Scoring rules and Information Quantities in Bayesian Analysis — Why I am never going to submit this? —

Ferenc Huszár

April 15, 2013

Contents

Chapter 1

Introduction

Foor bar Lorem ipsum.

Part I

Scoring rules and Bayesian analysis

Chapter 2

An introduction to scoring rules

In this section I describe scoring rules that are 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 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 $S(x, P)$. In this thesis I follow a convention by which scoring rules are interpreted as losses, so lower values are associated with better predictions.

A well known 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 **??**), which made it popular in the probabilistic machine learning community. But it is not the only one, and there are situations in which it is more convenient or efficient to use alternative scoring rules. Mathematically, a scoring rule can be any measurable function that maps an outcome-probability distribution pair onto real numbers: $S : \mathcal{X} \times \mathcal{M}^1_{\mathcal{X}} \mapsto \mathcal{R} \cup \{\infty\}$. I will give further examples of scoring rules in section **??**.

2.1 Information quantities

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

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:*

$$\mathbb{H}_S[P] = \mathbb{E}_{x\sim P} S(x, P) \tag{2.1}$$

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:*

$$d_S[P\|Q] = \mathbb{E}_{x\sim P} S(x, Q) - \mathbb{E}_{x\sim P} S(x, P). \tag{2.2}$$

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:*

$$\mathbb{E}_{x\sim P} S(x, Q) \geq \mathbb{E}_{x\sim P} S(x, P), \tag{2.3}$$

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

$$d_S[P\|Q] \geq 0. \tag{2.4}$$

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.**TODO: what does this paragraph say?**

Definition (2.2) suggests that the the first argument, P, should take the role of the true distribution, and Q the approximate. **TODO: elaborate on this.**

The divergence defined in (2.2) is a special case of Bregman divergences. Bregman divergences are an important class of divergence functions on complex domains, and include well known distances such as the Eulidean distance.

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

$$d_{Bregman,H}[P\|Q] = H(P) - H(Q) + \langle \nabla H(Q), Q - P\rangle \tag{2.5}$$

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 $\mathbb{H}_S[P]$:

$$\mathbb{H}_S[P] = \mathbb{E}_{x\sim P} S(x, P) = \langle P, S(\cdot, P)\rangle \tag{2.6}$$

Using this notation

$$\begin{aligned} \nabla \mathbb{H}_S[P] &= \nabla \langle P, S(\cdot, P)\rangle && (2.7)\\ &= S(\cdot, P) + \langle P, \nabla S(\cdot, P)\rangle && (2.8) \end{aligned}$$

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

$$\begin{aligned} d_{Bregman,\mathbb{H}_S}[P\|Q] &= \mathbb{H}_S[Q] + \langle \nabla \mathbb{H}_S[Q], P - Q\rangle - \mathbb{H}_S[P] && (2.9)\\ &= \mathbb{H}_S[Q] + \langle S(\cdot, Q), P - Q\rangle - \mathbb{H}_S[P] && (2.10)\\ &= \langle S(\cdot, Q), P\rangle - \mathbb{H}_S[P] && (2.11)\\ &= d_S[P\|Q] && (2.12) \end{aligned}$$

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

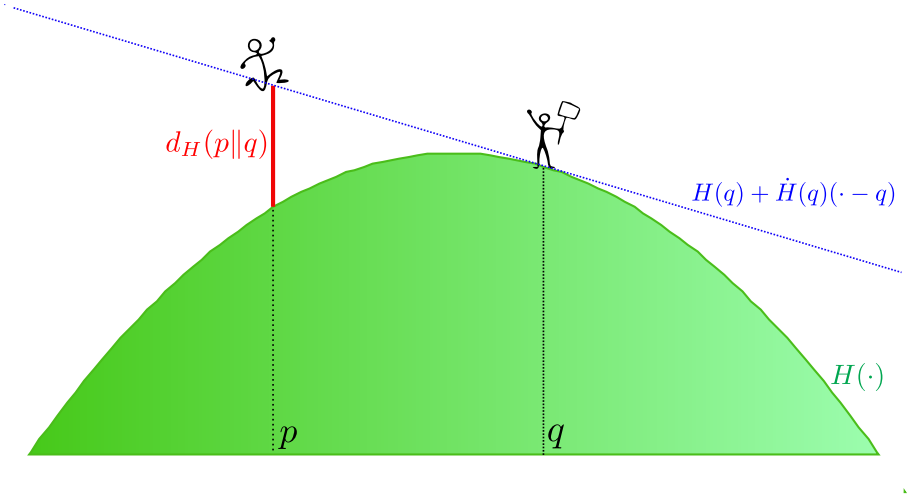

Figure 2.1: Pictorial illustration of Bregman divergences. Peter and Quentin are points who live on a convex hill, whose surface is described by the concave function $H(p)$. Peter lives at $(Q, H(P))$, Quentin at $(Q, H(Q))$. Because the hill is convex and they are both points, they cannot normally see 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.

For a more elaborate proof and discussion of Bregman divergences and scoring rules please refer to [Amari and Cichocki, 2010, Dawid, 2007]. An intuitive explanation of Bregman divergences is given in Figure **??**.

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 two variables:

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*

$$\mathbb{I}_S[X \leftarrow Y] = \mathbb{E}_{x\sim P_X} S(x, P_X) - \mathbb{E}_{y\sim P_Y}\mathbb{E}_{x\sim P_{X|Y=y}} S(x, P_{X|Y=y}) \tag{2.13}$$

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

$$\mathbb{I}_S[X \leftarrow Y] = \mathbb{H}_S[P_X] - \mathbb{E}_{y\sim P_Y}\mathbb{H}_S\left[P_{X|Y=y}\right] \tag{2.14}$$

$$= \mathbb{E}_{y\sim P_Y} d_S\left[P_X \middle\| P_{X|Y=y}\right] \tag{2.15}$$

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 $S : \mathcal{X} \times \mathcal{M}_{\mathcal{X}}^1 \mapsto \mathbb{R}$ be a strictly proper scoring rule with respect to probability distributions $\mathcal{M}_{\mathcal{X}}^1$, and $P \in \mathcal{M}_{\mathcal{X}\times\mathcal{Y}}^1$ 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 $\mathbb{I}_S [X \leftarrow Y] = 0$.

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

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

Corollary 1 (Weak equivalence of strictly proper scoring rules)**.** *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 $\mathbb{I}_{S_1} [X \leftarrow Y] > 0$ if and only if $\mathbb{I}_{S_2} [X \leftarrow Y] > 0$.*

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

Corollary 2 (Weak symmetry of information)**.** *Let $S_X : \mathcal{X} \times \mathcal{M}_{\mathcal{X}}^1 \mapsto \mathbb{R}$ be a strictly proper scoring rule over X and $S_Y : \mathcal{Y} \times \mathcal{M}_{\mathcal{Y}}^1 \mapsto \mathbb{R}$ be a strictly proper scoring rule over Y. Then $\mathbb{I}_{S_X} [X \leftarrow Y] > 0$ if and only if $\mathbb{I}_{S_Y} [Y \leftarrow X] > 0$.*

2.2 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 known scoring rules, the logarithmic, Brier (quadratic) and spherical scores. Then I present the kernel scoring rule, which is lesser known in the statistics literature. I establish the connections between the kernel scoring rule to the maximum mean discrepancy, a divergence measure that has gained popularity recently in the machine learning community over the past years.

Following the discussion of kernel scoring rules 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.

2.2.1 The logarithmic score

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

$$S_{log}(x, P) = -\log P(x) \tag{2.16}$$

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

$$\theta_{ML} = \underset{\theta}{\operatorname{argmax}} \sum_{n=1}^{N} \log P(x_i|\theta) \tag{2.17}$$

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

$$\mathbb{H}_{Shannon} [P] = -\mathbb{E}_{x\sim P} \log P(x) \tag{2.18}$$

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

$$d_{KL}\left[P\|Q\right] = \mathbb{E}_{x\sim P}\frac{\log P(x)}{\log Q(x)} \tag{2.19}$$

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 are related to 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 for scoring rules other than the logarithmic [Parry et al., 2012, Dawid et al., 2012].

The value of information becomes Shannon's mutual information, a crucial quantity in channel coding [Shannon, 1948, MacKay, 2002]. Interestingly, Shannon's mutual information can be rewritten as the KL divergence between the joint distribution and the product of marginals:

$$\begin{aligned}
\mathbb{I}_{Shannon}\left[X \leftarrow Y\right] &= \mathbb{H}_{Shannon}\left[X\right] - \mathbb{E}_{y\sim P_Y}\mathbb{H}_{Shannon}\left[P_{X|Y=y}\right] && (2.20)\\
&= \mathbb{E}_{y\sim P_Y} d_{KL}\left[P_X \middle\| P_{X|Y=y}\right] && (2.21)\\
&= \mathbb{E}_{y\sim P_Y}\left[\mathbb{E}_{x\sim P_{X|Y=y}} \log \frac{P_{X|Y=y}(x)}{P_X(x)}\right] && (2.22)\\
&= \mathbb{E}_{(x,y)\sim P} \log \frac{P(x,y)}{P_X(x)P_Y(y)} && (2.23)\\
&= d_{KL}\left[P(x,y)\|P_X(x)P_Y(y)\right] && (2.24)
\end{aligned}$$

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, I note here that some authors have generalised Shannon's mutual information along the lines of (2.24), by replacing the KL divergence with a more general divergence d:

$$\mathbb{J}_d(X, Y) = \mathrm{d}\left[P(x,y)\|P_X(x)P_Y(y)\right] \tag{2.25}$$

Examples of information functionals defined this way are described in [Póczos and Schneider, 2011]. 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.

2.2.2 The pseudolikelihood

The idea of maximum pseudolikelihood estimation was introduced originally by [Besag, 1977] to estimate parameters of Gaussian random fields. Later it was popularised in the context if parameter

estimation general Markov random fields [Comets, 1992] and in Boltzmann machines [Hyvärinen, 2006]. The pseudolikelihood is particularly useful for estimating parameters of statistical models with intractable normalisation constants.

$$S_{\text{pseudo}}(x, P) = -\sum_{d=1}^{D} \log P(x_d | x_{\neg d}), \tag{2.26}$$

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

$$p(x_d | x_{\neg d}) = \frac{P(x)}{\int P(X_d = y, x_{\neg d}) dy} = \frac{C \cdot P(x)}{\int C \cdot P(X_d = y, x_{\neg d}) dy} \tag{2.27}$$

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.

$$P(x) = \frac{1}{Z} \exp(x^T W x + b^T x), x \in \{0, 1\}^D, \tag{2.28}$$

where $Z = \sum_{x \in \{0,1\}^D} \exp(x^T W x + b^T x)$ 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:

$$P(x_d | x_{\neg d}, W, b) = \frac{p(x)}{\int p(x_d = y, x_{\neg d}) dy} \tag{2.29}$$

$$= \frac{\frac{1}{Z} \exp(x^T W x + b^T x)}{\sum_{x_d \in \{0,1\}} \frac{1}{Z} \exp(x^T W x + b^T x)} \tag{2.30}$$

$$= \frac{\exp(x^T W x + b^T x)}{\sum_{x_d \in \{0,1\}} \exp(x^T W x + b^T x)} \tag{2.31}$$

$$= \frac{\exp\left(x_d \left(W_{d,d} + 2W_{d,\neg d}^T x_{\neg d} + b_d\right)\right)}{\exp(W_{d,d} + 2W_{d,\neg d}^T x_{\neg d} + b_d) + 1} \tag{2.32}$$

$$\tag{2.33}$$

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. [Hyvärinen, 2006] showed that pseudo-likelihood estimation is consistent for fully visible Boltzmann machines. **TODO: For this to make sense we have to first talk about score matching...**

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:

$$S_{\log}(x, p) = -\log P(x) = -\sum_{d=1}^{D} \log P(x_d | x_{1:d-1}) \tag{2.34}$$

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 $X_{\pi_d} \perp\!\!\!\perp X_{\pi_{d+1}}, \dots, X_{\pi_D} | X_{\pi_1}, \dots, X_{\pi_{d-1}}$.

