Scoring rules and divergences in Bayesian analysis

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:*

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

TODO: Mention Bregman divergences [?].

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.

The divergence defined in (1.2) is a special case of Bregman divergences:

Definition 4 (Bregman divergence)**.** *Let H be a differentiable, strictly concave function on a convex domain Θ. For $P, Q \in \Theta$*

$$d_{Bregman,H}[P\|Q] = H(P) - H(Q) + \langle \nabla H(Q), Q - P\rangle \tag{1.5}$$

Statement 1 (Generalised divergences d_S for strictly proper S are Bregman divergences)**.** *Let S be a strictly proper scoring rule, with generalised entropy $\mathbb{H}_S[P]$. If $\mathbb{H}_S[P]$ is differentiable with respect to P, then the generalised divergence $d_S[P\|Q] = \mathbb{E}_{x\sim P} S(x, Q) - \mathbb{H}_S[P]$ is a Bregman divergence with $H(\cdot) = \mathbb{H}_S[\cdot]$.*

Proof. Review the definition of the entropy $\mathbb{H}_S[P]$:

$$\mathbb{H}_S[P] = \mathbb{E}_{x\sim P} S(x, P) = \langle P, S(\cdot, P)\rangle \tag{1.6}$$

Using this notation

$$\begin{aligned} \nabla \mathbb{H}_S[P] &= \nabla \langle P, S(\cdot, P)\rangle & (1.7)\\ &= S(\cdot, P) + \langle P, \nabla S(\cdot, P)\rangle & (1.8) \end{aligned}$$

The second term $\langle P, \nabla S(\cdot, P)\rangle = 0$ because of strictly proper property of S. Thus

$$\begin{aligned} d_{Bregman,\mathbb{H}_S}[P\|Q] &= \mathbb{H}_S[Q] + \langle \nabla \mathbb{H}_S[Q], P - Q\rangle - \mathbb{H}_S[P] & (1.9)\\ &= \mathbb{H}_S[Q] + \langle S(\cdot, Q), P - Q\rangle - \mathbb{H}_S[P] & (1.10)\\ &= \langle S(\cdot, Q), P\rangle - \mathbb{H}_S[P] & (1.11)\\ &= d_S[P\|Q] & (1.12) \end{aligned}$$

Concavity of $\mathbb{H}_S[P]$ also follows from strictly proper property $d_S[P\|Q] > 0, P \neq Q$. □

An intuitive explanation of Bregman divergences is given in Figure **??**.

[**?**, **?**] 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.

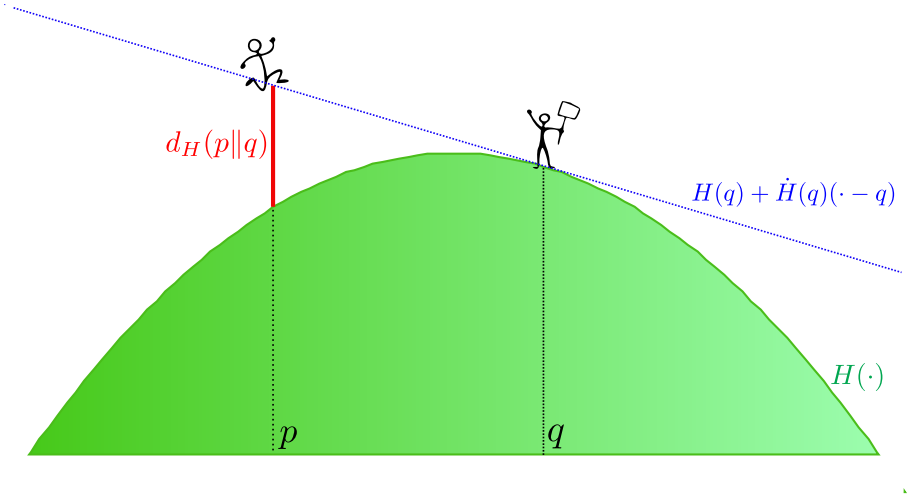

Figure 1.1: Pictorial illustration of Bregman divergences. Peter and Quentin are points who live on a convex hill, whose surface is described by the concave function $H(p)$. Peter lives at $(Q, H(P))$, Quentin at $(Q, H(Q))$. Because the hill is convex and they are both points, they cannot normally see each other, unless $P = Q$. Anyone above the tangential line $H[Q] + \dot{H}(Q)(\cdot - Q)$ can see Quentin, but Peter is normally below this line. If Peter wants to see Quentin, he has to jump up. The Bregman divergence $d_H[P\|Q]$ measures how high Peter has to jump to see Quentin. In this example H was chosen to be the Brier (quadratic) entropy, so here the divergence is symmetric, but this is not generally the case.

Definition 5 (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*

$$\mathbb{I}_S[X \leftarrow Y] = \mathbb{E}_{x\sim P_X} S(x, P_X) - \mathbb{E}_{y\sim P_Y}\mathbb{E}_{x\sim P_{X|Y=y}} S(x, P_{X|Y=y}) \tag{1.13}$$

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

$$\begin{aligned}\mathbb{I}_S[X \leftarrow Y] &= \mathbb{H}_S[P_X] - \mathbb{E}_{y\sim P_Y}\mathbb{H}_S\left[P_{X|Y=y}\right] &(1.14)\\ &= \mathbb{E}_{y\sim P_Y} d_S\left[P_X \middle\| P_{X|Y=y}\right] &(1.15)\end{aligned}$$

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. $\mathbb{I}_S[X \leftarrow Y] = 0$

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 $\mathbb{I}_S\left[X \leftarrow Y\right] = 0$.

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

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

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

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. I will review three widely used scoring rules, the logarithmic, Brier (quadratic) and spherical scores. Then I present the kernel scoring rule, and point out its connections to the maximum mean discrepancy, a divergence measure that gained popularity recently in the machine learning community. Then I define a novel scoring rule, called *kernel spherical scoring rule*, examine its properties, and provide a proof that it is strictly proper. Finally, I show the connections between scoring rules and Bayesian decision theory, and explain how decision problems give rise to scoring rules and associated information quantities.

1.2.1 The logarithmic score

The most straightforward, and most widely used scoring rule is the logarithmic score which is of the form:

$$S_{log}(x, P) = -\log P(x) \tag{1.16}$$

This score is widely used, most notably in maximum likelihood estimation of parametric models:

$$\theta_{ML} = \underset{\theta}{\operatorname{argmax}} \sum_{n=1}^{N} \log P(x_i|\theta) \tag{1.17}$$

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

$$\mathbb{H}_{Shannon}\left[P\right] = -\mathbb{E}_{x\sim P} \log P(x) \tag{1.18}$$

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

$$d_{KL}\left[P\|Q\right] = \mathbb{E}_{x\sim P} \frac{\log P(x)}{\log Q(x)} \tag{1.19}$$

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 continuous as well for the KL divergence to be defined. Therefore we cannot compute the KL divergence between, say, an empirical distribution of samples and 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 logarithmic 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 logarithmic score. Any strictly proper local scoring rule is analogous to the logarithmic score. Note, that there are weaker definitions of locality of scoring rules, which hold scoring rules other than the logarithmic [].

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:

$$\begin{aligned}
\mathbb{I}_{Shannon}\left[X \leftarrow Y\right] &= \mathbb{H}_{Shannon}\left[X\right] - \mathbb{E}_{y\sim P_Y}\mathbb{H}_{Shannon}\left[P_{X|Y=y}\right] && (1.20)\\
&= \mathbb{E}_{y\sim P_Y} d_{KL}\left[P_X \middle\| P_{X|Y=y}\right] && (1.21)\\
&= \mathbb{E}_{y\sim P_Y}\left[\mathbb{E}_{x\sim P_{X|Y=y}} \log \frac{P_{X|Y=y}(x)}{P_X(x)}\right] && (1.22)\\
&= \mathbb{E}_{(x,y)\sim P} \log \frac{P(x,y)}{P_X(x)P_Y(y)} && (1.23)\\
&= d_{KL}\left[P(x,y) \| P_X(x)P_Y(y)\right] && (1.24)
\end{aligned}$$

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 (1.24), by replacing the KL divergence with a more general divergence d:

$$\mathbb{J}_d(X, Y) = \mathrm{d}\left[P(x,y) \| P_X(x)P_Y(y)\right] \tag{1.25}$$

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.