[Csiszár and Talata, 2004] 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:

$$S_{\mathrm{DLP12}}(x, P) = -\sum_{d=1}^{D} S_d \left(x_d, P_{X_d | X_{\neg d} = x_{\neg d}}\right), \tag{2.35}$$

where S_d are strictly proper scoring rules for each dimension

2.2.3 The Brier (quadratic) score

Another widely used scoring rule is the so-called *Brier score* or quadratic score, originally introduced in [Brier, 1950]. It was first applied to evaluating probabilistic weather forecasts and it is still used in meteorology [Ferro, 2007] as well as in medicine [Spiegelhalter, 2006] and epidemiology [Redelmeier et al., 1991]. It is also related to the (root) mean squared error of probabilistic binary classifiers, which is a commonly used loss function for training neural networks [Rumelhart et al., 1988].

We will define the Brier score in terms of the L^2 norm of a probability distributions, which we define as:

$$\|P\|_2 = \sqrt{\mathbb{E}_{x \sim P} P(x)} \tag{2.36}$$

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:

$$\begin{aligned} S_{Brier}(x, P) &= \|P - \delta_x\|_2^2 && (2.37)\\ &= \|P\|_2^2 - 2P(x) + 1 && (2.38)\\ &= \mathbb{E}_{x' \sim P} P(x') - 2P(x) + 1 && (2.39)\\ & && (2.40) \end{aligned}$$

The score gives rise to the following entropy function.

$$\begin{aligned} \mathbb{H}_{Brier}[P] &= \mathbb{E}_{x \sim P} \left[\mathbb{E}_{x' \sim P} P(x') - 2P(x) + 1\right] && (2.41)\\ &= 1 - \mathbb{E}_{x \sim P} P(x) && (2.42)\\ &= 1 - \|P\|_2^2 && (2.43) \end{aligned}$$

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 equals $1 - \sum_{d=1}^{D} \frac{1}{D^2} = 1 - \frac{1}{D}$. 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, and thus symmetric. The divergence is analogous to the squared Euclidean distance.

$$d_{Brier}\left[P\|Q\right] = \mathbb{E}_{x\sim Q}\left[\|P\|_2^2 - 2P(x) + 1\right] - \mathbb{H}_{Brier}\left[P\right] \tag{2.44}$$
$$= \|P\|_2^2 - 2\mathbb{E}_{x\sim Q}P(x) + \|P\|_2^2 \tag{2.45}$$
$$= \|P\|_2^2 - 2\left\langle P, Q\right\rangle + \|Q\|_2^2 \tag{2.46}$$
$$= \|P - Q\|_2^2 \tag{2.47}$$

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

$$\mathbb{I}_{Brier}\left[X \leftarrow Y\right] = \mathbb{E}_{y\sim P_Y}\|P_X - P_{X|Y=y}\|_2^2 \tag{2.48}$$
$$\tag{2.49}$$

2.2.4 Spherical and pseudo-spherical scoring rules

Another example of strictly proper scoring rules, introduced in [Good, 1971] is the spherical scoring rule [Dawid, 2007, Dawid et al., 2012]. The spherical score is defined as follows:

$$S_{spherical}(x, P) = 1 - \frac{P(x)}{\|P\|_2} \tag{2.50}$$

This gives rise to the following entropy and divergence functions.

$$\mathbb{H}_{spherical}\left[P\right] = 1 - \mathbb{E}_{x\sim P}\frac{P(x)}{\|P\|_2} \tag{2.51}$$
$$= 1 - \|P\|_2 \tag{2.52}$$

$$d_{spherical}\left[P\|Q\right] = -\mathbb{E}_{x\sim P}\frac{Q(x)}{\|Q\|_2} + \|P\|_2 \tag{2.53}$$
$$= \|P\|_2 - \frac{\langle Q, P\rangle}{\|Q\|_2} \tag{2.54}$$
$$= \|P\|_2\left(1 - \cos(P, Q)\right), \tag{2.55}$$

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, $d_{spherical}\left[P\|c\cdot Q\right] = d_{spherical}\left[P\|Q\right]$ and $d_{spherical}\left[c\cdot P\|Q\right] = c\cdot d_{spherical}\left[P\|Q\right]$. This means that when approximating a fixed distribution P by Q via minimising $d_{spherical}\left[P\|Q\right]$ we only need to know P and Q up to a normalising constant.

The value of information under the spherical score is

$$\mathbb{I}_{spherical}\left[X \leftarrow Y\right] = \|P_X\|_2\mathbb{E}_{y\sim P_Y}\left(1 - \cos(P_X, P_{X|Y=y})\right) \tag{2.56}$$

Gneiting and Raftery [2007] and Jose et al. [2008] also introduce generalisations of the spherical score, where the L_2 norm is replaced by a general L_γ norm:

$$\mathbb{H}_{\gamma,pseudospherical}\left[P\right] = -\|P\|_\gamma \tag{2.57}$$

2.2.5 The kernel scoring rule

The kernel scoring rule first appeared in the statistics literature in [Eaton et al., 1996], although the name *kernel scoring rule* was only used in more recent references [Dawid and Sebastiani, 1999, Dawid, 2007, Gneiting and Raftery, 2007].

Recently, a related concept, derived from different first principles, has become known in the machine learning community as *maximum mean discrepancy* (MMD, [Sriperumbudur et al., 2008]). As we will see, MMD is closely related to the kernel scoring rule. MMD has been adopted in a variety of modern applications in machine learning and statistics, including two sample tests [Gretton et al., 2012], kernel moment matching [Song et al., 2008], embedding of probability distributions[Smola et al., 2007] and the kernel-based message passing [Fukumizu et al., 2010].

MMD measures the divergence or distance between two distributions, P and Q. It belongs to a rich class of divergences called integral probablity metrics [Sriperumbudur et al., 2009], which define the distance between p and q, with respect to a class of integrand functions $\mathcal{F}$ as follows:

$$d_{\mathcal{F}}\left[P\|Q\right] = \sup_{f\in\mathcal{F}} \left|\mathbb{E}_{x\sim P} f(x) - \mathbb{E}_{x\sim Q} f(x)\right| \tag{2.58}$$

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 expectation under P and Q should be small. This class of divergences include Wasserstein distance [del Barrio et al., 1999], Dudley metric [Dudley, 1974] and MMD, which differ only 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 [**?**]:

$$\begin{aligned}
d_k\left[P\|Q\right] &:= \mathrm{MMD}^2\left(P, Q\right) && (2.59)\\
&= \sup_{\substack{f\in\mathcal{H}\\ \|f\|_{\mathcal{H}}=1}} \left(\mathbb{E}_{x\sim P} f(x) - \mathbb{E}_{x\sim Q} f(x)\right)^2 && (2.60)\\
&= \sup_{\substack{f\in\mathcal{H}\\ \|f\|_{\mathcal{H}}=1}} \left|\mathbb{E}_{x\sim P} \left\langle f, k(\cdot, x)\right\rangle - \mathbb{E}_{x\sim Q} \left\langle f, k(\cdot, x)\right\rangle\right|^2 && (2.61)\\
&= \sup_{\substack{f\in\mathcal{H}\\ \|f\|_{\mathcal{H}}=1}} \left|\left\langle f, \mathbb{E}_{x\sim P} k(\cdot, x) - \mathbb{E}_{x\sim Q} k(\cdot, x)\right\rangle\right|^2 && (2.62)\\
&= \sup_{\substack{f\in\mathcal{H}\\ \|f\|_{\mathcal{H}}=1}} \left\langle f, \mu_P - \mu_Q\right\rangle^2 && (2.63)\\
&= \|\mu_P - \mu_Q\|_{\mathcal{H}}^2 && (2.64)\\
&= \mathbb{E}_{x,x'\sim P} k(x, x') - 2\mathbb{E}_{x\sim P}\mathbb{E}_{x'\sim Q} k(x, x') + \mathbb{E}_{x,x'\sim Q} k(x, x'), && (2.65)
\end{aligned}$$

In the derivation above $\mu_P(\cdot) = \int k(\cdot, x)P(x)dx$ is called the mean element or RKHS-embedding of the probability distribution p. The kernel score is simply the squared Hilbert norm of the difference between mean elements.

The most interesting kernels for the purposes of Hilbert-space embedding of distributions are those called *characteristic* [Sriperumbudur et al., 2008]. If the kernel k is characteristic, the mapping from Borel probability measures to mean elements 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 [Ord et al., 1999]. The characteristic function of a probability distribution p over the real line is defined as follows:

$$\phi_p(t) = \mathbb{E}_{x\sim p}\left[e^{itx}\right] = \int e^{itx} p(x)dx, \tag{2.66}$$

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 (2.65) assumed a real valued kernel function, therefore it is not valid for the special case of 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. [Gretton et al., 2012, 2009] developed and analysed empirical estimators of MMD for this problem.

Herding [Welling, 2009, Chen et al., 2012], 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 [Song et al., 2008] 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 (2.2), and corresponds to the following scoring rule:

$$S_k(x, P) := k(x, x) - 2\mathbb{E}_{x' \sim P} k(x, x') + \mathbb{E}_{x', x'' \sim Q} k(x', x'') \tag{2.67}$$

$$= \sup_{\substack{f \in \mathcal{H} \\ \|f\|_{\mathcal{H}} = 1}} \left(f(x) - \mathbb{E}_{x \sim Q} f(y)\right)^2 \tag{2.68}$$

This scoring rule is analogous to the kernel scoring rule introduced originally in [**?**]. The original definition differed from the formula by a factor of two, and it did not have 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, and are hence equivalent.

The connection between the kernel scoring rule as it is known in statistics and the maximum mean discrepancy has been first pointed out in [**?**], and it is one of the original contributions in this thesis. This interpretation allows one to uncover previously unknown connections between existing machine learning methods and to provide a solid theoretical framework for understanding and generalising them.

[Gneiting and Raftery, 2007] provide a proof of the propriety of the kernel scoring rule for Borel probability measures whenever the expectation $\mathbb{E}_{x, x' \sim P} k(x, x')$ is finite. Using the theory developed to study properties of MMD and characteristic kernels we can also see that the scoring rule is strictly proper whenever the kernel is characteristic [**?**]. [Gneiting and Raftery, 2007] showed that many examples of scoring rules, among them the Brier score (see section **??**), can be interpreted as special cases of the kernel scoring rule.

The generalised entropy defined by this scoring rule becomes:

$$\mathbb{H}_k[P] := \mathbb{E}_{x \sim P} k(x, x) - \mathbb{E}_{x, x' \sim P} k(x, x') \tag{2.69}$$

This entropy function is very general and is concave for all positive definite kernel k. Importantly, it 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. Thus, in most practical cases the entropy functional is bounded. Irrespective of kernel choice, the entropy is zero for delta distributions, that is when the distribution P is concentrated on a single point. If the kernel satisfies the strict inequality $\forall x, y : k(x, x) > k(x, y)$, the entropy is strictly positive for all other probability distributions.

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 defined. This property is useful in applications such as quasi-Monte Carlo in chapter **??**.

Thirdly, the entropy function has the kernel as free parameter, which is in fact 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 following derivation shows that the Bregman divergence between two distributions P and Q under the kernel scoring rule becomes the squared maximum mean discrepancy defined in equation (6.6).

$$\begin{aligned} d_{S_k}[P\|Q] &= \mathbb{E}_{x\sim P}[S_k(x,Q)] - \mathbb{E}_{x\sim P}[S_k(x,Q)] && (2.70)\\ &= \mathbb{E}_{x\sim P}k(x,x) - 2\mathbb{E}_{x\sim P}\mathbb{E}_{x'\sim Q}k(x,x') + \mathbb{E}_{x,x'\sim Q}k(x,x') && (2.71)\\ &\quad - (\mathbb{E}_{x\sim P}k(x,x) - \mathbb{E}_{x,x'\sim P}k(x,x')) && (2.72)\\ &= \mathbb{E}_{x,x'\sim P}k(x,x') - 2\mathbb{E}_{x\sim P}\mathbb{E}_{x'\sim Q}k(x,x') + \mathbb{E}_{x,x'\sim Q}k(x,x') && (2.73)\\ &= d_k[P\|Q] && (2.74) \end{aligned}$$

TODO: this paragraph assumes more than previous paragraphs... 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:

$$\begin{aligned} \mathbb{I}_k[X \leftarrow Y] &= \mathbb{E}_{y\sim P_Y} d_k\left[P_X \middle\| P_{X|Y=y}\right] && (2.75)\\ &= \mathbb{E}_{y\sim P_Y}\|\mu_{X|Y=y} - \mu_X\|^2_{\mathcal{H}} && (2.76)\\ &= k(P_X,P_X) - 2 * \mathbb{E}_{y\sim P_Y}k(P_X,P_{X|Y=y}) + \mathbb{E}_{y\sim P_Y}k(P_{X|Y=y},P_{X|Y=y}) && (2.77)\\ &= \mathbb{E}_{y\sim P_Y}\mathbb{E}_{P_{x_1,x_2\sim P_{X|Y=y}}}k(x_1,x_2) - \mathbb{E}_{x_1,x_2\sim P_X}k(x_1,x_2) && (2.78) \end{aligned}$$

To my knowledge, this kernel-based measure of information has not been defined or 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, which 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 $g \in \mathcal{H}_{\mathcal{Y}}$*

$$\langle f, C_{XY}g\rangle_{\mathcal{H}_{\mathcal{X}}} = \mathbb{E}_{(x,y)\sim P}\left(f(x) - \mathbb{E}_{x'\sim P_X}f(x')\right)\left(g(y) - \mathbb{E}_{y'\sim P_Y}g(y')\right) \quad (2.79)$$

Based on the cross-covariance operator, we can define various measures of dependence and information. Here I only define the simplest one, the constrained covariance, or COCO:

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

$$COCO_{XY} = \sup_{\substack{f\in\mathcal{H}_\mathcal{X}, g\in\mathcal{H}_\mathcal{Y} \\ \|f\|_{\mathcal{H}_\mathcal{X}}=1, \|g\|_{\mathcal{H}_\mathcal{Y}}=1}} \mathbb{C}ov_{(x,y)\sim P}\left[f(x), g(y)\right] \tag{2.80}$$

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

$$COCO_{XY} = \|C_{XY}\|_2, \tag{2.81}$$

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

A more robust measure of dependence, the Hilbert Schmidt Information Criterion (HSIC) uses the Hilbert-Schmidt norm of the cross-covatiance operator[**?**]. Just as generalisations of Shannon's mutual information (eqn. (2.25)), kernel measures of dependence like COCO and HSIC have several useful properties. They are symmetric, and can be effectively estimated from empirical data [**?**].

However, as with eqn. (2.25), 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 to be defined over both $\mathcal{X}$ and $\mathcal{Y}$, and properties of the functional depend on both choices of kernels. In contrast (2.76) only requires a single kernel over $\mathcal{X}$.

Interestingly, the kernel value of information $\mathbb{I}_k\left[X \leftarrow Y\right]$ that I introduced based on the kernel score can also be interpreted in terms of a linear operator in the Hilbert space. I am not aware of any previous use of this operator before, and in referencing it I will use the name diversity operator.

Definition 8 (Diversity operator)**.** *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 $f, g\in\mathcal{H}$*

$$\left\langle f, D_{X|Y} g\right\rangle_\mathcal{H} = \mathbb{C}ov_{y\sim P_Y}\left[\mathbb{E}_{X|Y=y} f, \mathbb{E}_{X|Y=y} g\right] \tag{2.82}$$

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

$$\left\langle f, D_{X|Y} f\right\rangle_\mathcal{H} = \mathbb{V}_{y\sim P_Y}\left[\mathbb{E}_{x\sim P_{X|Y=y}} f(x)\right] \tag{2.83}$$

Equivalently, the operator can be defined in terms of mean elements or Hilbert-space embedding of the conditional and marginal distributions as follows:

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

$$D_{X|Y} = \mathbb{E}_{y\sim P_Y}\left(\mu_{X|Y=y} - \mu_X\right)\otimes\left(\mu_{X|Y=y} - \mu_X\right) \tag{2.84}$$

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

$$\begin{aligned}
&\left\langle f, \left(\mathbb{E}_{y\sim P_Y}\left(\mu_{X|Y=y} - \mu_X\right)\otimes\left(\mu_{X|Y=y} - \mu_X\right)\right) g\right\rangle && (2.85)\\
=&\mathbb{E}_{y\sim P_Y}\left\langle f, \left(\left(\mu_{X|Y=y} - \mu_X\right)\otimes\left(\mu_{X|Y=y} - \mu_X\right)\right) g\right\rangle && (2.86)\\
=&\left\langle f, \left(\mu_{X|Y=y} - \mu_X\right)\right\rangle\left\langle g, \left(\mu_{X|Y=y} - \mu_X\right)\right\rangle && (2.87)\\
=&\mathbb{E}_{y\sim P_Y}\left(\mathbb{E}_{X|Y=y} f(x) - \mathbb{E}_{x\sim P_X} f(x)\right)\left(\mathbb{E}_{X|Y=y} g(x) - \mathbb{E}_{x\sim P_X} g(x)\right) && (2.88)\\
=&\mathbb{C}\mathrm{ov}_{y\sim P_Y}\left[\mathbb{E}_{X|Y=y} f, \mathbb{E}_{X|Y=y} g\right] && (2.89)\\
=&\left\langle f, D_{X|Y} g\right\rangle_\mathcal{H} && (2.90)
\end{aligned}$$

□

Using this alternative definition it is easy to see that the kernel value of information as defined in eqn. (2.76) can be expressed as the trace of the diversity operator (which in turn is the same as the Hilbert-Schmidt norm of the squareroot of the operator):

$$\begin{aligned}
\mathbb{I}_k\left[X \leftarrow Y\right] &= \mathbb{E}_{y\sim\mu_Y}\|\mu_X - \mu_{X|Y=y}\|_2^2 && (2.91)\\
&= \mathbb{E}_{y\sim\mu_Y}\operatorname{trace}\left\langle \mu_X - \mu_{X|Y=y}, \mu_X - \mu_{X|Y=y}\right\rangle && (2.92)\\
&= \mathbb{E}_{y\sim\mu_Y}\operatorname{trace}\left(\mu_X - \mu_{X|Y=y}\right)\otimes\left(\mu_X - \mu_{X|Y=y}\right) && (2.93)\\
&= \operatorname{trace}\mathbb{E}_{y\sim\mu_Y}\left(\mu_X - \mu_{X|Y=y}\right)\otimes\left(\mu_X - \mu_{X|Y=y}\right) && (2.94)\\
&= \operatorname{trace} I_{X|Y} && (2.95)\\
&= \|I_{X|Y}^{1/2}\|_{HS} && (2.96)
\end{aligned}$$

It would be interestnig future direction to investigate whether this information criterion has any connections to COCO and HSIC, or indeed if it inherits any of their nice convergence properties.

2.2.6 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

$$\begin{aligned}
S_{k,spherical}(x,P) &:= \|\mu_{\delta_x}\|_{\mathcal{H}} - \frac{\mu_P(x)}{\|\mu_P\|_{\mathcal{H}}} && (2.97)\\
&= \|\mu_{\delta_x}\|_{\mathcal{H}}\left(1-\cos(\mu_{\delta_x},\mu_P)\right) && (2.98)\\
&= \sqrt{k(x,x)} - \frac{\mathbb{E}_{x'\sim P}k(x,x')}{\sqrt{\mathbb{E}_{x,x'\sim P}k(x,x')}}, && (2.99)
\end{aligned}$$

The scoring rule gives rise to the following entropy functional:

$$\begin{aligned}
\mathbb{H}_{k,spherical}\left[P\right] &= \mathbb{E}_{x\sim P}\|\mu_{\delta_x}\|_{\mathcal{H}} - \|\mu_P\|_{\mathcal{H}} && (2.100)\\
&= \mathbb{E}_{x\sim P}\sqrt{k(x,x)} - \sqrt{\mathbb{E}_{x,x'\sim P}k(x,x')} && (2.101)
\end{aligned}$$

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:

$$\begin{aligned}
d_{k,spherical}\left[P\|Q\right] &= -\mathbb{E}_{x\sim Q}\frac{\mu_P}{\|\mu_P\|_{\mathcal{H}}} + \|\mu_P\|_{\mathcal{H}} && (2.102)\\
&= \|\mu_P\|_{\mathcal{H}}\left(1-\cos(\mu_P,\mu_Q)\right) && (2.103)\\
&= \sqrt{\mathbb{E}_{x,x'\sim P}k(x,x')} - \frac{\mathbb{E}_{x\sim P}\mathbb{E}_{x'\sim Q}k(x,x')}{\sqrt{\mathbb{E}_{x,x'\sim Q}k(x,x')}} && (2.104)
\end{aligned}$$

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

$$0 < d_k\left[P \| Q\right] \tag{2.105}$$

$$0 < \|\mu_P\|_{\mathcal{H}}^2 + \|\mu_Q\|_{\mathcal{H}}^2 - 2\left\langle \mu_P, \mu_Q \right\rangle_{\mathcal{H}} \tag{2.106}$$

$$\left\langle \mu_P, \mu_Q \right\rangle < \frac{1}{2}\left(\|\mu_P\|_{\mathcal{H}}^2 + \|\mu_Q\|_{\mathcal{H}}^2\right) \leq \|\mu_P\|_{\mathcal{H}}\|\mu_Q\|_{\mathcal{H}} \tag{2.107}$$

$$\cos(\mu_P, \mu_Q) = \frac{\left\langle \mu_P, \mu_Q \right\rangle_{\mathcal{H}}}{\|\mu_P\|_{\mathcal{H}}\|\mu_Q\|_{\mathcal{H}}} < 1 \tag{2.108}$$

Thus,

$$d_{k,spherical}\left[P \| Q\right] = \|\mu_P\|_{\mathcal{H}}\left(1 - \cos(\mu_P, \mu_Q)\right) > 0 \tag{2.109}$$

□

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.

The spherical kernel score also has an interesting intuitive meaning in terms of test functions Gaussian processes

Proposition 1. *Let P, Q be probability distribuitons over the domain $\mathcal{X}$ and $\mathcal{H}$ a RKHS with associated kernel function k. Let GP denote a standard Gaussian process in the Hilbert space. Then the following equality holds:*

$$d_{k,spherical}\left[P \| Q\right] = \|\mu_P\|_{\mathcal{H}} \mathbb{P}_{f \sim GP}\left[sign(\mathbb{E}_{x \sim P} f(x)) \neq sign(\mathbb{E}_{x \sim Q} f(x))\right] \tag{2.110}$$

Thus, the divergence function (2.104) is related to the probability that the expectation of a randomly drawn test function f has the same sign when the expectation is taken under P or under Q. Intuitively, the more smooth functions one can find whose expectation under P is positive but under Q is negative, the more different P and Q are.

I am not aware of any previous definition or mention of the spherical scoring rule or it's associated divergence in either the statistics or machine learning literature. It is unclear whether this intuitive divergence function provides any advantages over, say MMD, in practical applications, or whether efficient empirical estimators exist.

2.2.7 Scoring rules and Bayesian 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 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. This belief is usually formed by probabilistic inference. Given our forecast P we can choose an action that minimises the expected loss:

$$a_P^* = \operatorname*{argmin}_{a \in \mathcal{A}} \mathbb{E}_{x \sim P} \ell(x, a) \tag{2.111}$$

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 $X = x$.

$$S_\ell(x, P) = \ell(x, a_P^*) \tag{2.112}$$

This function only depends on the true state x and the forecast P, hence it is a scoring rule. The generalised entropy that this scoring rule defines is otherwise known as the Bayes-risk of the decision problem:

$$\mathbb{H}_\ell[P] := \mathbb{E}_{x\sim P}\ell(x, a^*_P) \tag{2.113}$$
$$= \min_{a\in\mathcal{A}} \mathbb{E}_{x\sim P}\ell(x, a) \tag{2.114}$$

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:

$$d_\ell[P\|Q] = \mathbb{E}_{x\sim P}\ell(x, a^*_Q) - \min_{a\in\mathcal{A}} \mathbb{E}_{x\sim P}\ell(x, a) \tag{2.115}$$

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:

$$\mathbb{E}_{x\sim P}|f(x)| \geq \mathbb{H}_{Shannon}[P]. \tag{2.116}$$

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, and it becomes more exact as the dictionary size increases.

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.

$$\ell(x, a) = \sup_{f\in\mathcal{F}} (f(x) - a(f))^2 \tag{2.117}$$

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

$$a^*_P(f) = \mathbb{E}_{x\sim P} f(x) \tag{2.118}$$

Thus, we can define the following scoring rule S:

$$S(x, P) = \ell(x, a^*_P) = \sup_{f\in\mathcal{F}} (f(x) - \mathbb{E}_{x\sim P} f(x))^2 \tag{2.119}$$

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.

2.3 Summary

In this chapter I introduced the framework of scoring rules and strictly proper scoring rules. The framework allows us to define meaningful generalisations of entropy, divergence and the value of information, which are useful in a variety of tasks such as approximate inference and experiment design. I have also shown how the framework of scoring rules and Bayesian decision theory are intimately connected.