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

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

$$p(x_d|x_{\neg d}) = \frac{P(x)}{\int P(X_d = y, x_{\neg d})dy} = \frac{C \cdot P(x)}{\int C \cdot P(X_d = y, x_{\neg d})dy} \tag{1.27}$$

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.

$$P(x) = \frac{1}{Z}\exp(x^T W x + b^T x), x \in \{0,1\}^D, \tag{1.28}$$

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:

$$P(x_d|x_{\neg d}, W, b) = \frac{p(x)}{\int p(x_d = y, x_{\neg d})dy} \tag{1.29}$$

$$= \frac{\frac{1}{Z}\exp(x^T W x + b^T x)}{\sum_{x_d\in\{0,1\}} \frac{1}{Z}\exp(x^T W x + b^T x)} \tag{1.30}$$

$$= \frac{\exp(x^T W x + b^T x)}{\sum_{x_d\in\{0,1\}} \exp(x^T W x + b^T x)} \tag{1.31}$$

$$= \frac{\exp\left(x_d\left(W_{d,d} + 2W_{d,\neg d}^T x_{\neg d} + b_d\right)\right)}{\exp(W_{d,d} + 2W_{d,\neg d}^T x_{\neg d} + b_d) + 1} \tag{1.32}$$

$$\tag{1.33}$$

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:

$$S_{\log}(x, p) = -\log P(x) = -\sum_{d=1}^{D} \log P(x_d|x_{1:d-1}) \tag{1.34}$$

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:

$$S_{\text{DLP12}}(x, P) = -\sum_{d=1}^{D} S_d\left(x_d, P_{X_d|X_{\neg d}=x_{\neg d}}\right), \tag{1.35}$$

Where S_d are strictly proper scoring rules for each dimension

1.2.3 The Brier (quadratic) score

Another widely used scoring rule is the so-called *Brier score* or quadratic score, originally introduced in [**?**]. It is used in **TODO: where?**

We will define the Brier score in terms of the L^2 norm of the probability distribution, that is:

$$\|P\|_2 = \sqrt{\mathbb{E}_{x\sim P} P(x)} \tag{1.36}$$

The above definition, albeit slightly informal, makes sense for most classes of probability distributions we are concerned with. For continuous distributions, $P(x)$ denotes the probability density, for discrete distributions $P(x)$ denotes the probability of outcome x. Using this notion we can define the Brier score as follows:

$$\begin{aligned} S_{Brier}(x, P) &= \|P - \delta_x\|_2^2 && (1.37)\\ &= \|P\|_2^2 - 2P(x) + 1 && (1.38)\\ &= \mathbb{E}_{x'\sim P} P(x') - 2P(x) + 1 && (1.39)\\ & && (1.40) \end{aligned}$$

The score gives rise to the following entropy function.

$$\begin{aligned} \mathbb{H}_{Brier}[P] &= \mathbb{E}_{x\sim P}\left[\mathbb{E}_{x'\sim P} P(x') - 2P(x) + 1\right] && (1.41)\\ &= 1 - \mathbb{E}_{x\sim P} P(x) && (1.42)\\ &= 1 - \|P\|_2^2 && (1.43) \end{aligned}$$

For discrete distributions when $\dim \mathcal{X} = D$, the quadratic entropy function is bounded. It's maximum value is attained when P is the D dimensional uniform distribution, then it's maximal value is $1 - \sum_{d=1}^{D} \frac{1}{D^2} = 1 - \frac{1}{D}$. The upper bound is 1 if $\dim \mathcal{X} = \infty$. The entropy function is also non-negative for discrete distributions, with $\mathbb{H}_{Brier}[P] = 0$ only for atomic distributions $P = \delta_{x_0}$.

In uncountable domains, just like Shannon's entropy, The entropy function becomes unbounded from below. For atomic distributions it takes value $-\infty$. Unlike Shannon's entropy, it still is bounded from above.

The Brier divergence function becomes simply the squared norm of the difference between the distribution functions.

$$\begin{aligned} d_{Brier}[P\|Q] &= \mathbb{E}_{x\sim Q}\left[\|P\|_2^2 - 2P(x) + 1\right] - \mathbb{H}_{Brier}[P] && (1.44)\\ &= \|P\|_2^2 - 2\mathbb{E}_{x\sim Q} P(x) + \|P\|_2^2 && (1.45)\\ &= \|P\|_2^2 - 2\langle P, Q\rangle + \|Q\|_2^2 && (1.46)\\ &= \|P - Q\|_2^2 && (1.47) \end{aligned}$$

The value of information under the Brier score becomes the following straightforward quantity.

$$\begin{aligned} \mathbb{I}_{Brier}[X \leftarrow Y] &= \mathbb{E}_{y\sim P_Y} \|P_X - P_{X|Y=y}\|_2^2 && (1.48)\\ & && (1.49) \end{aligned}$$

1.2.4 Spherical and pseudo-spherical scoring rules

Another example of strictly proper scoring rules mentioned in [**?**, **?**], introduced in [**?**, **?**]. The spherical score is defined as follows.

$$S_{spherical}(x, P) = 1 - \frac{P(x)}{\|P\|_2} \tag{1.50}$$

This gives rise to the following entropy and divergence functions.

$$\mathbb{H}_{spherical}[P] = 1 - \mathbb{E}_{x\sim P}\frac{P(x)}{\|P\|_2} \tag{1.51}$$

$$= 1 - \|P\|_2 \tag{1.52}$$

$$d_{spherical}[P\|Q] = -\mathbb{E}_{x\sim P}\frac{Q(x)}{\|Q\|_2} + \|P\|_2 \tag{1.53}$$

$$= \|P\|_2 - \frac{\langle Q, P\rangle}{\|Q\|_2} \tag{1.54}$$

$$= \|P\|_2 \left(1 - \cos(P, Q)\right), \tag{1.55}$$

where $\cos(P, Q) = \frac{\langle P,Q\rangle}{\|P\|_2\|Q\|_2}$ is the cosine similarity between P and Q.

An interesting property of the spherical score is that it is agnostic to scaling of P. That is $S_{spherical}(x, c \cdot P) = S_{spherical}(x, P)$. Similarly, $d_{spherical}[P\|c \cdot Q] = d_{spherical}[P\|Q]$. and $d_{spherical}[c \cdot P\|Q] = c \cdot d_{spherical}[P\|Q]$. This means that when approximating P by Q via minimising $d_{spherical}[P\|Q]$ we only need to know P and Q up to a normalising constant.

The value of information under the spherical score is

$$\mathbb{I}_{spherical}[X \leftarrow Y] = \|P_X\|_2 \mathbb{E}_{y\sim P_Y}\left(1 - \cos(P_X, P_{X|Y=y})\right) \tag{1.56}$$

Pseudo-spherical scoring rules are generalisations of the spherical score of the form: **TODO: check if formulæare correct**.

$$\mathbb{H}_{\gamma,pseudospherical}[P] = -\|P\|_\gamma \tag{1.57}$$

For more details about pseudospherical scores see [].

1.2.5 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:

$$d_{\mathcal{F}}[p\|q] = \sup_{f\in\mathcal{F}}\left|\int f(x)p(x)dx - \int f(x)q(x)dx\right| \tag{1.58}$$

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 [**?**]:

$$
\begin{aligned}
d_k\left[P\|Q\right] &:= \mathrm{MMD}^2\left(P, Q\right) && (1.59)\\
&= \sup_{\substack{f\in\mathcal{H}\\ \|f\|_{\mathcal{H}}=1}} \left(\mathbb{E}_{x\sim P} f(x) - \mathbb{E}_{x\sim Q} f(x)\right)^2 && (1.60)\\
&= \sup_{\substack{f\in\mathcal{H}\\ \|f\|_{\mathcal{H}}=1}} \left|\mathbb{E}_{x\sim P} \left\langle f, k(\cdot, x)\right\rangle - \mathbb{E}_{x\sim Q} \left\langle f, k(\cdot, x)\right\rangle\right|^2 && (1.61)\\
&= \sup_{\substack{f\in\mathcal{H}\\ \|f\|_{\mathcal{H}}=1}} \left|\left\langle f, \mathbb{E}_{x\sim P} k(\cdot, x) - \mathbb{E}_{x\sim Q} \int k(\cdot, x)\right\rangle\right|^2 && (1.62)\\
&= \sup_{\substack{f\in\mathcal{H}\\ \|f\|_{\mathcal{H}}=1}} \left\langle f, \mu_P - \mu_Q\right\rangle^2 && (1.63)\\
&= \left\|\mu_P - \mu_Q\right\|_{\mathcal{H}}^2 && (1.64)\\
&= \mathbb{E}_{x,x'\sim P} k(x, x') - 2\mathbb{E}_{x\sim P}\mathbb{E}_{x'\sim Q} k(x, x') + \mathbb{E}_{x,x'\sim Q} k(x, x'), && (1.65)
\end{aligned}
$$

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 $d_k\left[P\|Q\right] = 0 \iff Q = P$ 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:

$$
\phi_p(t) = \mathbb{E}_{x\sim p}\left[e^{itx}\right] = \int e^{itx} p(x)dx, \tag{1.66}
$$

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 1.65 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.

The squared MMD in fact conforms to our definition of a generalised divergence in equation (1.2), and corresponds to the following scoring rule:

$$S_k(x, P) := k(x, x) - 2\mathbb{E}_{x' \sim P} k(x, x') + \mathbb{E}_{x', x'' \sim Q} k(x', x'') \tag{1.67}$$

$$= \sup_{\substack{f \in \mathcal{H} \\ \|f\|_{\mathcal{H}} = 1}} \left(f(x) - \mathbb{E}_{x \sim Q} f(y)\right)^2 \tag{1.68}$$

This scoring rule is analogous to the kernel scoring rule introduced originally in [**?**]. The difference is scaling by a factor of two, and the leading $k(x, x)$ term. These differences do not make any practical difference: scoring rules that are equal up to scaling and an additive term that depends only on x but not on the distribution P give rise to exactly the same generalised entropy and divergence functionals.

[**?**] give a proof of the propriety of this scoring rule for Borel probability measures for which the expectation $\mathbb{E}_{x, x' \sim P} k(x, x')$ is finite. The scoring rule is also strictly proper whenever the kernel is characteristic [**?**]. [**?**] showed particular examples of scoring rules, among them the Brier score (see section **??**), that can be interpreted as special cases of the kernel scoring rule.

The generalised entropy defined by this scoring rule becomes:

$$\mathbb{H}_k [P] := \mathbb{E}_{x \sim P} k(x, x) - \mathbb{E}_{x, x' \sim P} k(x, x') \tag{1.69}$$

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 bounded, then the entropy functional is also bounded. If we further assume that the kernel satisfies $\forall x, y : k(x, x) \geq k(x, y)$, then the score is also non-negative. Irrespective of kernel choice, the entropy is zero for delta distributions, that is when the distribution Q is concentrated on a single point. If the kernel satisfies the strict inequality $\forall x, y : k(x, x) > k(x, y)$, the kernel is positive for any other distribution.