In addition to the classic examples – logarithmic, Brier, spherical scores – I reviewed information quantities that one can define based on reproducing kernel Hilbert spaces. These rich classes of scoring rules have been used by the machine learning community, where they were derived from different first principles.

Establishing connections between these quantities and strictly proper scoring rules allows us to understand their general properties, and to introduce generalisations such as the kernel value of information or the spherical kernel score. In the following chapter I further examine the properties of these scoring rules in terms of the Riemannian geometry they imply over probability distributions.

Chapter 3

Information geometry

Strictly proper scoring rules and associated Bregman divergences determine an *information geometry* of probability distributions. In this section I aim to develop an understanding of the differences between various scoring rules and by visualising these geometric structures each of them give rise to.

The central object of interest in information geometry is the smooth Riemannian manifold of probability distributions that the divergence function induces, called the *statistical manifold*. In this section, our goal is to create low-dimensional maps of these statistical manifolds in such a way that distances measured between points on the map correspond to geodesic distances measured on the manifold as precisely as possible. In particular, we will focus on one and two-dimensional maps of families of distributions parametrised by at most two continuous parameters.

First, it is important to note that a perfect embedding of this sort does not always exist. As an illustration, think of the well-known practical problem of creating a two-dimensional map of the surface of the Earth. The surface of Earth is approximately a sphere, which is a smooth two-dimensional Riemannian manifold, just as the statistical manifolds we would like to map in this chapter. The sphere 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 up at certain places, but even then, the embedding is only approximate, and distances are only correct locally. This is why on Google maps Finland appears about twice as large as France, even though in reality it is only about half the size of France. 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 developeded in the machine learning community. These approximate embedding procedures numerically find a low-dimensional map that best represents distances on the statistical manifold defined by a particular scoring rule and divergence, optimising an appropriately defined objective function. In this chapter I will employ an algorithm analogous to the ISOMAP algorithm [**?**], originally developed for dimensionality reduction and visualisation of high-dimensional data.

This chapter is organised as follows. I will first review crucial mathematical concepts in information geometry. Then I will introduce a general algorithm for numerically mapping out Riemannian manifolds induced by strictly proper scoring rules. Finally, I show maps of one- and two-parameter exponential families of distributions with respect to various divergence metrics introduced in chapter **??**.

3.1 Information 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

$$ds^2(P) = \frac{1}{2}\left\langle P, \ddot{H}(P)P \right\rangle, \tag{3.1}$$

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

$$ds^2 = \frac{1}{2}\sum_{i,j}\frac{\partial^2 H}{\partial p_i \partial p_j} dp_i dp_j. \tag{3.2}$$

The local distance between distributions is therefore controlled by the curvature of the entropy function: the higher the curvature, the more amplified the distances are locally. The following Taylor expansion shows how this local distance is related to the Bregman divergences defined by the entropy function:

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:*

$$d_H[P\|P + dP] \approx \frac{1}{2}\sum_{i,j}\frac{\partial^2 \mathbb{H}_S}{\partial p_i \partial p_j} dp_i dp_j \approx d_H[P + dP\|P] \tag{3.3}$$

Proof. We prove the left-hand equation first:

$$\begin{aligned}
\frac{\partial}{\partial q_i} d_H[P\|Q] &= -\frac{\partial}{\partial q_i} H(Q) + \frac{\partial}{\partial q_i}\langle \nabla_Q H(Q), Q - P\rangle && (3.4)\\
&= -\frac{\partial}{\partial q_i} H(Q) + \frac{\partial}{\partial q_i}\sum_j \frac{\partial}{\partial q_j} H(Q)(q_j - p_j) && (3.5)\\
&= -\frac{\partial}{\partial q_i} H(Q) + \frac{\partial}{\partial q_i} H(Q) + \sum_j \frac{\partial^2}{\partial q_i \partial q_j} H(Q)(q_j - p_j) && (3.6)\\
&= \sum_j \frac{\partial^2}{\partial q_i \partial q_j} H(Q)(q_j - p_j) && (3.7)
\end{aligned}$$

hence by first order Taylor expansion around P:

$$\begin{aligned}
d_H[P\|P + dP] &\approx d_H[P\|P + dP] + \frac{1}{2}\left\langle dP, \nabla_Q d_H[P\|Q]|_{Q=P} \right\rangle && (3.8)\\
&= \frac{1}{2}\sum_i dp_i \sum_j \frac{\partial^2 H(P)}{\partial p_i \partial p_j} dp_j && (3.9)\\
&= \frac{1}{2}\sum_{i,j}\frac{\partial^2 H(P)}{\partial p_i \partial p_j} dp_i dp_j && (3.10)
\end{aligned}$$

Similarly in the other direction

$$\frac{\partial}{\partial q_i} d_H\left[Q\|P\right] = \frac{\partial}{\partial q_i} H(Q) + \frac{\partial}{\partial q_i}\left\langle \nabla_P H(P), P - Q\right\rangle \tag{3.11}$$

$$= \frac{\partial}{\partial q_i} H(Q) + \frac{\partial}{\partial q_i}\sum_j \frac{\partial}{\partial p_j} H(P)(p_j - q_j) \tag{3.12}$$

$$= \frac{\partial}{\partial q_i} H(Q) - \frac{\partial}{\partial p_i} H(P) \tag{3.13}$$

$$= \dot{H}(Q) - \dot{H}(P) \tag{3.14}$$

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

$$\frac{\partial}{\partial dP} d_H\left[P + dP\|P\right] = \dot{H}(P + dP) - \dot{H}(P) \approx \ddot{H}(P)dP \tag{3.15}$$

therefore, via Taylor expansion we get that for small dP

$$d_H\left[P\|P + dP\right] \approx \frac{1}{2}\left\langle dP, \ddot{H}(P)dP\right\rangle \tag{3.16}$$

$$= \frac{1}{2}\sum_{i,j}\frac{\partial^2 H(P)}{\partial p_i \partial p_j} dp_i dp_j \tag{3.17}$$

□

Hence, the distance on the manifold can be approximated locally as half the squareroot of the divergence function. Even though a Bregman divergence function is generally asymmetric, for infinitesimally small differences it becomes symmetric, and therefore it does not matter which direction we use if we want to approximate local distances on the manifold. Below we will use the following local approximation:

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.*

$$\textit{distance}(P, Q) \approx \sqrt{d_S\left[P\|Q\right] + d_S\left[Q\|P\right]} \tag{3.18}$$

Another core concept in information geometry is that of geodesics and geodesic distances between distributions:

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*

$$l(\mathcal{P}) = \int_0^1 \sqrt{\left\langle \dot{P}(t), \ddot{H}(P(t))\dot{P}(t)\right\rangle} dt \tag{3.19}$$

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.

3.2 Approximate embedding of Riemannian manifolds

Our goal in the rest of this chapter is going to be to create maps of statistical manifolds, such that the Euclidean distances between distributions on the maps approximate geodesic distances on the manifold as faithfully as possible. However, geodesic distances on general, non-trivial 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. (3.19)) 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, if we want to create maps of arbitrarily complex statistical manifolds, we will have to resort to numerical approximations to geodesic distances. To sidestep both computational problems at once, we are going to restrict geodesic paths between distributions P and Q, to paths on a graph of neighbouring points on the manifold.

Consider a graph, whose vertices are points on the manifold and we draw an edge between pairs of points that are close enough to each other. We will refer to this as a local neighbourhood graph. As neighbours in this graph are assumed to be close, the geodesic distance between neighbours is approximately the same as the local squared distance, and can be approximated using Eqn. 3.18. Let us define this approximate distance between neighbours as the weight of the edge between them. The length of any path that travels trough a series of vertices of the neighbourhood graph can then be approximated as the sum of edge weights between subsequent points the path travels trough. It is easy to see that if the vertices of this neighbourhood graph cover the manifold densely enough, path lengths on this graph can be used to approximate the length of any smooth path on the manifold. Computing geodesic distances then amounts to finding the shortest path on the neighbourhood graph, for which polynomial time algorithms exist.

This idea of using shortest paths in a local neighbourhood graph as approximation to geodesic distances has been used in the context of manifold learning and forms the basis of the ISOMAP algorithm [**?**]. In ISOMAP, a set of points that are assumed to conform to a manifold are given to us as the input to the algorithm, and we have to recover the latent geometric structure. In this section, we are free to chose a set of distributions that will constitute the vertices of the neighbourhood graph. In most cases as we will visualise manifolds of parametric classes of distributions it is practical to choose a uniformly or logarithmically spaced grid in parameter space, where the neighbourhood structure is naturally defined by the grid itself.

We will follow the following procedure to produce a map of the statistical manifolds 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 is the squareroot of the symmetrised divergence between the two distributions, as in Eqn. (3.18).

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

3. use metric multidimensional scaling with the approximate geodesic distance matrix as input to embed the set of distributions as points a low-dimensional Euclidean space.

3.2.1 Bernoulli distributions

Let us first look at the simple and special case of one dimensional statistical manifolds of Bernoulli distributions. Bernoulli random variables, often referred to as biased coin-flips, have a binary outcome: positive with probability p and negative with probability $1 - p$. The probability p is a real valued parameter, hence Bernoulli distributions conform to a one-dimensional manifold.

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, never to a circle. So the only difference between various statistical manifolds is how the real line is stretched and compressed at various locations.

In Figure 3.1 I illustrate the differences between the statistical manifolds induced by the logarithmic, Brier and spherical scoring rules using the numerical embedding technique outlined above. As KL divergence between p and q is not bounded and diverges for $q \to 0$ and $q \to 1$, the statistical manifold corresponding to this divergence will span the whole extended real line $[-\infty, \infty]$. The Brier and spherical divergences are bounded, hence the manifold becomes an finite interval of $\mathbb{R}$.

To visualise the differences between these manifolds, I started with a linearly spaced grid of 33 parameter values in the interval $[0+\epsilon, 1-\epsilon]$, with $\epsilon = 10^{-3}$. In this interval of parameter values all three divergences are bounded, so when applying the ISOMAP procedure, this part of the manifold gets mapped to a finite segment in each of the three cases. As scoring rules - and divergences - are equivalent up to a multiplicative constant, we can scale the resulting intervals arbitrarily to be of the same length. Figure 3.1 shows the resulting manifold structure for the three divergences. As the Brier divergence between p and q is the squared Euclidean distance between p and q, the geodesic distance on this statistical manifold is simply the Euclidean distance. Therefore, as expected, the uniformly spaced grid of probabilities is represented as a uniformly spaced grid of points on this map of the manifold.

As we can see, compared to the Brier score, the KL divergence is more sensitive to differences in very small (close to 0) and large (close to 1) probabilities, but puts less emphasis on discriminating between intermediate values close to $p = 0.5$. Remember that the statistical manifold corresponding to the KL divergence extends to $-\infty$ and ∞ and in Figure 3.1 we only show a segment from it.

When using the KL divergence or the log-score in practical situations, such as to train binary classifiers, we should therefore expect that much of the statistical power is going to be spent on faithfully representing small probabilities, as this is where the resolution of the divergence is highest. This behaviour 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 modelling 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 would spend most effort on modelling how small the small probabilities are exactly.

Figure 3.1 also shows the statistical manifold induced by the spherical score. As we can see, relative to the Brier score, the spherical score has a larger resolution among intermediate probabilities close to 0.5 than around small probabilities closer to 0 and 1. Therefore in applications where modelling probabilities closer to 0.5 is important, the spherical score may be an even more appropriate choice than the Brier score.

In Figure 3.2 I plotted the local distance $\sqrt{\ddot{\mathbb{H}}_S[p]}$ as a function of p for the three different scores illustrated in Figure **??**. Higher value of this curve means that the scoring rule has a "higher resolution" locally. It is another visualisation that allows us to observe that relative to the Brier score, the logarithmic score focuses more on probabilities close to 0 and 1, whilst the spherical divergence focuses more on probabilities close to 0.5.

3.2.2 Gaussian distributions

Gaussian distributions are probably the most important family of distributions due to their convenient analytical properties. They are often used in density estimation, regression, approximate inference and more advanced non-parametric models such as Gaussian process regression.