Secondly, The only requirement for the distribution Q is that we can compute expectations with respect to it. This means that any probability distribution, and indeed any 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. This property is going to be useful in applications to quasi-Monte Carlo in chapter **??**.

Thirdly, the entropy function has the kernel as free parameter, which is a mixed blessing. On one hand, this provides extra flexibility: even if we commit to a particular family of kernels, like the square exponential, we can fine-tune the entropy function to our needs by adjusting parameters, such as the length-scale parameter [**?**]. On the other hand there is no principled, general 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 squared maximum mean discrepancy defined in equation (1.60).

$$d_{S_k} [P \| Q] = \mathbb{E}_{x \sim P} [S_k(x, Q)] - \mathbb{E}_{x \sim P} [S_k(x, Q)] \tag{1.70}$$

$$= \mathbb{E}_{x \sim P} k(x, x) - 2\mathbb{E}_{x \sim P} \mathbb{E}_{x' \sim Q} k(x, x') + \mathbb{E}_{x, x' \sim Q} k(x, x') \tag{1.71}$$

$$- \left(\mathbb{E}_{x \sim P} k(x, x) - \mathbb{E}_{x, x' \sim P} k(x, x')\right) \tag{1.72}$$

$$= \mathbb{E}_{x, x' \sim P} k(x, x') - 2\mathbb{E}_{x \sim P} \mathbb{E}_{x' \sim Q} k(x, x') + \mathbb{E}_{x, x' \sim Q} k(x, x') \tag{1.73}$$

$$= d_k [P \| Q] \tag{1.74}$$

It is easy to show that the Brier (quadratic) score is a special case of the kernel score when the kernel is chosen to be the trivial $k(x, x') = \delta(x - x')$, where δ is the Dirac delta function. This insight allows us to understand why the Brier score is so impoverished when applied to continuous domains $\mathcal{X}$ such as the real line $\mathbb{R}$: Just as the KL divergence, it does not incorporate any notion of smoothness or similarity of neighbouring points. Two point masses on neighbouring points x and $x + \epsilon$ are maximally dissimilar, irrespective of how small the difference ϵ is. The kernel scoring rule

overcomes this strict limitation by allowing us to engineer a kernel with appropriate smoothness assumptions built in.

This makes it particularly hard to estimate Brier divergences from sampled data.

We can use the generalised entropy and divergence defined by the kernel scoring rule to define the value of information a random variable provides about another one:

$$\begin{aligned}
\mathbb{I}_k\left[X \leftarrow Y\right] &= \mathbb{E}_{y\sim P_Y} d_k\left[P_X \middle\| P_{X|Y=y}\right] && (1.75)\\
&= \mathbb{E}_{y\sim P_Y} \|\mu_{X|Y=y} - \mu_X\|^2_{\mathcal{H}} && (1.76)\\
&= k(P_X, P_X) - 2 * \mathbb{E}_{y\sim P_Y} k(P_X, P_{X|Y=y}) + \mathbb{E}_{y\sim P_Y} k(P_{X|Y=y}, P_{X|Y=y}) && (1.77)\\
&= \mathbb{E}_{y\sim P_Y} \mathbb{E}_{P_{x_1,x_2\sim P_{X|Y=y}}} k(x_1, x_2) - \mathbb{E}_{x_1,x_2\sim P_X} k(x_1, x_2) && (1.78)
\end{aligned}$$

To my knowledge, this kernel-based measure of information has not been used in the machine learning or statistics literature before. It is interesting to contrast this to other kernel measures of dependence developed recently in statistics. These are largely based on the cross-covariance operator between Hilbert space embedding of the two distributions.

Definition 6 (kernel Cross-covariance operator)**.** *Let X and Y be two random variables with joint distribution $P \in \mathcal{M}^1_{\mathcal{X}\times\mathcal{Y}}$, and marginals P_X and P_Y. Let $k_{\mathcal{X}} : \mathcal{X}\times\mathcal{X} \mapsto \mathbb{C}$ and $k_{\mathcal{Y}} : \mathcal{Y}\times\mathcal{Y} \mapsto \mathbb{C}$ be positive definite kernels with associated reproducing kernel Hilbert spaces $\mathcal{H}_{\mathcal{X}}$ and $\mathcal{H}_{\mathcal{Y}}$, respectively. Let us define the kernel cross-covariance operator C_{XY} between X and Y so that for all $f \in \mathcal{H}_{\mathcal{X}}$ and $g \in \mathcal{H}_{\mathcal{Y}}$*

$$\langle f, C_{XY} g\rangle_{\mathcal{H}_{\mathcal{X}}} = \mathbb{E}_{(x,y)\sim P}\left(f(x) - \mathbb{E}_{x'\sim P_X} f(x')\right)\left(g(y) - \mathbb{E}_{y'\sim P_Y} g(y')\right) \tag{1.79}$$

TODO: Figure out if there is any connection between COCO and my criterion, maybe with a stupid choice of $k_{\mathcal{Y}}$ Based on the cross-covariance operator, we can define various measures of dependence or information, the simplest of which is constrained covariance, or COCO:

Definition 7 (Constrained covariance, see [**?**, **?**])**.** *In the same notation as above let us define define the constrained covariance between X and Y, $COCO_{XY}$, as*

$$COCO_{XY} = \sup_{\substack{f\in\mathcal{H}_{\mathcal{X}}, g\in\mathcal{H}_{\mathcal{Y}}\\ \|f\|_{\mathcal{H}_{\mathcal{X}}}=1, \|g\|_{\mathcal{H}_{\mathcal{Y}}}=1}} \mathbb{C}ov_{(x,y)\sim P}\left[f(x), g(y)\right] \tag{1.80}$$

It can be shown that, $COCO$ is the matrix norm of the cross-covariance operator:

$$COCO_{XY} = \|C_{XY}\|_2, \tag{1.81}$$

where $\|\cdot\|_2$ denotes the matrix norm, that is the modulus of largest eigenvalue. ***TODO: Check if statements are correct***

COCO is only one of several independence measures that have been developed based on kernels, see [**?**] for an overview of variants. Just as generalisations of Shannon's mutual information (eqn. (1.25)), these measures of dependence have several useful properties. They are symmetric, and can be effectively estimated from empirical data [].

However, as with eqn. (1.25), COCO and its variants do not have an interpretation as "the extent to which knowing Y is useful for predicting X". Also, they require a kernel on both $\mathcal{X}$ and $\mathcal{Y}$, and properties of the functional depend on both choices of kernels. In contrast (1.25) only requires a kernel over X.

The diversity operator TODO: find a better name for it

The kernel value of information $\mathbb{I}_k\left[X \leftarrow Y\right]$ can also be interpreted as the norm of an operator, that we shall name the diversity operator.

Definition 8. *Given two random variables X and Y with joint distribution P, and a positive definite kernel $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{C}$ with associated Hilbert space $\mathcal{H}$, let us define the 'diversity operator' of Y over X, $D_{X|Y} : \mathcal{H} \mapsto \mathcal{H}$ such that for all $f, g \in \mathcal{H}$*

$$\left\langle f, D_{X|Y} g \right\rangle_{\mathcal{H}} = \mathbb{C}ov_{y \sim P_Y} \left[\mathbb{E}_{X|Y=y} f, \mathbb{E}_{X|Y=y} g \right] \tag{1.82}$$

Consequently for all $f \in \mathcal{H}$

$$\left\langle f, D_{X|Y} f \right\rangle = \mathbb{V}_{y \sim P_Y} \left[\mathbb{E}_{x \sim P_{X|Y=y}} f(x) \right] \tag{1.83}$$

Statement 2 (Alternative definition of $D_{X|Y}$)**.** *$D_{X|Y}$ admits the following equivalent definition*

$$D_{X|Y} = \mathbb{E}_{y \sim P_Y} \left(\mu_{X|Y=y} - \mu_X \right) \otimes \left(\mu_{X|Y=y} - \mu_X \right) \tag{1.84}$$

Proof. Let $f, g \in \mathcal{H}$, then

$$\begin{aligned}
&\left\langle f, \left(\mathbb{E}_{y \sim P_Y} \left(\mu_{X|Y=y} - \mu_X \right) \otimes \left(\mu_{X|Y=y} - \mu_X \right) \right) g \right\rangle && (1.85)\\
=&\mathbb{E}_{y \sim P_Y} \left\langle f, \left(\left(\mu_{X|Y=y} - \mu_X \right) \otimes \left(\mu_{X|Y=y} - \mu_X \right) \right) g \right\rangle && (1.86)\\
=&\left\langle f, \left(\mu_{X|Y=y} - \mu_X \right) \right\rangle \left\langle g, \left(\mu_{X|Y=y} - \mu_X \right) \right\rangle && (1.87)\\
=&\mathbb{E}_{y \sim P_Y} \left(\mathbb{E}_{X|Y=y} f(x) - \mathbb{E}_{x \sim P_X} f(x) \right) \left(\mathbb{E}_{X|Y=y} g(x) - \mathbb{E}_{x \sim P_X} g(x) \right) && (1.88)\\
=&\mathbb{C}\mathrm{ov}_{y \sim P_Y} \left[\mathbb{E}_{X|Y=y} f, \mathbb{E}_{X|Y=y} g \right] && (1.89)
\end{aligned}$$