The KL divergence between two univariate Gaussian distributions is available in a closed form and is given by the following formula:

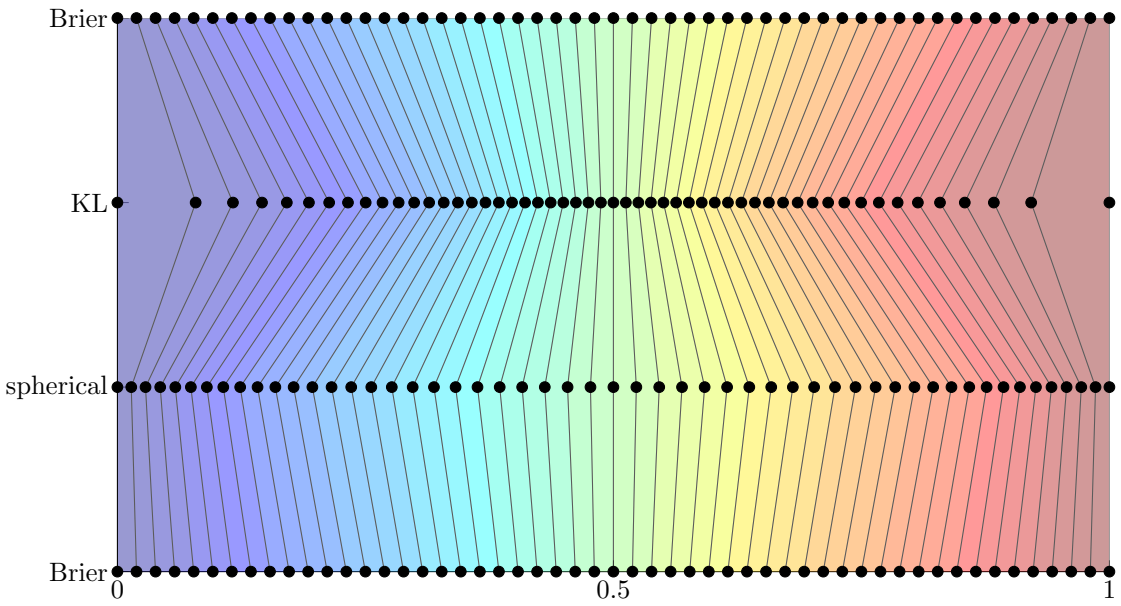

Figure 3.1: Illustration of the differences between the Brier, spherical, and KL divergences between single parameter Bernoulli distributions. Each horizontal line of dots shows the embedding Bernoulli distributions corresponding to an uniform grid of parameter values between $0 + \epsilon$ and $1 - \epsilon$ on the statistical manifold induced by (from top to bottom) the Brier, KL, spherical and again the Brier score. Dots representing the same distributions on the different manifolds are connected. This, together with colouring, highlights the differences between the manifolds. The Brier divergence is equivalent to the squared Euclidean distance between parameter values, therefore when mapped by Brier divergence, parameters are evenly spaced along the line segment *(top, bottom)*. The KL divergence places emphasis on discriminating 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 full manifold of Bernoulli distributions stretches to the entire real line. By contrast, the spherical score focuses more on probability values around 0.5.

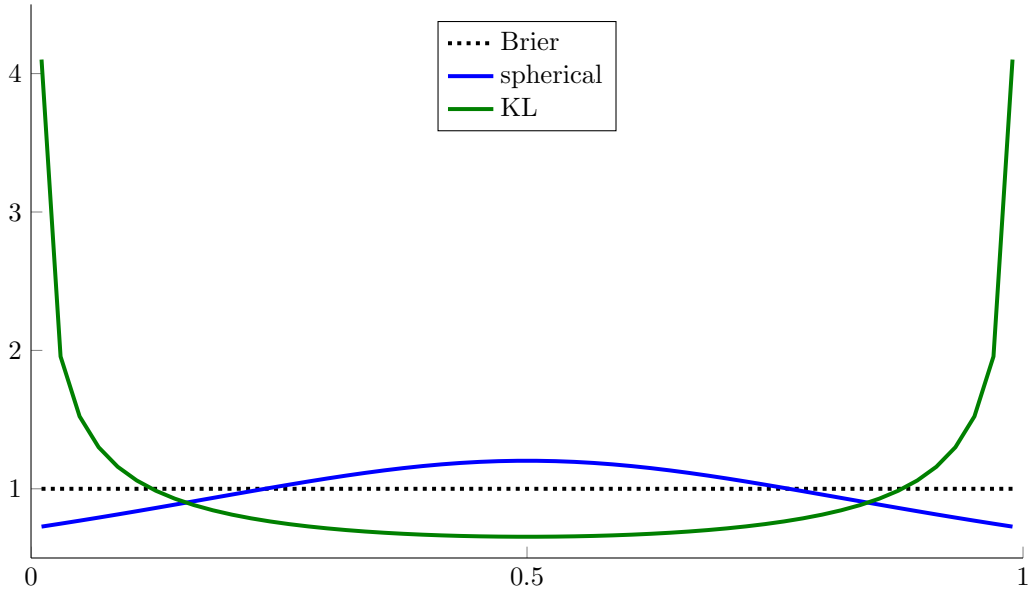

Figure 3.2: Illustration of the differences between local distances on statistical manifolds of Bernoulli distributions. Each line shows the magnitude of the local distance on the manifold relative to the Euclidean distance as a function of the parameter value. Distance on the Brier manifold is equivalent to the Euclidean distance, hence it's relative magnitude is constant **??**. The KL divergence gives rise to increasing local distances as the parameter approaches 0 and 1 **??**. The spherical score induces a local distance that is largest at 0.5 **??**.

$$d_{KL}\left[\mathcal{N}_{\mu_1,\sigma_1} \| \mathcal{N}_{\mu_2,\sigma_2}\right] = \frac{(\mu_1-\mu_2)^2}{2\sigma_2^2} + \frac{1}{2}\left(\frac{\sigma_1^2}{\sigma_2^2} - 1 - \log\frac{\sigma_1^2}{\sigma_2^2}\right) \tag{3.20}$$

In this case as Gaussian distributions have two parameters, the distributions are going to conform to a two dimensional statistical manifold, as illustrated in Figure **??**. We used the ISOMAP technique on a linearly spaced grid of parameters to produce this approximate embedding. We can observe that assuming that P and Q have the same mean, the larger their variance, the easier it becomes to distinguish between them. Otherwise the manifold structure is symmetrical.

The main purpose of this section is to visualise differences between the geometries induced by various divergence measures over the same set of distributions. A particularly interesting divergence that we will use in subsequent chapters is that induced by the (quadratic) kernel scoring rule from (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 divergence between univariate Gaussians can be computed in closed form.[**?**] For the squared exponential kernel $k_\ell(x,x') = {}^1\!/\!_\ell \exp(-\frac{(x-x')^2}{\ell^2})$ the divergence is given by the following formula:

$$d_{k_\ell}\left[\mathcal{N}_{\mu_1,\sigma_1} \| \mathcal{N}_{\mu_2,\sigma_2}\right] = \frac{1}{\sqrt{\ell^2+2\sigma_1^2}} + \frac{1}{\sqrt{\ell^2+2\sigma_2^2}} - \frac{2}{\sqrt{\ell^2+\sigma_1^2+\sigma_2^2}} \exp\left(-\frac{(\mu_1-\mu_2)^2}{2(\ell^2+\sigma_1^2+\sigma_2^2)}\right) \tag{3.21}$$

The above formula can be derived from the following general expression for the inner product between mean embeddings:

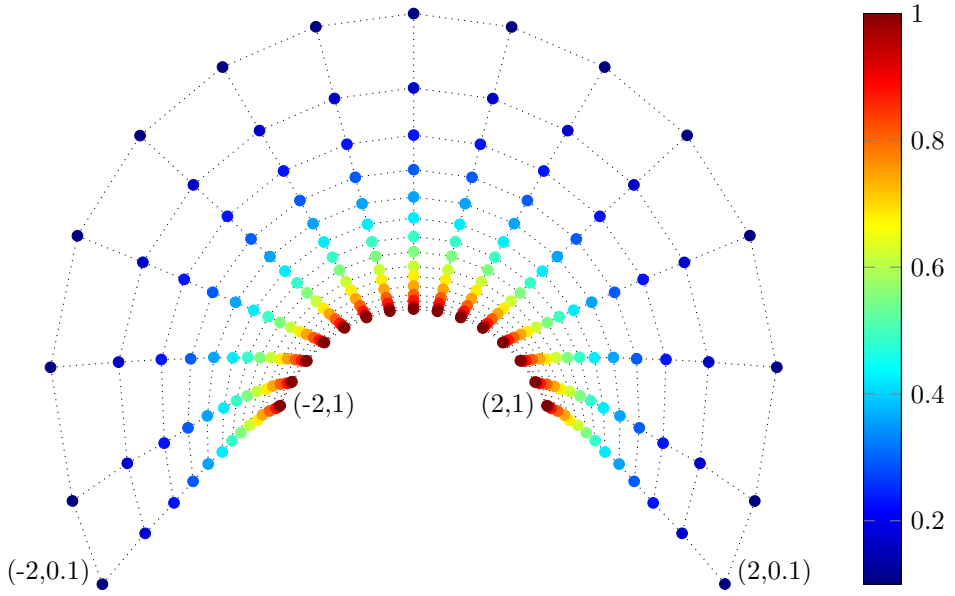

Figure 3.3: Map of Normal distributions on the statistical manifold induced by the logarithmic score and KL divergence. Distributions are chosen from a uniform grid in parameter space, with mean ranging between -2 and 2 *(left to right)*, and standard deviation between 0.1 and 1 *(from outside inwards)*. The labels show distributions at the corners of this grid. Dots of the same colour show 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 characteristic fan-like structure.

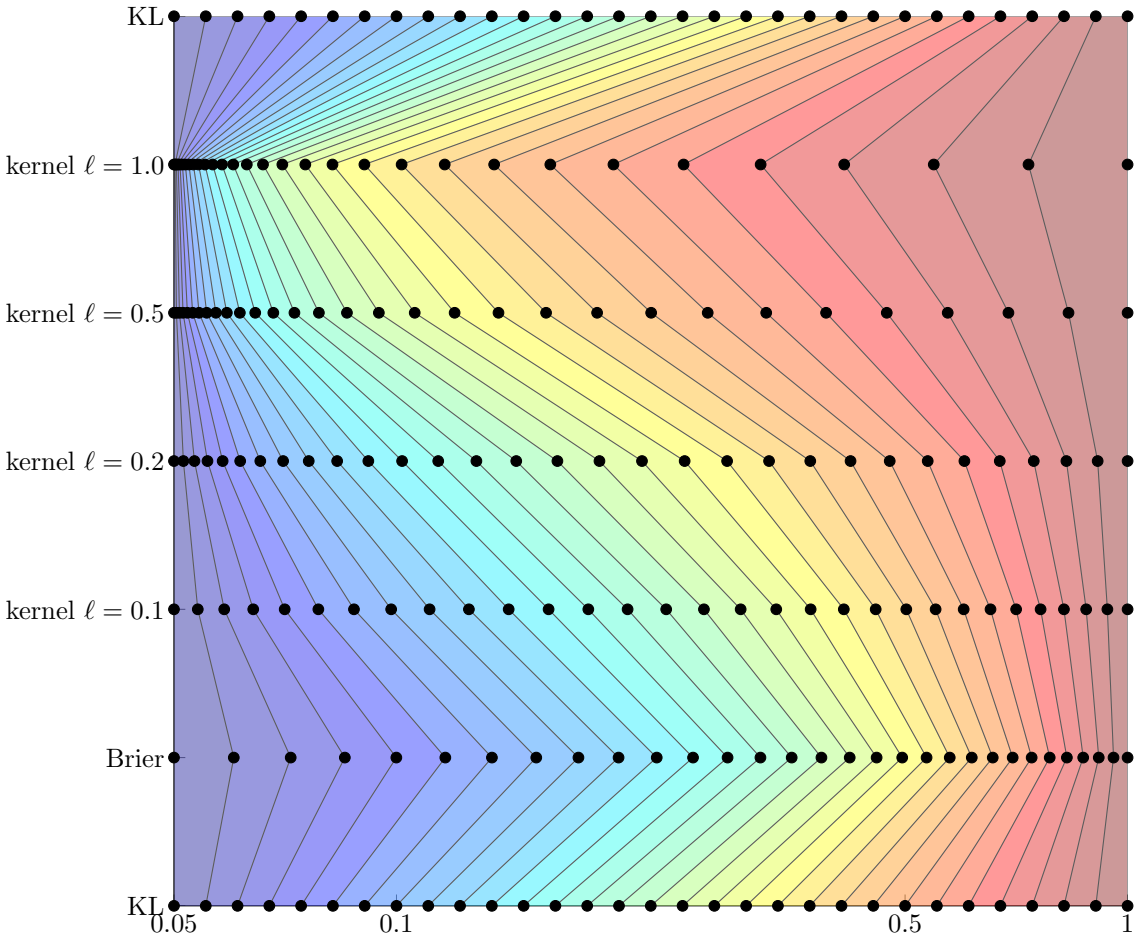

Figure 3.4: Illustration of the differences between the statistical manifolds of Normal distributions induced by the KL, Brier and kernel divergences. Each horizontal line of points shows the one-dimensional manifold of zero-mean Gaussians. The dots correspond to distributions with logarithmically spaced variance between $\sigma_{min} = 0.05$ and $\sigma_{max} = 1$. When mapped according to the KL divergence, these distributions become evenly spaced *(top,bottom)*. Compared to the KL, the Brier score *(second from bottom)* places more emphasis on disriminating between narrower distributions. In this range $0.05 \leq \sigma \leq 1$ the kernel divergence with bandwidth $\ell = 0.1$ *(third from bottom)* approximately mimics the behaviour of the KL divergence. As we use the kernel score with increasing brandwith *(from bottom to top)*, we can see that the focus shifts from narrow distributions towards distributions with larger variance.