□

The information is the trace of this operator:

$$\begin{aligned}
\mathbb{I}_k \left[X \leftarrow Y \right] &= \mathbb{E}_{y \sim P_Y} \| P_X - P_{X|Y=y} \|_2^2 && (1.90)\\
&= \mathbb{E}_{y \sim P_Y} \mathrm{trace} \left\langle P_X - P_{X|Y=y}, P_X - P_{X|Y=y} \right\rangle && (1.91)\\
&= \mathbb{E}_{y \sim P_Y} \mathrm{trace} \left(P_X - P_{X|Y=y} \right) \otimes \left(P_X - P_{X|Y=y} \right) && (1.92)\\
&= \mathrm{trace}\, \mathbb{E}_{y \sim P_Y} \left(P_X - P_{X|Y=y} \right) \otimes \left(P_X - P_{X|Y=y} \right) && (1.93)\\
&= \mathrm{trace}\, I_{X|Y} && (1.94)
\end{aligned}$$

See definition of HSIC in [**?**]

1.2.6 The spherical kernel score

Seeing how the Brier score is a special case of the kernel scoring rule, one might wonder whether the spherical scoring rule has a similar generalisation. It turns out it does, and it gives rise to a very intuitive divergence. Consider the following scoring rule

$$\begin{aligned}
S_{k,spherical}(x, P) &:= \| \mu_{\delta_x} \|_{\mathcal{H}} - \frac{\mu_P(x)}{\| \mu_P \|_{\mathcal{H}}} && (1.95)\\
&= \| \mu_{\delta_x} \|_{\mathcal{H}} \left(1 - \cos(\mu_{\delta_x}, \mu_P) \right) && (1.96)\\
&= \sqrt{k(x,x)} - \frac{\mathbb{E}_{x' \sim P} k(x, x')}{\sqrt{\mathbb{E}_{x,x' \sim P} k(x, x')}}, && (1.97)
\end{aligned}$$

The scoring rule gives rise to the following entropy functional:

$$\begin{aligned}
\mathbb{H}_{k,spherical} \left[P \right] &= \mathbb{E}_{x \sim P} \| \mu_{\delta_x} \|_{\mathcal{H}} - \| \mu_P \|_{\mathcal{H}} && (1.98)\\
&= \mathbb{E}_{x \sim P} \sqrt{k(x,x)} - \sqrt{\mathbb{E}_{x,x' \sim P} k(x, x')} && (1.99)
\end{aligned}$$

Whenever $k(x,x) = c$ this entropy is non-negative, and bounded from above. For characteristic kernels it is only zero for delta distributions. The entropy is very scoring rule leads to the following divergence:

$$d_{k,spherical}\left[P\|Q\right] = -\mathbb{E}_{x\sim Q}\frac{\mu_P}{\|\mu_P\|_{\mathcal{H}}} + \|\mu_P\|_{\mathcal{H}} \tag{1.100}$$

$$= \|\mu_P\|_{\mathcal{H}}\left(1 - \cos(\mu_P, \mu_Q)\right) \tag{1.101}$$

$$= \sqrt{\mathbb{E}_{x,x'\sim P}k(x,x')} - \frac{\mathbb{E}_{x\sim P}\mathbb{E}_{x'\sim Q}k(x,x')}{\sqrt{\mathbb{E}_{x,x'\sim Q}k(x,x')}} \tag{1.102}$$

Unlike MMD and $d_k\left[\cdot\|\cdot\right]$, this divergence is asymmetric because of the leading $\|\mu_P\|_{\mathcal{H}}$ factor. Also, just like the spherical score, it is agnostic to scaling of Q, that is $d_{k,spherical}\left[P\|c\cdot Q\right] = d_{k,spherical}\left[P\|Q\right]$. Furthermore, $d_{k,spherical}\left[c\cdot P\|Q\right] = c\cdot d_{k,spherical}\left[P\|Q\right]$. Whenever the kernel is characteristic, this scoring rule is strictly proper with respect to Borel probability distributions, whose mean embedding $\mu_P(x)$ is bounded.

Theorem 2 (The spherical kernel score is strictly proper)**.** *Proof.* Suppose $P \neq Q$, then by the strict propriety of the kernel score

$$0 < d_k\left[P\|Q\right] \tag{1.103}$$

$$0 < \|\mu_P\|_{\mathcal{H}}^2 + \|\mu_Q\|_{\mathcal{H}}^2 - 2\left\langle\mu_P, \mu_Q\right\rangle_{\mathcal{H}} \tag{1.104}$$

$$\left\langle\mu_P, \mu_Q\right\rangle < \frac{1}{2}\left(\|\mu_P\|_{\mathcal{H}}^2 + \|\mu_Q\|_{\mathcal{H}}^2\right) \leq \|\mu_P\|_{\mathcal{H}}\|\mu_Q\|_{\mathcal{H}} \tag{1.105}$$

$$\cos(\mu_P, \mu_Q) = \frac{\left\langle\mu_P, \mu_Q\right\rangle_{\mathcal{H}}}{\|\mu_P\|_{\mathcal{H}}\|\mu_Q\|_{\mathcal{H}}} < 1 \tag{1.106}$$

Thus,

$$d_{k,spherical}\left[P\|Q\right] = \|\mu_P\|_{\mathcal{H}}\left(1 - \cos(\mu_P, \mu_Q)\right) \geq 0 \tag{1.107}$$

□

Just as it is the case with the Brier score and the kernel scoring rule, the spherical kernel rule reduces to the spherical score whenever the trivial kernel $k(x,x') = \delta_x(x')$ is used.

TODO: Figure out if this interpretation is useful:

$$1 - \cos(\mu_P, \mu_Q) = \mathbb{E}_{f\sim GP}\mathbb{1}\left[\text{sign}(\mathbb{E}_{x\sim P}f(x)) = \text{sign}(\mathbb{E}_{x\sim Q}f(x))\right] \tag{1.108}$$

1.2.7 Scoring rules and 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 $X = x$, and we choose action a, we incur a loss $\ell(x,a)$. Let us assume we have a probabilistic forecast or belief P about the state of the environment X. Given this we can choose an action that minimises the expected loss:

$$a_P^* = \operatorname*{argmin}_{a\in\mathcal{A}} \mathbb{E}_{x\sim P}\ell(x,a) \tag{1.109}$$

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 $X = x$.

$$S_\ell(x, P) = \ell(x, a^*_P) \tag{1.110}$$

As scoring rule of this form defines a generalised entropy, otherwise known as the Bayes-risk of the decision problem:

$$\begin{aligned} \mathbb{H}_\ell[P] :=& \mathbb{E}_{x\sim P}\ell(x, a^*_P) && (1.111)\\ =& \min_{a\in\mathcal{A}} \mathbb{E}_{x\sim P}\ell(x, a) && (1.112)\end{aligned}$$

The associated divergence can be interpreted as the excess loss we incur by using the suboptimal action a^*_Q computed on the basis of Q, when in fact the true distribution of X is P:

$$d_\ell[P\|Q] = \mathbb{E}_{x\sim P}\ell(x, a^*_Q) - \min_{a\in\mathcal{A}} \mathbb{E}_{x\sim P}\ell(x, a) \tag{1.113}$$

Several scoring rules can be interpreted as special cases of this loss-calibrated framework.

Logarithmic score and Shannon entropy

Shannon's entropy has an intuitive operational meaning as minimum description length. We are given a random variable X with distribution P over a finite, discrete dictionary $\mathcal{X}$. We would like to encode symbols in $\mathcal{X}$ by binary sequences, in such a way, that any sequence composed by concatenating codewords is uniquely decodable. It can be shown that the expected code-length of any uniquely decodable code $f : \mathcal{X} \mapsto \{0, 1\}^*$ under the distribution P is lower bounded by the entropy of P:

$$\mathbb{E}_{x\sim P}|f(x)| \geq \mathbb{H}_{Shannon}[P]. \tag{1.114}$$

TODO: define logarithmic score with base 2 logarithm so that this makes sense

Let us consider the following decision problem: Let $\mathcal{A}$ be the set of all uniquely decodable binary codes, so that $a : \mathcal{X} \mapsto \{0, 1\}^*$ maps X to a binary codeword of variable length. Let the loss ℓ be the length of the codeword assigned to X: $\ell(x, a) = |a(x)|$.

The scoring rule defined by this decision problem is approximately the same as the logarithmic score.

Kernel scoring rule

Let's say your task is to estimate value of a set of functions $f \in \mathcal{F}$ evaluated at X. The action can be interpreted as a functional $a : \mathcal{F} \mapsto \mathbb{R}$, that gives an estimated value of $f(X)$ 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.

$$\ell(\theta, a) \sup_{f\in\mathcal{F}} (f(\theta) - a(f))^2 \tag{1.115}$$

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

$$a^*_P = \mathbb{E}_{x\sim P} f(x) \tag{1.116}$$

Thus, we can define the following scoring rule S:

$$S(\theta, P) = \ell(x, a^*_P) \sup_{f\in\mathcal{F}} \left(f(\theta) - \int f(\theta)p(\theta)\, d\theta \right)^2 \tag{1.117}$$

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.

As the Brier score is a special case of the kernel scoring rule, it can also be derived from the same decision problem.

Chapter 2

Information geometry

Riemannian manifold, local metric, geodesic distance .

In this section I aim to develop an understanding of the differences between various scoring rules and corresponding divergence metrics by visualising the geometric structure they give rise to.

TODO: A sentence about manifolds here Our goal is to create a low-dimensional map of particular manifolds in such a way that distances measured between points on the map correspond to distances measured in the abstract geometry the scoring rule defines 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 Riemannian manifold, which can be parametrised by two parameters, 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.

Having understood that a perfect map of two-dimensional statistical manifolds cannot necessarily be produced, I 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.

2.1 Riemannian geometry

Strictly proper scoring rules and their associated divergence functions induce a geometry over probability distributions, that we will call the information geometry. Under suitable smoothness assumptions, probability distributions form a smooth Riemannian manifold [**?**, **?**], on which the squared local distance is

$$ds^2 = \frac{1}{2}\left\langle P, \ddot{H}(P)P \right\rangle, \tag{2.1}$$

Where $\ddot{H(P)}$ is the Hamiltonian of the entropy $H(P) = \mathbb{H}_S$ at P. For discrete distributions, when $\mathcal{X} = 1, 2, \ldots$, denoting $p_i := P(X = i)$ we can write this squared distance as

$$ds^2 = \frac{1}{2}\sum_{i,j} \frac{\partial^2 H}{\partial p_i \partial p_j} dp_i dp_j. \tag{2.2}$$

Definition 9 (Riemannian geodesic)**.** *Let P_1 and P_2 be two probability distributions and d_H a Bregman divergence. Let $\mathcal{P} = \{P(t), t \in [0,1]\}$ a smooth, differentiable path on the manifold such that $P(0) = P_1$ and $P(1) = P_2$. The length of the curve $\mathcal{P}$ is defined as*

$$l(\mathcal{P}) = \int_0^1 \sqrt{\left\langle \dot{P}(t), \ddot{H}(P(t))\dot{P}(t) \right\rangle} dt \tag{2.3}$$

A Riemannian geodesic between P_1 and P_2 is a path, whose length is minimal. The length of such a path is called the geodesic distance between P_1 and P_2.

Geodesic distances in non-trivial, general Riemannian manifolds are hard to compute analytically. There are two main technical difficulties that arise:

1. The integral defining the Riemannian length of a given path (eqn. (2.3)) can be hard to compute analytically, even if an analytical expression for the local squared distance ds^2 exists.

2. The geodesic distance between P and Q is the minimum of the length of any path that connects P and Q. This minimisation over all paths is a non-trivial one and is very hard to carry out exactly, even if an analytical expression for the length existed.

Therefore in the following chapter I will resort to numerical approximations. To solve the first problem, I note that geodesic distances between close distributions P and $P + dP$ are the same as the local distance ds. Furthermore the local distance can be approximated by computing the square root of the symmetrised divergence between the two points, using the following observations.

Statement 3 (Taylor expansion of Bregman divergences)**.** *Let $H : \Theta \mapsto \mathbb{R}$ be a smooth, strictly concave function and $d_H[\cdot\|\cdot]$ the Bregman divergence it induces. For infinitesimally small $dP \in \Theta$ the following approximation holds:*

$$d_H[P\|P + dP] = \frac{1}{2}\sum_{i,j} \frac{\partial^2 \mathbb{H}_S}{\partial p_i \partial p_j} dp_i dp_j = d_H[P + dP\|P] \tag{2.4}$$

Proof. We prove the left-hand equation first:

$$\frac{\partial}{\partial q_i} d_H[P\|Q] = -\frac{\partial}{\partial q_i} H(Q) + \frac{\partial}{\partial q_i} \langle \nabla_Q H(Q), Q - P \rangle \tag{2.5}$$

$$= -\frac{\partial}{\partial q_i} H(Q) + \frac{\partial}{\partial q_i} \sum_j \frac{\partial}{\partial q_j} H(Q)(q_j - p_j) \tag{2.6}$$

$$= -\frac{\partial}{\partial q_i} H(Q) + \frac{\partial}{\partial q_i} H(Q) + \sum_j \frac{\partial^2}{\partial q_i \partial q_j} H(Q)(q_j - p_j) \tag{2.7}$$

$$= \sum_j \frac{\partial^2}{\partial q_i \partial q_j} H(Q)(q_j - p_j) \tag{2.8}$$

hence by first order Taylor expansion around P:

$$d_H[P\|P + dP] \approx d_H[P\|P + dP] + \frac{1}{2}\left\langle dP, \nabla_Q d_H[P\|Q]|_{Q=P} \right\rangle \tag{2.9}$$

$$= \frac{1}{2}\sum_i dp_i \sum_j \frac{\partial^2 H(P)}{\partial p_i \partial p_j} dp_j \tag{2.10}$$

$$= \frac{1}{2}\sum_{i,j} \frac{\partial^2 H(P)}{\partial p_i \partial p_j} dp_i dp_j \tag{2.11}$$

Similarly in the other direction

$$\frac{\partial}{\partial q_i} d_H\left[Q\|P\right] = \frac{\partial}{\partial q_i} H(Q) + \frac{\partial}{\partial q_i}\left\langle \nabla_P H(P), P - Q\right\rangle \tag{2.12}$$

$$= \frac{\partial}{\partial q_i} H(Q) + \frac{\partial}{\partial q_i}\sum_j \frac{\partial}{\partial p_j} H(P)(p_j - q_j) \tag{2.13}$$

$$= \frac{\partial}{\partial q_i} H(Q) - \frac{\partial}{\partial p_i} H(P) \tag{2.14}$$

$$= \dot{H}(Q) - \dot{H}(P) \tag{2.15}$$

Note that for small deviation dP, the derivative can be written as

$$\frac{\partial}{\partial dP} d_H\left[P + dP\|P\right] = \dot{H}(P + dP) - \dot{H}(P) \approx \ddot{H}(P)dP \tag{2.16}$$

therefore, via Taylor expansion we get that for small dP

$$d_H\left[P\|P + dP\right] \approx \frac{1}{2}\left\langle dP, \ddot{H}(P)dP\right\rangle \tag{2.17}$$

$$= \frac{1}{2}\sum_{i,j} \frac{\partial^2 H(P)}{\partial p_i \partial p_j} dp_i dp_j \tag{2.18}$$

□

Corollary 3 (Local approximation to geodesic distance)**.** *The geodesic distance between distributions P and Q on the statistical manifold defined by the scoring rule S can be approximated as follows.*

$$\textit{distance}(P, Q) \approx \sqrt{d_S\left[P\|Q\right] + d_S\left[Q\|P\right]} \tag{2.19}$$

Now that we have a decent local approximation to geodesic distances, we can approximate the length of longer, smooth paths on the manifold by approximating the path as a series of small segments, and then using the above approximation to compute the length of each segment.

Now we have to solve the second problem of approximating the minimisation over all paths between P and Q. To do this, we will use ideas from ISOMAP, an isometric embedding technique developed in the machine learning community [**?**]: we restrict the paths under consideration and only compute the minimum among those paths that travel trough a sequence of neighbouring points from a predefined set. This approximate shortest path can be computed in polynomial time in the number of nodes in the neighbourhood graph. The length of the paths can be approximated as the neighbours are assumed to be close enough for the local approximation to work.

We will follow the following procedure to produce a map of the statistical manifold induced by scoring rules.

1. take a set of probability distributions, preferably such that they relatively densely cover an interesting region on the manifold. In most cases we will choose a square grid in an appropriately chosen parameter-space.

2. compute approximate geodesics:

   (a) construct a graph over the sampled distributions as nodes, such that we draw edges between each distribution and its k nearest neighbours. The weight of each edge being the squareroot of the symmetrised divergence between the two distributions, as in Eqn. (2.19).

   (b) compute the shortest path on the resulting graph between every pair of points on the graph

3. use metric multidimensional scaling to embed the set of distributions as points a low-dimensional Euclidean space.

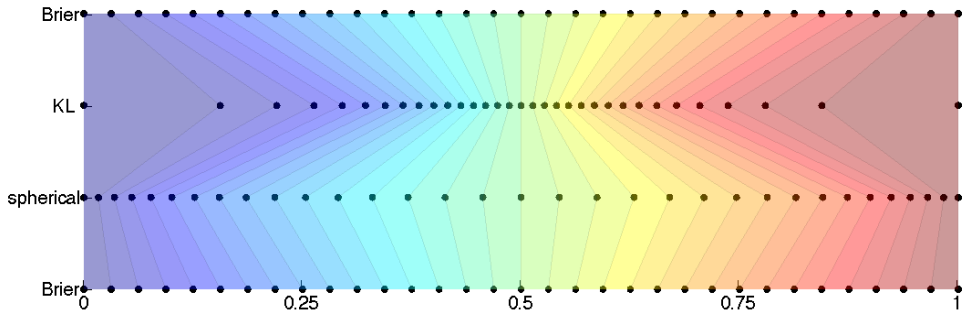

Figure 2.1: **TODO: to be replaced with nicer graphics** Illustration of the differences between the Brier, spherical, and KL divergences between Bernoulli distributions. The Brier divergence is equivalent to the Euclidean distance between parameter values, therefore when mapped by Brier divergence, parameters are evenly spaced on the real line *(top, bottom)*. The KL divergence places emphasis on small differences between small probabilities, therefore the manifold is stretched out as the parameter approaches 0 and 1. In fact the KL divergence is not bounded, and the manifold of Bernoulli distributions stretches to the entire real line.

2.1.1 Bernoulli distributions

Let us first look at the simple and special case of one dimensional statistical manifolds of Bernoulli distributions (biased coinflips). Bernoulli random variables have a binary outcome, positive with probability p and negative with probability $1 - p$.

One dimensional Riemannian manifolds are special, as these are always homeomorphic to either the real line $\mathbb{R}$, or the circle. In addition, one dimensional statistical manifolds induced by strictly proper scoring rules are always homeomorphic to the real line. So the only difference between various manifolds is how to real line is stretched and compressed at various parts.