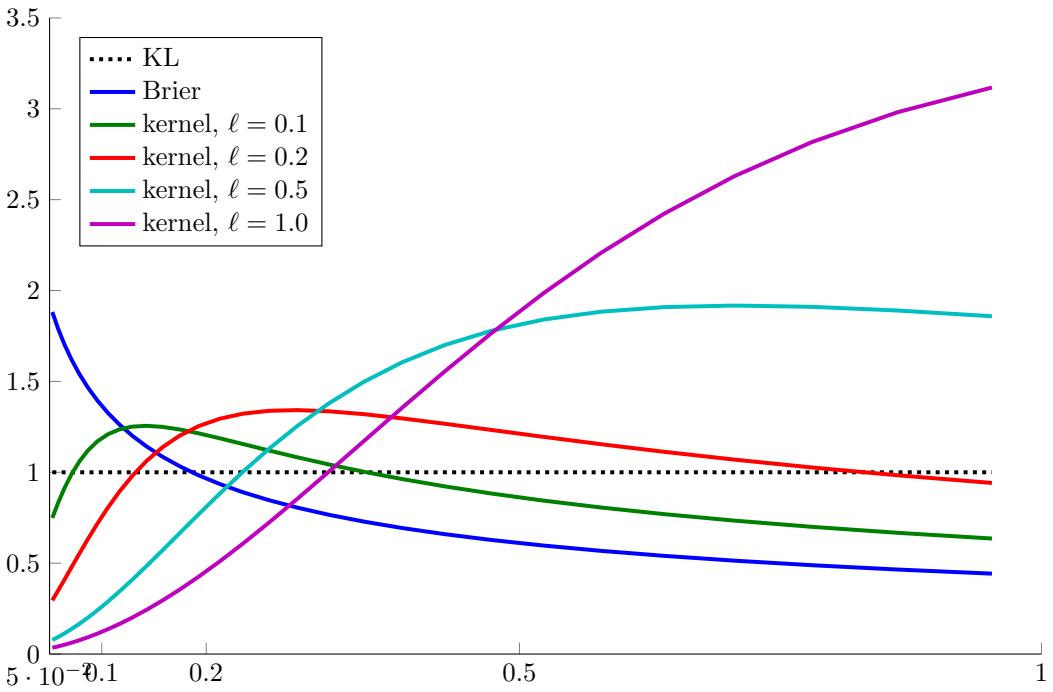

Figure 3.5: Illustration of the differences between local distances induced by various scoring rules on the statistical manifold of zero-mean Normal distributions. Each line shows the magnitude of the local distance on each manifold relative to that induced by the KL divergence as a function of variance. Relative to the KL, the kernel divergence induces distances that are magnified around a region depending on the bandwidth of the kernel. As the bandwidth increases, this magnified region shifts towards distributions with larger variance.

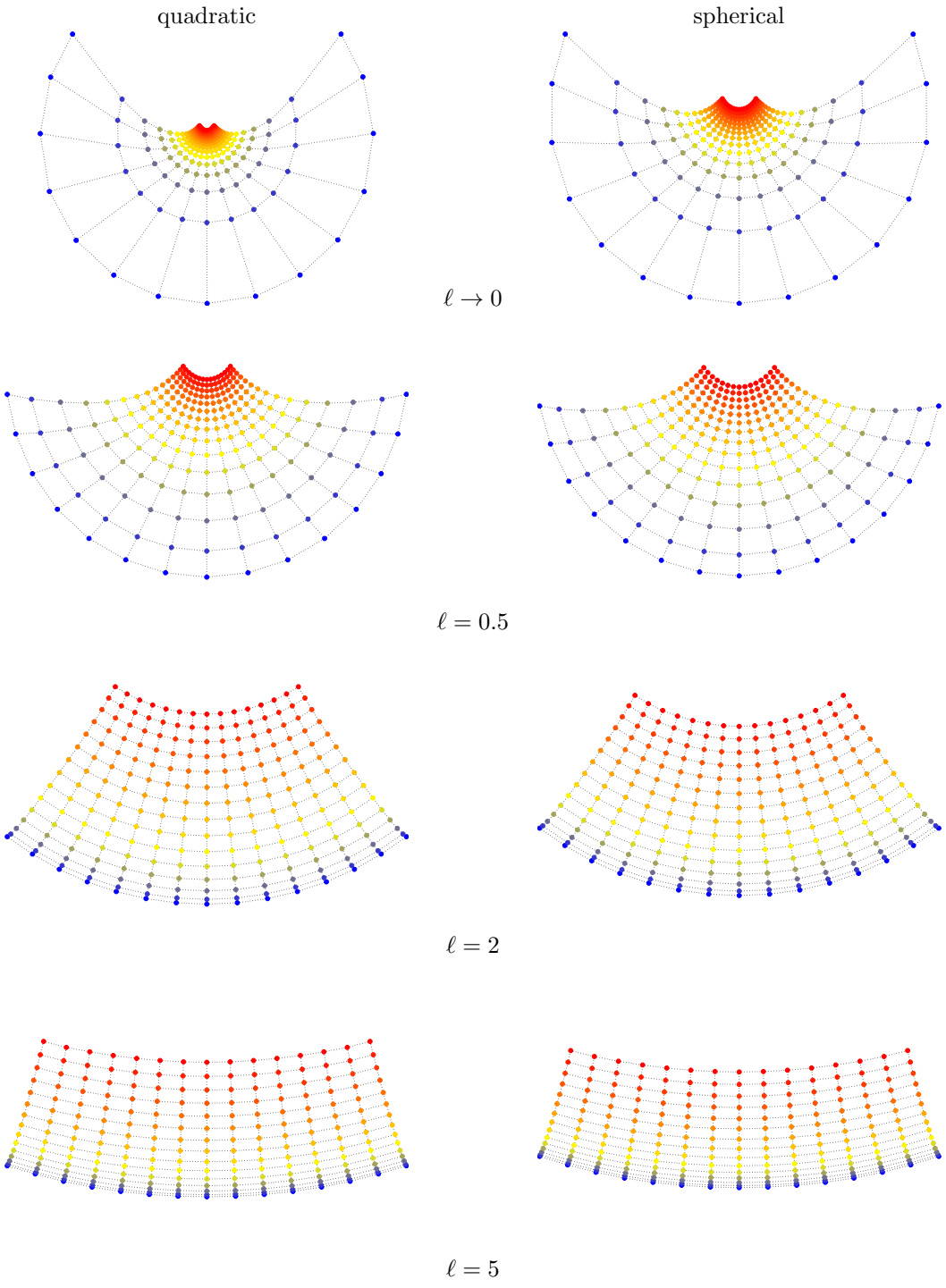

Figure 3.6: Maps of the statistical manifold induced by the (quadratic) kernel score and the spherical kernel score over Gaussian distributions for different setting of the kernel bandwidth parameter. The two panels in the top row $\ell \to 0$ correspond to the limiting cases of the Brier score and the spherical score. It can be seen that as the bandwidth increases, both scores shift their sensitivity to distributions with higher variance (red). For equal bandwidth, the spherical kernel score is more sensitive to distributions with larger standard deviation.

$$\left\langle \mu_{\mathcal{N}(\mu_1,\sigma_1)}, \mu_{\mathcal{N}(\mu_2,\sigma_2)} \right\rangle_{k_\ell} = \mathbb{E}_{x\sim\mathcal{N}(\mu_1,\sigma_1)}\mathbb{E}_{x'\sim\mathcal{N}(\mu_1,\sigma_1)} k_\ell(x,x') \tag{3.22}$$

$$= \frac{1}{\sqrt{\ell^2+\sigma_1^2+\sigma_2^2}} \exp\left(-\frac{(\mu_1-\mu_2)^2}{2(\ell^2+\sigma_1^2+\sigma_2^2)}\right) \tag{3.23}$$

The first fact one may observe is that unlike the KL divergence, the kernel divergence is bounded from above by $2/\ell$. This upper bound is approached when computing divergence between two infinitesimally narrow Gaussians $\sigma_1, \sigma_2 \approx 0$ that are far apart $|\mu_1 - \mu_2| > 0$. The divergence is also bounded from below by 0 and it is 0 exactly when the two distributions are identical, confirming that this kernel function gives rise to a strictly proper scoring rule.

The Brier score is a special case of this divergence as the lengthscale of the kernel ℓ decreases to 0. In that case we obtain the following expression:

$$d_{Brier}\left[\mathcal{N}_{\mu_1,\sigma_1} \| \mathcal{N}_{\mu_2,\sigma_2}\right] = \frac{1}{\sqrt{2\sigma_1^2}} + \frac{1}{\sqrt{2\sigma_2^2}} - \frac{2}{\sqrt{\sigma_1^2+\sigma_2^2}} \exp\left(-\frac{(\mu_1-\mu_2)^2}{2(\sigma_1^2+\sigma_2^2)}\right) \tag{3.24}$$

We can immediately see that unlike the kernel score with a positive lengthscale, the Brier score is not bounded from above. It diverges for very small values of the variances σ_1 and σ_2. It is still non-negative and strictly proper.

To illustrate the differences between the various divergences between Gaussian distributions, we first applied the ISOMAP embedding technique to the one-dimensional manifold of zero-mean Gaussians, whose sole free parameter is the standard deviation. I chose a logarithmically spaced grid of standard deviation values, then used the ISOMAP algorithm to embed the distributions on the real line. The logarithmic spacing is useful as the KL divergence now depends only on the difference in the logarithm of variances, therefore when these distributions are embedded according to the KL divergence, we expect to get a uniform, linearly spaced grid.

Figure **??** compares the statistical manifold induced by the KL and Brier divergences, as well as by the kernel divergence with different choices of the kernel bandwidth parameter ℓ. As expected, when the KL divergence is used the numerical algorithm spreads the distributions uniformly on the real line. We can see that compared to the KL divergence, the Brier divergence is more sensitive to differences between narrow distributions, whose standard deviation is small. In case of the kernel score, with increasing kernel brandwith the focus shifts from narrow distributions towards distributions with larger variance. In the range mapped in this figure ($0.05 \leq \sigma \leq 1$) the kernel bandwidth $\ell = 0.1$ mimics the behaviour of the KL divergence the best.

For these distributions the KL divergence is scale-free: the divergence between two zero-mean Gaussians with variance $\sigma_1 = 0.05$ and $\sigma_2 = 0.1$ is the same as the divergence between $\sigma_1 = 0.5$ and $\sigma_2 = 1$. The kernel score on the other hand has a characteristic brandwith, and is therefore not scale free: when the brandwith is chosen to be $\ell = 1$, the largest shown in Figure **??**, the distance between $\sigma_1 = 0.05$ and $\sigma_2 = 0.1$ is only about one tenth of the distance between $\sigma_1 = 0.5$ and $\sigma_2 = 1$.

In Figure **??** I plotted the local distances on the various manifolds relative to distances induced by the logarithmic score. Higher values on the plot indicate a region where local distances are magnified in comparison to the KL divergence, which can be interpreted as a region in which the particular scoring rule is more sensitive to small differences. Observe how changing the kernel bandwidth shifts the most sensitive region of the kernel scoring rule.