2.1.2 Visualising the Shannon-information geometry

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

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:

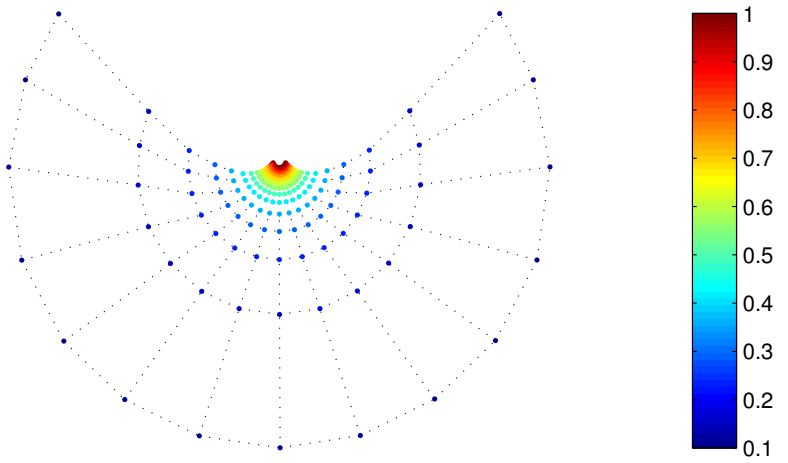

Figure 2.2: Map of Normal distributions on the statistical manifold defined by the log-score, or KL divergence. Dots of the same color show Normal distributions with the same standard deviation. It can be clearly seen that distributions with lower standard deviation are spread out more than those with a higher standard deviation, giving rise to a fan-like structure.

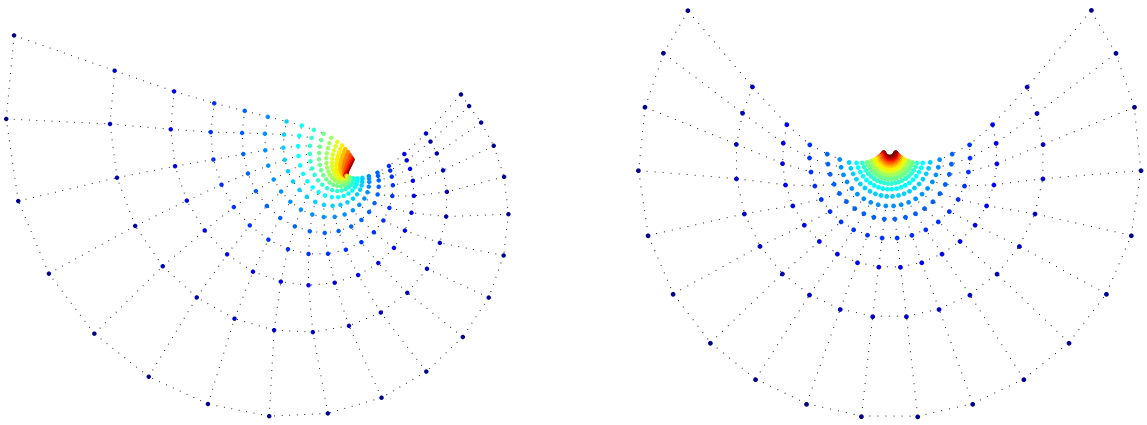

Figure 2.3: Comparing the statistical manifolds defined by the KL divergence between Gamma *(left)* and Normal *(right)* distributions. Here the mean and variance of each point on the left is matched to the corresponding point on the right (see text for details). For large values of variance (yellow and red) the two manifolds are very dissimilar, the Gaussian one is symmetric, while the Gamma one is asymmetric. However, as variance decreases (blue), by the central limit theorem Gamma distributions are more alike Gaussians of the same mean and variance, thus the manifold conforms to the fan-like shape that is characteristic of Gaussian distributions.

$$d_{KL}\left[\mathcal{N}_{\mu_1,\sigma_1}\|\mathcal{N}_{\mu_2,\sigma_2}\right] = \frac{\left(\mu_1-\mu_2\right)^2}{2\sigma_2^2} + \frac{1}{2}\left(\frac{\sigma_1^2}{\sigma_2^2} - 1 - \log\frac{\sigma_1^2}{\sigma_2^2}\right) \tag{2.20}$$

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:

$$(x) = \beta^\alpha \frac{1}{\Gamma(\alpha)} x^{\alpha-1}\exp(-\beta x) \tag{2.21}$$

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:

$$d_{KL}\left[\Gamma_{\alpha_1,\beta_1}\|\Gamma_{\alpha_2,\beta_2}\right] = \left(\alpha_1-\alpha_2\right)\psi\left(\alpha_1\right) - \log\frac{\Gamma(\alpha_1)}{\Gamma(\alpha_2)} + \alpha_1\log\frac{\beta_1}{\beta_2} + \alpha_1\frac{\beta_2-\beta_1}{\beta_1} \tag{2.22}$$

Figure **??** shows the manifold of Gamma distributions for parameters $a \le \alpha \le b, c \le \beta \le d$. 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:

$$\mu = \frac{\alpha}{\beta} \tag{2.23}$$

$$\sigma^2 = \frac{\alpha}{\beta^2} \tag{2.24}$$

Solving for α and β in these equations we get

$$\alpha = \frac{\mu^2}{\sigma^2} \tag{2.25}$$

$$\beta = \frac{\mu}{\sigma^2} \tag{2.26}$$

Plugging these into Eqn. (2.22) 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. (2.22) becomes numerically imprecise, as it relies on look-up-table implementation of the Gamma (Γ) and bigamma (ψ) functions.

2.1.3 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

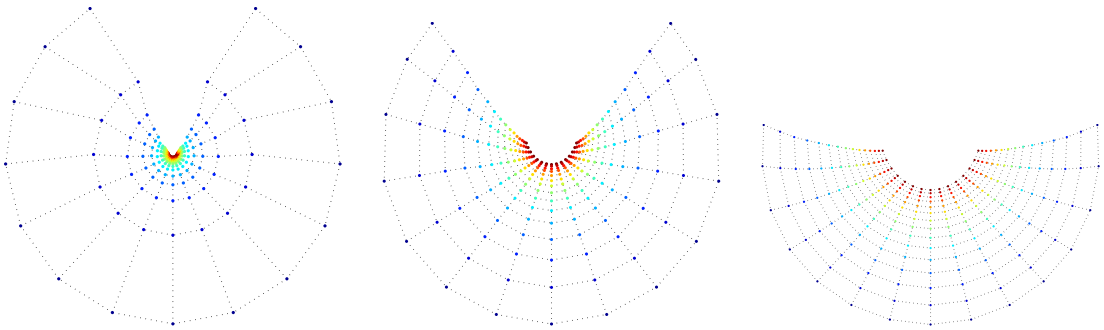

Figure 2.4: Map of the statistical manifold corresponding to the kernel score with a square exponential kernel. The lengthscale parameter was set to *(from left to right)* $\ell = 0.5, 2, 5$. For small lengthscales, the divergence is very sensitive to distributions with small standard deviation. As the lengthscale increases, this sensitivity is less and less dominant, and we start to observe comparable resolution amongst distributions with larger variance.

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:

$$\left\langle \mu_{\mathcal{N}(\mu_1,\sigma_1)}, \mu_{\mathcal{N}(\mu_2,\sigma_2)} \right\rangle_{k_\ell} = \frac{\ell}{\sqrt{\ell^2 + \sigma_1^2 + \sigma_2^2}} \exp\left(-\frac{(\mu_1 - \mu_2)^2}{2(\ell^2 + \sigma_1^2 + \sigma_2^2)} \right) \tag{2.27}$$

TODO: correct expression below!!!

$$d_k\left[\mathcal{N}_{\mu_1,\sigma_1} \| \mathcal{N}_{\mu_2,\sigma_2}\right] = \ell \left(\frac{1}{\sqrt{\ell^2 + 2\sigma_1^2}} + \frac{1}{\sqrt{\ell^2 + 2\sigma_2^2}} - \frac{2}{\sqrt{\ell^2 + \sigma_1^2 + \sigma_2^2}} \exp\left(-\frac{(\mu_1 - \mu_2)^2}{2(\ell^2 + \sigma_1^2 + \sigma_2^2)} \right) \right) \tag{2.28}$$

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".

$$d_{reactor}\left[p\|q\right] = \begin{cases} 0 & p(\{t \geq T_{crit}\}) \geq \ell \text{ and } q(\{t \geq T_{crit}\}) \geq \ell \\ \ell_1 & p(\{t \geq T_{crit}\}) \geq \ell \text{ and } q(\{t \geq T_{crit}\}) \leq \ell \\ \ell_2 & p(\{t \geq T_{crit}\}) \leq \ell \text{ and } q(\{t \geq T_{crit}\}) \geq \ell \end{cases} \tag{2.29}$$

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 $d_{reactor}\left[\cdot\|\cdot\right]$ 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 $d_{KL}\left[\cdot\|\cdot\right]$ and $d_{reactor}\left[\cdot\|\cdot\right]$.

Chapter 3

Scoring rules for processes

The scoring rule framework presented in chapter **??** is quite general and accommodates a large number of traditional and more modern estimation and scoring techniques. However, there are certain limitations as to what the framework can describe. In this chapter I extend the scoring rule framework further and consider scoring rules for general stochastic processes, in other words, scoring rules in infinite dimensional sampling spaces. I introduce the notion of marginal scoring rules, and show a number of examples of estimation techniques that do not fit into the traditional scoring rule framework, but can be accommodated in this more general treatment. I will also argue that the notion of strictly proper scoring rules is insufficient in certain situations, and introduce an analogous property I call very strictly proper scoring. I give examples of scoring rules for processes.

Let us recall the definition of a strictly proper scoring rule: A score $S(x, P)$ is called strictly proper if for any two distributions P, Q the following holds:

$$\mathbb{E}_{x\sim P} S(x, P) \leq \mathbb{E}_{x\sim P} S(x, Q), \tag{3.1}$$

with equality only when $P = Q$.

In this definition we are concerned with the scoring rule's behaviour in expectation over the distribution P. In empirical terms taking an expectation is analogous to studying the limit as we observe multiple independent copies of the distribution x sampled from P.

Indeed, the strictly proper property of a scoring rule under suitable regularity conditions ensures that as we observe infinitely many independent and identically distributed samples from a parametric distribution $P_{X|\theta}$, the following estimate is consistent.

Definition 10 (Score matching estimate)**.** *Let $\{P_{X|\theta,\theta\in\Theta}\}$ be a parametric family of distributions and S a strictly proper scoring rule with respect to this class. The following estimator is called the score matching estimate:*

$$\hat{\theta}_N(x_1, \ldots, x_N) = \operatorname*{argmin}_{\theta\in\Theta} \sum_{n=1}^{N} S(x_n, P_{X|\theta}) \tag{3.2}$$

The above equation is an unbiased estimating equation, and under suitable regularity conditions $\hat{\theta}_N(x_1, \ldots, x_N)$ is a consistent estimator, that is if $x_1, \ldots \sim P_{X|\theta_0}$ i. i. d.

$$\lim_{N\to\infty} \hat{\theta}_N(x_1, \ldots, x_N) = \theta_0 \qquad P_{X|\theta_0}\text{- almost surely} \tag{3.3}$$

Score matching has been used to fit parametric models . Maximum likelihood estimation is a typical example when the scoring rule is chosen to be the logarithmic score. In [**?**], the kernel scoring rule is used to fit parametric distributions; the technique is referred to as kernel moment matching. Score matcing with the Brier and spherical scores is used in meteorology and **TODO: Where exactly?**.

However, the i. i. d. case is not the only limiting behaviour that is interesting. Often we want to score non-i. i. d. sequences of variables. Examples include:

1. Pseudo-likelihood estimation on a large Markov random field, studying the limit as the image size grows
2. Parameter estimation in phylogenetic tree models, studying behaviour as the number of species grows
3. Parameter estimation in non-parametric models, such as Gaussian process regression
4. Bayesian model selection in hierarchical Bayesian models, where instead of i. i. d. assumption we often find exchangeability assumptions

In all of the examples above, we do not observe a growing number of independent copies of the same random variable. Instead, we observe larger and larger marginals from a single draw or trajectory of a random process. In this section I devise a framework of scoring rules for stochastic processes, and define stronger equivalents of strictly proper scoring rules. Let us first consider the following definition.

Definition 11 (Marginal scoring rule)**.** *Let $\mathcal{X}$ be a countably infinite index set, $I \subseteq \mathcal{X}$ a subset of indices. A marginal scoring rule R_I over this index set is a function that assigns a real value, $R_I(y_I, \Pi)$, to a process measure $\Pi \in \mathcal{M}^1_{\mathcal{Y}^\mathcal{X}}$ and an observed marginal $y_I \in \mathcal{Y}^I$.*

If $\mathcal{X}$ is the set of natural numbers, and Π is an i. i. d. process such that $\Pi_I(y_I) = \prod_{i \in I} P(y_i)$, then we can readily construct marginal scoring rules of the following form:

$$R_I(y_I, Q) = \frac{1}{|I|} \sum_{i \in I_n} S(y_i, P), \tag{3.4}$$

where S is a strictly proper scoring rule over $\mathcal{Y}$. These kind of scoring rules are useful in the following sense.

Statement 4 (Consistency of marginal scoring for i. i. d. processes)**.** *Let us take a monotonically increasing sequence of index sets $I_1 \subseteq \ldots \subseteq I_N \subseteq \ldots \subseteq \mathcal{X}$, such that $\cup_{N=1}^{\infty} I_N = \mathcal{X}$ and $S(y, P)$ be a strictly proper scoring rule over $\mathcal{Y}$ with respect to the class of marginals P_θ. Let Π_θ denote the i. i. d. process over $\mathcal{Y}^\mathcal{X}$ with marginal P_θ. Then "typically" the following limit holds*

$$\operatorname*{argmin}_\theta R_{I_n}(y_{I_n}, \Pi_{\theta^*}) = \operatorname*{argmin}_\theta \frac{1}{|I_N|} \sum_{i \in I_N} S(y_i, P_\theta) \xrightarrow[N \to \infty]{P_{\theta^*}} \theta^* \tag{3.5}$$

The above equation is the same as score matching, but now we look at it slightly differently. Intuitively, consistency in this general sense means that as larger and larger marginals of a trajectory drawn from the process Π_θ are revealed, the scoring rule can identify the true parameter θ of the process from which the trajectory was drawn. When a general family marginal scoring rules has this property, we will call it the very strictly proper property:

Definition 12 (Very strictly proper marginal scoring)**.** *Let R_I be a family of marginal scoring rules over processes on $\mathcal{Y}^\mathcal{X}$, $\mathcal{Q} = \{\Pi_\theta, \theta \in \Theta\}$ a family of process measures over $\mathcal{Y}^\mathcal{X}$. R_I is very strictly proper with respect to $\mathcal{Q}$ if for any monotonically increasing sequence of index sets $I_1 \subseteq \ldots \subseteq I_N \subseteq \ldots \subseteq \mathcal{X}$, such that $\cup_{N=1}^{\infty} I_N = \mathcal{X}$ the following limit holds:*

$$\lim_{N \to \infty} \operatorname*{argmin}_\theta R_{I_N}(y_{I_N}, \Pi_\theta) = \theta^* \qquad \Pi_{\theta^*} \text{ almost surely} \tag{3.6}$$

In the following sections I show families of marginal scoring rules R_I, that do not simply decompose into sums of scoring rules on $\mathcal{Y}$ for marginals Π, but most of which still have the very strictly proper property.

3.0.4 Maximum product of spacings score

The first example, *maximum product of spacings (MPS) estimation* is not normally considered in the context of scoring rules, but we can now interpret it as part of this general framework of marginal scoring rules. It is an estimation technique based on the observation that if a set of samples are sampled from a one dimensional uniform distribution, then the spacing between neighbouring samples should be roughly equal. Thus, a viable "scoring rule" for the uniform distribution measures how uneven the spacing between neighbouring samples are. The evenness of a set of numbers can be measured by the difference between arithmetic and geometric means. The above argument can be extended to arbitrary real probability distributions by noting that transforming a general non-uniform random variable by its cumulative distribution function results in a uniform random variable between 0 and 1.

Definition 13 (Product of spacings score)**.** *Let $y_1, \dots, y_N$ independent random variables on the real line $\mathbb{R}$. Let P be a distribution over $\mathbb{R}$ with cumulative distribution function F_P. Let $\pi : \{1, \dots, N\} \mapsto \{1, \dots, N\}$ a rank ordering such that $n < m \implies y_{\pi_n} \leq y_{\pi_m}$. For convenience of notation let us further define $y_{\pi_0} := 0$ and $y_{\pi_{N+1}} := 1$.*

$$R_N^{PS}(y_1, \dots, y_N, P) = -\frac{1}{N+1} \sum_{n=0}^{N} \log \left(F_P \left(y_{\pi_{n+1}} \right) - F_P \left(y_{\pi_n} \right) \right) \tag{3.7}$$

It is shown in [**?**] that subject to smoothness assumptions, MPS estimation, which minimises R_N^{PS} is consistent when y_n are sampled i. i. d. from a distribution. That is

$$\operatorname*{argmin}_{\theta} R_N^{PS}(y_1, \dots, y_N, \Pi_{\theta^*}) \xrightarrow[N \to \infty]{P_{\theta^*}} \theta^* \tag{3.8}$$

Furthermore, the estimator is often statistically more efficient than maximum likelihood estimation (MLE) [**?**]. The drawback of the product of spacings score is that it does not readily generalise to multivariate distributions, and it requires the knowledge of the cumulative distribution function to be calculated.

3.0.5 Decision theoretic scoring and F_β scores

Further examples of scores that do not decompose as in (3.4) can be constructed on decision theoretic grounds when the action and the loss that we minimise itself depends on multiple outcomes. For example in a binary case we may want to penalise imbalanced forecasters, so we prefer forecasters that produce about as many false positives as false negatives while at the same time minimise the total number of false predictions.

Most of these complicated objectives can be achieved by optimising something like the F_β score, which is applied in a variety of statistical applications []:

$$F_\beta = (1 - \beta^2) \frac{\text{precision} \cdot \text{recall}}{\beta^2 \text{precision} + \text{recall}} \tag{3.9}$$

The F score depends on a whole sample and cannot be decomposed as a sum of terms that score each sample independently. Hence it is a good example of general

3.1 Non-i. i. d. processes

In the examples I gave above, the scoring rule did not decompose into a sum of univariate scores as in (3.4), but we would still use these scores to score and estimate i. i. d. processes. The present framework also accommodates more general cases when we want to score non-i. i. d. processes, such as hierarchical Bayesian models, Markov random fields, processes over tree structures or Gaussian processes.

The simplest and most widely adopted scoring rule of this form is log-score (assuming marginals of Π have densities P_{I_n}), which is also called (negative) evidence:

$$R_{I_N}(x_{I_N}, \Pi) = -\frac{1}{|I_N|} \log P_{I_n}(y_{I_n}), \tag{3.10}$$

3.1.1 Bayesian model selection

The consistency of the maximum likelihood estimator has been studied with respect to particular classes of models and processes. A particularly interesting case is when the process Π is exchangeable.