These figures highlight how the choice of the kernel allows us to fine-tune properties of the divergences and the corresponding manifold. We can use this flexibility to tailor the divergence to our application [**?**]. However, as discussed in chapter **??** this flexibility also poses a challenge in applications where there is no principled way of choosing kernel hyperparameters.

3.2.3 Gamma distributions

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

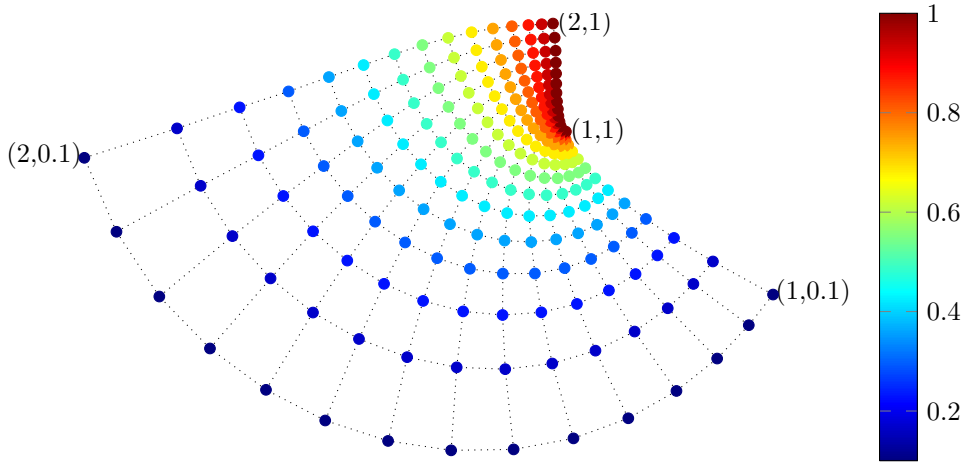

Figure 3.7: Map of Gamma distributions on the statistical manifold induced by the logarithmic score and KL divergence. To be comparable to the manifold of Normal distributions in Figure 3.3, the distributions are parametrised by their mean and standard deviation. Distributions are chosen from a uniform grid in this non-standard parameter-space, with their mean ranging between 1 and 2, and standard deviation between 0.1 and 1. For large values of variance *(yellow and red)* the manifold is asymmetric and dissimilar to that of Normal distributions. However, as variance decreases *(blue)*, by the central limit theorem Gamma distributions approach Gaussians of the same mean and variance, thus the manifold conforms to the fan-like shape that is characteristic of Gaussian distributions.

probability density function of Gamma distributions is as follows:

$$(x) = \beta^{\alpha} \frac{1}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\beta x) \tag{3.25}$$

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:

$$d_{KL}\left[\Gamma_{\alpha_1,\beta_1} \| \Gamma_{\alpha_2,\beta_2}\right] = (\alpha_1 - \alpha_2)\,\psi(\alpha_1) - \log\frac{\Gamma(\alpha_1)}{\Gamma(\alpha_2)} + \alpha_1 \log\frac{\beta_1}{\beta_2} + \alpha_1 \frac{\beta_2 - \beta_1}{\beta_1} \tag{3.26}$$

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

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

$$\mu = \frac{\alpha}{\beta} \tag{3.27}$$

$$\sigma^2 = \frac{\alpha}{\beta^2} \tag{3.28}$$

Solving for α and β in these equations we get

$$\alpha = \frac{\mu^2}{\sigma^2} \tag{3.29}$$

$$\beta = \frac{\mu}{\sigma^2} \tag{3.30}$$

Plugging these into Eqn. (3.26) 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. (3.26) becomes numerically imprecise, as it relies on look-up-table implementation of the Gamma (Γ) and bigamma (ψ) functions.

Chapter 4

Scoring rules for processes

The scoring rule framework presented in chapter **??** is already 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.

Stochastic processes have enjoyed a surge of interest from the machine learning community thanks to their use in nonparametric Bayesian inference. Nonparametric techniques based on processes like the Gaussian process, Dirichlet process, Indian buffet process have become ubiquitious in modern statistical models. It is interesting to investigate whether and how the scoring rule framework extends to these more general, infinite dimensional statistical models.

To analyse stochastic processes in the scoring rule framework I introduce a novel concept of marginal scoring rules. I show a number of examples of estimation techniques that do not fit the traditional scoring rule framework, but can be accommodated readily in this more general treatment. I will also argue that the notion of strictly proper scoring rules is insufficient when dealing with stochastic processes, and introduce an analogous property I call very strictly proper scoring.

4.1 Extensions of score matching for i. i. d. data

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:

$$\mathbb{E}_{x\sim P} S(x, P) \leq \mathbb{E}_{x\sim P} S(x, Q), \tag{4.1}$$

with equality only when $P = Q$.

In this definition we are concerned with the scoring rule's behaviour in expectation under the distribution P. In empirical terms taking an expectation is analogous to studying the limit of the empirical mean, 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 (i. i. d.) samples from a parametric distribution $P_{X|\theta}$, the parameter of the distribution can be consistently identified:

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:*

$$\hat{\theta}_N(x_1, \ldots, x_N) = \operatorname*{argmin}_{\theta\in\Theta} \sum_{n=1}^{N} S(x_n, P_{X|\theta}) \tag{4.2}$$

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.*

$$\lim_{N\to\infty} \hat{\theta}_N(x_1, \ldots, x_N) = \theta_0 \qquad P_{X|\theta_0}\text{- almost surely} \tag{4.3}$$

Score matching has been used to fit parametric models for decades. Maximum likelihood estimation is a typical example when the scoring rule is chosen to be the logarithmic score. In more modern work like [**?**], 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 has also been used in meteorology and epidemiology.

The score matching procedure assumes that one observes a sequence of i. i. d. observations repeatedly sampled from the same distribution P. However, observed data is not always i. i. d. . Often we want to score non-i. i. d. sequences or sets of variables, and we would still like our parameter estimates to be consistent in some sense. Examples of such non-i. i. d. estimation scenarios include:

1. Parameter estimation in non-parametric process models, such as in Gaussian process regression, studying the limit as the size of the training data grows
2. parameter estimation of a large Markov random field, studying constitency as the image size grows
3. Parameter estimation in phylogenetic tree models, studying behaviour as the number of species and the size of the tree grows
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 never observe a growing number of independent copies of the same random variable. Instead, we observe higher and higher dimensional marginals of the same single trajectory sampled randomly from a random process. We sample one infinitely large object from a random process once, and we study the behaviour as larger and larger parts of this object are revealed.

The above cases do not readily fit into the scoring rule framework which is specifically concerned with i. i. d. observations from the same, fixed dimensional distribution. To analyse these cases we can extend the scoring rule framework so that we allow scoring of partial observations. Let us consider the following definition of marginal scoring rules:

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* $y_I \in \mathcal{Y}^I$*.*

Marginal scoring rules allow one to calculate the score of a forecast on the basis of a partial observation y_I. For example, if Π is a distribution over images, a marginal scoring rule can assign a score to Π even if we only observe the right-hand corner of the image, and the rest remains hidden. This definition therefore allows us to study what happens as higher and higher dimensional marginals are revealed, as we desired.

Observe, that the traditional i. i. d. scoring rule framework can be expressed in terms of marginal scoring rules over i. i. d. processes. Instead of thinking about repeatedly sampling from a finite dimensional distribution, we can imagine sampling one infinitely long sequence first, but then revealing longer and longer sub-sequences from it. Let $\mathcal{X}$ be the set of natural numbers, and Π an i. i. d. process such that $\Pi_I(y_I) = \prod_{i\in I} P(y_i)$, then we can construct marginal scoring rules as follows:

$$R_I(y_I, \Pi) = \frac{1}{|I|} \sum_{i\in I_n} S(y_i, P), \tag{4.4}$$

where S is a strictly proper scoring rule over $\mathcal{Y}$.

The marginal scoring rules in eqn. (4.4) 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*

$$\operatorname*{argmin}_\theta R_{I_n}(y_{I_n}, \Pi_{\theta^*}) = \operatorname*{argmin}_\theta \frac{1}{|I_N|} \sum_{i \in I_N} S(y_i, P_\theta) \xrightarrow[N\to\infty]{P_{\theta^*}} \theta^* \tag{4.5}$$

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 of 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:*

$$\lim_{N\to\infty} \operatorname*{argmin}_\theta R_{I_N}(y_{I_N}, \Pi_\theta) = \theta^* \qquad \Pi_{\theta^*} \text{ almost surely} \tag{4.6}$$

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.

4.1.1 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 points are sampled from a one dimensional uniform distribution, then the spacing between neighbouring samples should be roughly the same. 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 $y_1, \ldots, y_N$ independent random variables on the real line $\mathbb{R}$. Let P be a distribution over $\mathbb{R}$ with cumulative distribution function F_P. Let $\pi : \{1, \ldots, N\} \mapsto \{1, \ldots, N\}$ 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 $y_{\pi_{N+1}} := 1$.*

$$R_N^{PS}(y_1, \ldots, y_N, P) = -\frac{1}{N+1} \sum_{n=0}^{N} \log\left(F_P\left(y_{\pi_{n+1}}\right) - F_P\left(y_{\pi_n}\right)\right) \tag{4.7}$$

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

$$\operatorname*{argmin}_\theta R_N^{PS}(y_1, \ldots, y_N, \Pi_{\theta^*}) \xrightarrow[N\to\infty]{P_{\theta^*}} \theta^* \tag{4.8}$$

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.

4.1.2 Decision theoretic scoring and F_β scores

Further examples of scores that do not decompose as in (4.4) 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 []:

$$F_\beta = (1-\beta^2)\frac{\text{precision}\cdot\text{recall}}{\beta^2\text{precision}+\text{recall}} \tag{4.9}$$

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 marginal scoring rules.

4.2 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 (4.4), 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:

$$R_{I_N}(x_{I_N}, \Pi) = -\frac{1}{|I_N|}\log P_{I_n}(y_{I_n}), \tag{4.10}$$

4.2.1 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:

$$\begin{aligned}\text{evidence}(\mathcal{M}) &= R_{log}(y_{1:N}, \Pi_{Y|\mathcal{M}}) && (4.11)\\ &= \log P_{Y_{1:N}|\mathcal{M}}(y_{1:N}) && (4.12)\\ &= \log \int \prod_{n=1}^{N} P_{Y_1|\theta}(y_n) P_{\theta|\mathcal{M}}(\theta) d\theta && (4.13)\end{aligned}$$

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.

$$
\begin{aligned}
d_k\left[\delta_y \middle\| \Pi_{Y|\mathcal{M}}\right] &= \|k(\cdot, y) - \mu_{Y|\mathcal{M}}\|^2_{\mathcal{H}} && (4.14)\\
&= \|k(\cdot, y) - \mathbb{E}_{\theta\sim p_{\theta|\mathcal{M}}}\mu_{Y|\theta}\|^2_{\mathcal{H}} && (4.15)\\
&= \mathbb{E}_{\theta\sim p_{\theta|\mathcal{M}}}\|k(\cdot, y) - \mu_{Y|\theta}\|^2_{\mathcal{H}} - \mathbb{E}_{\theta\sim p_{\theta|\mathcal{M}}}\|\mu_{X|\theta,\mathcal{M}} - \mu_{X|\mathcal{M}}\|^2_{\mathcal{H}} && (4.16)\\
&= \underbrace{\mathbb{E}_{\theta\sim p_{\theta|\mathcal{M}}} d_k\left[\delta_y \middle\| p_{X|\theta}\right]}_{\text{average accuracy}} - \underbrace{\mathbb{I}_k\left[X \leftarrow \theta\right]}_{\text{diversity}} && (4.17)
\end{aligned}
$$

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

4.2.2 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.

$$
R(y_{I_N}, \Pi) = -\frac{1}{|I_N|}\sum_{n\in I_N} \log P_{Y_n|Y_{I_n\setminus\{n\}}}(y_n) \tag{4.18}
$$

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.

4.2.3 Information quantities for stochastic 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*

$$\mathbb{H}_R[\Pi] = lim_{N\to\infty} R_{I_N}(X_{I_N}, \Pi), \text{ where } X \sim \Pi \tag{4.19}$$

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

$$d_R[\Pi\|\Xi] = lim_{N\to\infty} R_{I_N}(X_{I_N}, \Xi) - R_{I_N}(X_{I_N}, \Pi), \text{ where } X \sim \Pi \tag{4.20}$$

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.