By De Finetti's theorem, exchangeable sequences of random variables have a representation as mixtures of i. i. d. sequences, and can therefore be interpreted as hierarchical Bayesian models. Let's say we have a model $\mathcal{M}$, under which y_n are conditionally i. i. d. given some parameters θ. Given some observed data y, the model's evidence is given by the following formula:

$$\begin{aligned} \text{evidence}(\mathcal{M}) &= R_{log}(y_{1:N}, \Pi_{Y|\mathcal{M}}) && (3.11)\\ &= \log P_{Y_{1:N}|\mathcal{M}}(y_{1:N}) && (3.12)\\ &= \log \int \prod_{n=1}^{N} P_{Y_1|\theta}(y_n) P_{\theta|\mathcal{M}}(\theta) d\theta && (3.13) \end{aligned}$$

We can observe that even though conditioned on θ subsequent y_n variables are independent, marginally they are dependent, therefore the evidence does not decompose into a sum of terms per data-point as it would for marginally i. i. d. models.

The evidence is accepted as a very robust criterion for Bayesian model selection and is thought to be particularly robust against over-fitting. On the other hand it is often intractable to compute exactly for complicated models, and one has to rely on various approximations such as the Bayesian information criterion (BIC)[], Akaike information criterion (AIC)[], variational lower bounds, or expectation-propagation (EP)[].

We can generalise Bayesian model selection by replacing the log-score with another family of marginal scoring rulea, such as the quadratic kernel rule from Eqn. (**??**). When we use the kernel scoring rule for Bayesian model selection, it decomposes clearly into two terms, which intuitively represent the trade-off between accuracy and model flexibility.

$$\begin{aligned} d_k\left[\delta_y \middle\| \Pi_{Y|\mathcal{M}}\right] &= \|k(\cdot, y) - \mu_{Y|\mathcal{M}}\|^2_{\mathcal{H}} && (3.14)\\ &= \|k(\cdot, y) - \mathbb{E}_{\theta\sim p_{\theta|\mathcal{M}}} \mu_{Y|\theta}\|^2_{\mathcal{H}} && (3.15)\\ &= \mathbb{E}_{\theta\sim p_{\theta|\mathcal{M}}} \|k(\cdot, y) - \mu_{Y|\theta}\|^2_{\mathcal{H}} - \mathbb{E}_{\theta\sim p_{\theta|\mathcal{M}}} \|\mu_{X|\theta,\mathcal{M}} - \mu_{X|\mathcal{M}}\|^2_{\mathcal{H}} && (3.16)\\ &= \underbrace{\mathbb{E}_{\theta\sim p_{\theta|\mathcal{M}}} d_k\left[\delta_y \middle\| p_{X|\theta}\right]}_{\text{average accuracy}} - \underbrace{\mathbb{I}_k\left[X \leftarrow \theta\right]}_{\text{diversity}} && (3.17) \end{aligned}$$

After [] one may call this the ambiguity decomposition of the quadratic kernel score. As the Brier score is a special case of the kernel scoring rule, it naturally admits the same decomposition. The decomposition suggests that a good model is

accurate: it forecasts observed data relatively accurately on average for all parameters, and

diverse: it is capable of expressing a diverse range of behaviours via its parameters

3.1.2 Pseudo-likelihood

We can also study the pseudo-likelihood score in this context. In machine learning this score may be called the leave-one-out cross-validation error.

$$R(y_{I_N}, \Pi) = -\frac{1}{|I_N|} \sum_{n \in I_N} \log P_{Y_n | Y_{I_n \setminus \{n\}}}(y_n) \tag{3.18}$$

We already noted that the pseudo-likelihood score is strictly proper, so under multiple sampling from the same size marginal, pseudo-likelihood it is typically consistent. But we can also study the limit as larger and larger marginals are considered.

The most typical application of pseudo-likelihood estimation is parameter estimation in Markov random fields. A finite graph Markov-random field is a stochastic process, where each variable is conditionally independent of the rest of the trajectory given a finite number of other variables, known as the Markov-blanket of the variable. A common example of a finite graph Markov random field is the two-dimensional square lattice Isling model [**?**]. This model and its generalisations have found several applications in computer vision and image processing [].

Typically when studying consistency properties of pseudo-likelhiood, authors show that the pseudo-likelihood score is strictly proper and thus pseudo-likelihood estimation is consistent under repeated sampling from a finite dimensional Markov random field. In this chapter we are interested in consistency of estimation from a single draw from an infinite MRF, in the limit as larger and larger sub-graphs are observed. It turns out pseudo-likelihood estimation is consistent in this stricter sense, too [**?**], therefore we can call it a very strictly proper scoring rule.

Pseudo-likelihood estimation can be used in other cases where computing the logarithmic score is intractable, such as estimating phylogenetic trees in models such as the infinite sites model [**?**, **?**] . In this application the observed data are a set of related genomes, that are thought to have evolved from a common ancestor genome through a process of random mutations and recombinations and can therefore be related via a phylogenetic tree. The task is to infer the latent phylogenetic tree underlying the data, and to estimate model parameters such as the relative rate of mutation and recombination events.

This estimation problem is hard because of the exponential number of phylogenetic trees consistent with any one dataset. Computing the marginal likelihood (or logarithmic score, as in Eqn. (**??**)) is intractable as it involves computing a non-trivial sum over phylogenetic trees. However, computing the conditional distribution of just one gene (or segregating site on the DNA), conditioned on the observed values of other genes can be performed relatively efficiently in certain models, and thus pseudo-likelihood estimation may be efficiently applied to these models. It is an open question whether pseudo-likelihood estimation, or indeed maximum likelihood estimation is consistent in the limit of increasing number of alleles and species.

3.1.3 Information quantities for processes

Finally, it is a natural question to ask, whether it possible, or if at all useful to define information quantities, such as entropy and divergence, between stochastic processes. Following definition 12 it makes sense to replace expectations with limits in these definitions.

Definition 14 (Entropy and divergence for processes)**.** *Let R_I be a family of marginal scoring rules over processes on $\mathcal{Y}^{\mathcal{X}}$, R_I is very strictly proper family of marginal scoring rules.Consider a monotonically increasing sequence of index sets $I_1 \subseteq \ldots \subseteq I_N \subseteq \ldots \subseteq \mathcal{X}$. Let us define the entropy or a random process Π as follows*

$$\mathbb{H}_R[\Pi] = lim_{N \to \infty} R_{I_N}(X_{I_N}, \Pi), \text{ where } X \sim \Pi \tag{3.19}$$

Similarly, let us define the divergence between two processes Π and Ξ as

$$d_R[\Pi \| \Xi] = lim_{N \to \infty} R_{I_N}(X_{I_N}, \Xi) - R_{I_N}(X_{I_N}, \Pi), \text{ where } X \sim \Pi \tag{3.20}$$

In general the quantities defined above are random quantities. When the scoring rule R is the logarithmic score, and Π is a stationary stochastic process, the Shannon-McMillan-Breiman theorem ensures that the entropy defined this way converges almost surely to deterministic value [], called the source entropy of the probabilistic source Π. Under the same conditions the divergence $d_R[\Pi \| \Xi]$ also converges and is strictly positive for $\Pi \neq \Xi$.

We can conclude that in certain special cases the entropy and divergence quantities defined in definition 14 make sense and are useful, but very likely these generalisations are too general for practical purposes.

Chapter 4

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 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 some data or dataset $\mathcal{D} \in \mathcal{O}$ 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.

Variational methods find the optimal approximation q^* by maximising a lower bound to the marginal likelihood as follows.

$$\begin{aligned}
\log p(\mathcal{D}) &= \log \int p(\mathcal{D}|\theta) p(\theta) d\theta && (4.2)\\
&= \log \int \frac{p(\mathcal{D}|\theta)p(\theta)}{q(\theta)} q(\theta) d\theta && (4.3)\\
&\geq \int \log \frac{p(\mathcal{D}|\theta)p(\theta)}{q(\theta)} q(\theta) d\theta && (4.4)\\
&= \log p(\mathcal{D}) - d_{KL}\left[p_\mathcal{D} \| q\right] && (4.5)
\end{aligned}$$

by minimising the Kullback-Leibler divergence (Eqn. (**??**)) between the approximate distribution q and the true posterior $p_\mathcal{D}$. The KL divergence is non-symmetric, therefore the order of arguments matter. As I argued in section **??**, in the scoring rule interpretation suggests that the *right* way to use divergence is when it's first argument is the true distribution we want to approximate, and the second argument the approximation q. So we can conclude that variational Bayesian inference minimises KL divergence in the wrong direction. This has been noted previously by [**?**, **?**] and many other authors. This does not mean that variational inference does not work, it just means that by performing variational inference we loose the intuitive meaning of KL divergences.

We can try to generalise variational inference to general scoring rules along two lines. Firstly, the log likelihood can be replaced by the average score

However, generally, Bregman divergences are convex in their first parameter and not generally in the second, so a variational lower bound only holds if there is a Several approaches therefore tried to fix this conceptual issue, and minimise the divergence in the other direction. This is unfortunately very hard, as computing the divergence $d_S[p_{\mathcal{D}} \| q]$ requires an integral over the posterior $p_{\mathcal{D}}$, which is normally intractable, and this is why we perform approximate inference in the first place.

Assumed density filtering, and its generalisation, expectation propagation (EP) try to approximate the ideal method of minimising $d_{KL}[p_{\mathcal{D}} \| q]$ as follows. EP assumes the posterior can be written

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

4.4 Approximate Bayesian decision theory

Chapter 5

Bayesian experiment design

5.1 General framework for Bayesian experiment design

5.1.1 Shannon's entropy

5.1.2 Decision theoretic active classification

5.1.3 Bayesian optimisation

5.1.4 Bayesian quadrature

5.2 Bayesian active learning by Disagreement (BALD)