Part II

Approximate Bayesian analysis

Chapter 5

Decision theoretic approximate inference

In practically interesting Bayesian models, the posterior distribution 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 distribution, often of exponential family, 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 allow 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.

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 **?**.

5.1 Loss-calibrated approximate inference

TODO: General paragraph about Bayesian inference

In many practically relevant cases computing the posterior is not analytically tractable. This is predominantly due to the fact that the integral defining the marginal likelihood $\int p(\mathcal{D}|\theta)p(\theta)d\theta$ cannot be computed analytically in closed form, and therefore the normalisation of the posterior

cannot be computed. But sometimes the complexity of evaluating even the unnormalised posterior increases exponentially with the amount of observed data, as in the case of for example switching state space models. In either case, it is usual practice to approximate the intractable posterior by something simpler, an approximate distribution q. The problem of finding an approximate posterior q is referred to as approximate inference.

Over the years, two dominant branches of approximate inference emerged. The first branch, that I will refer to as parametric approximation schemes, includes variational inference, Laplace approximation and expectation propagation. The common theme in these techniques is that the complicated posterior is replaced by an approximate distribution chosen from a particular parametric family of distributions, usually from an exponential family. These methods differ in their objective functions they minimise, which measure discrepancy between the target posterior distribution and the approximation q.

5.1.1 Overview of variational methods and expectation propagation

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

$$\log p(\mathcal{D}) = \log \int p(\mathcal{D}|\theta)p(\theta)d\theta \tag{5.1}$$

$$= \log \int \frac{p(\mathcal{D}|\theta)p(\theta)}{q(\theta)} q(\theta)d\theta \tag{5.2}$$

$$\geq \int \log \frac{p(\mathcal{D}|\theta)p(\theta)}{q(\theta)} q(\theta)d\theta \tag{5.3}$$

$$= \log p(\mathcal{D}) - d_{KL}\left[p_{\mathcal{D}} \| q\right] \tag{5.4}$$

by minimising the Kullback-Leibler divergence (Eqn. (**??**)) between the approximate distribution q and the true posterior $p_{\mathcal{D}}$. A common practice in approximate inference is to choose the approximate posterior distribution q 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. When these assumptions are made, the solution to the above optimisation can often be expressed in closed form, or efficient iterative algorithms exist for finding a locally optimal solution numerically.

The KL divergence is non-symmetric, therefore the order of arguments matter. Variational methods minimise $d_{KL}\left[q \| p_{\mathcal{D}}\right]$, that is with the approximate distribution being the first argument. On the one hand, this is highly convenient as computing the divergence in this direction requires integration only over q, which is assumed simpler than the real posterior $p_{\mathcal{D}}$.On the other hand, as I argued in section **??**, in the scoring rule interpretation suggests that the *right* way to use divergence is $d_{KL}\left[p_{\mathcal{D}} \| q\right]$, i. e. when its first argument is the true distribution we want to approximate, and the second argument some approximation q. This has been pointed out previously by Csató and Opper [2002], Minka [2001] and many other authors. This does not mean that variational inference does not work, it just means that by performing variational inference we loose some of the intuitive interpretation of KL divergence as Bregman divergence under the logarithmic loss.

Several approaches therefore tried to fix this conceptual issue, and minimise KL divergence in the opposite direction. This is unfortunately a challenge, as computing the divergence $d_S\left[p_{\mathcal{D}} \| q\right]$ always 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 $d_{KL}\left[p_{\mathcal{D}} \| q\right]$ as follows. EP assumes the posterior can be written as a product of factors as such:

$$p_{\mathcal{D}}(\theta) = \frac{1}{Z} \prod t_i(\theta) \tag{5.5}$$

The terms t_i are assumed simple, and in most cases depend only on a few components of the multivariate parameter vector θ. What makes the posterior intractable is the normalisation constant Z, computing which would involve a very expensive integral. Expectation propagation approximates the posterior by substituting approximate factors $\tilde{t}_i$ for original factors t_i, in such a way that the product of approximate factors

$$q(\theta) = \prod \tilde{t}_i(\theta) \tag{5.6}$$

is tractable. The approximate factors are improved one-by-one using the following objective function:

$$\tilde{t}_i^{new} = \underset{t \in \text{approximate family}}{\operatorname{argmin}} d_{KL}\left[\frac{1}{\int q(\theta)\frac{t_i(\theta)}{\tilde{t}_i(\theta)}d\theta} q(\theta)\frac{t_i(\theta)}{\tilde{t}_i(\theta)} \middle\| q(\theta)\frac{t_i(\theta)}{t(\theta)}\right] \tag{5.7}$$

Essentially, in each iteration the algorithm replaces one of the approximate factors in the approximate posterior q with the real factor to construct a one-step-closer-to-exact approximation to the posterior $\tilde{q}$. Then it uses this $\tilde{q}$ as the target distribution and computes a new approximation by minimising KL divergence. This step is repeated until convergence, that is until no approximate factors can be further improved by the KL divergence metric.

Thus in expectation-propagation, the KL divergence is used in the right direction that is well motivated by the theory of scoring rules and Bregman divergences. However, as computing KL divergence in the right direction is intractable it has to use a roundabout method.

5.1.2 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. Approximate inference is concerned with approximating the posterior, but often ignores the fact that the posterior is then used in a wider context to make optimal decisions. In this section I review the theory of Bayesian decisions, and then devise a framework for addressing questions that arise when using approximate inference in the context of optimal decision making.

The ingredients of Bayesian decision theory are (see Ch. 2 of **?** or Ch. 1 of **?** for example):

- a loss $\ell(\theta, a)$ which quantifies 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* **?**)

$$\mathcal{R}_{p_\mathcal{D}}(a) \doteq \int_\Theta \ell(\theta, a)\, p(\theta|\mathcal{D}) d\theta \tag{5.8}$$

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 consecutive steps of 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. Crucially, the first step is independent of losses, the posterior can be computed irrespective of how the loss ℓ is defined. In fact, once we have computed the posterior, the same distribution can be used to solve different decision problems with different losses involved. This independent breakdown of computation is what makes the posterior distribution such an important object in Bayesian statistics.

But when the posterior is intractable to compute and approximations are needed - as it is the case most of the time - additional questions arise. Is this two-step breakdown of computations to inference and then risk minimisation still a sensible thing to do? How should we decide what approximate inference method to use? Can we still re-use the same approximate posterior with different loss functions just as we can if no approximations are needed. Is the choice of approximate inference technique independent of the loss function? This chapter introduces loss-calibrated approximate Bayesian inference is a theoretical framework for addressing these questions.

To illustrate the role and behaviour ot approximate inference in a Bayesian decision problem consider the following simple problem. Suppose that we control a nuclear power-plant which has an unknown temperature θ that we model with Bayesian inference based on some measurements $\mathcal{D}$. The plant is in danger of over-heating, and as the operator, we can take two actions: either shut it down or keep it running. Keeping it running while the temperature is above a critical threshold T_{crit} will cause a nuclear meltdown, incurring a large loss $L(\theta > T_{\text{crit}}, \text{'on'})$. On the other hand, shutting down the power plant incurs a moderate loss $L(\text{'off'})$, irrespective of the temperature. Suppose that our current observations yielded a complicated multi-modal posterior $p_{\mathcal{D}}(\theta)$ (**??**, solid curve) that we do not have comutational resources to represent. Thus we chose to approximate it with a simple Gaussian distribution.

Now consider how various approaches to approximate inference would perform in terms of their Bayesian posterior risk. Minimizing $d_{KL}\left[q \| p_{\mathcal{D}}\right]$, as in variational inference, yields candidate q_1 which concentrates around the largest mode, ignoring entirely the second small mode around the critical temperature **??**, dotted curve). Minimizing $d_{KL}\left[p_{\mathcal{D}} \| q\right]$ gives a more global approximation: q_2 matches moments of the posterior, but still underestimates the probability of the temperature being above T_{crit}, thereby leading to a suboptimal decision **??**, dashed curve).

TODO: Rewrite as divergences have not been definded yet q_3 is one of the minimizers of $d_L(p_{\mathcal{D}} \| q)$ in this setting, resulting in the same decision as $p_{\mathcal{D}}$ **??**, dash-dotted curve). Note that q_3 does not model all aspects of the posterior, but it estimates the Bayes-decision well. Because there are only two possible actions in this setup, the set $\mathcal{Q}$ is split in only two halves by the function $d_L(p_{\mathcal{D}}, q)$ and so there are infinitely many q_{opt}'s that are equivalent in terms of their risk. In contrast, in the predictive setting of section **??** where in addition we assume $\mathcal{X}$ and $p(x)$ to be continuous, we could obtain a finer resolution $d_L(p_{\mathcal{D}} \| q)$ which can potentially yield a unique optimizer.

This simple example already highlighted some features of the loss-calibrated framework. First of all, it is clear, that even in a simple example the choice of approximate inference methods matters, and has a great influence on risks and the final decisions made. In this case minimising $d_{KL}\left[p_{\mathcal{D}} \| q\right]$ yielded a solution superior to minimising the variational criterion $d_{KL}\left[q \| p_{\mathcal{D}}\right]$, but we could just as well construct another example, where it is the other way around. Even though $d_{KL}\left[p_{\mathcal{D}} \| q\right]$ is thought of as the more principled method, in the context of this decision problem neither of them is clearly better or more principled than the other.

5.1.3 The loss-calibrated approximate inference framework

In practice, one usually treats the approximate q as if it was the true posterior and chooses the action that minimizes what we will call the *q-risk*:

$$\mathcal{R}_q(h) \doteq \int_{\Theta} q(\theta) L(\theta, h) d\theta, \tag{5.9}$$

obtaining a *q-optimal* action h_q:

$$h_q \doteq \operatorname{argmin} h \in \mathcal{H} \mathcal{R}_q(h). \tag{5.10}$$

In this paper, we will assume that computing exactly the q-optimal action h_q for $q \in \mathcal{Q}$ is tractable, and focus on the problem of choosing a suitable q to approximate the posterior $p_{\mathcal{D}}$ *in order to yield a decision* h_q *with low posterior risk* $\mathcal{R}_{p_{\mathcal{D}}}(h_q)$, mimicking the standard methodology but

crystallizing the decision theoretic goal. Given this approach, a (usually non-unique) optimal $q \in \mathcal{Q}$ is clearly:

$$q_{\text{opt}} = \text{argmin}\, q \in \mathcal{Q}\mathcal{R}_{p_{\mathcal{D}}}(h_q), \tag{5.11}$$

though a practical algorithm might only be able to find an approximate minimizer to this quantity. In the case where $p_{\mathcal{D}} \in \mathcal{Q}$, $p_{\mathcal{D}}$ is obviously optimal according to this criterion.

We could interpret the above criterion as minimizing the following asymmetric non-negative discrepancy measure between distributions:

$$d_L(p\|q) \doteq \mathcal{R}_p(h_q) - \mathcal{R}_p(h_p). \tag{5.12}$$

Interestingly, the Kullback-Leibler divergence $KL(p\|q)$ can be interpreted as a special case of d_L for the task of posterior density estimation over Θ. In this task, an action h is a density over Θ and the standard density estimation statistical loss is $L(\theta, h) = -\log h(\theta)$. The q-risk $R_q(h)$ then becomes the cross-entropy $H(q, h) = -\int_\Theta q(\theta)\log(h(\theta))d\theta$, and so $h_q = q$ assuming that $q \in \mathcal{H}$. Under these assumptions, we obtain that $KL(p\|q) = d_L(p\|q)$ and so as was already known in statistics, $KL(p_{\mathcal{D}}\|\cdot)$ appears "loss-calibrated" for the task of posterior density estimation in our approximation framework. But this begs the natural question of whether minimizing d_L for a particular loss L provides optimal performance under other losses. We will show in **??** that even in the simple Gaussian linear regression setting, minimizing the KL divergence can be suboptimal in the squared loss sense, thus motivating us to seek loss-calibrated alternatives.

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.

Chapter 6

Quasi Monte Carlo and herding

A popular alternative to parametric approximation schemes, such as variational inference and expectation propagation are Monte Carlo methods.

Monte Carlo methods produce random samples from the posterior distribution $p_{\mathcal{D}}$ and then approximate relevant integrals by taking the empirical means over these samples. Subject to smoothness conditions, this non-deterministic estimate of any integral converges at a rate $\mathcal{O}(\frac{1}{\sqrt{N}})$, where N is the number of samples. This convergence is guaranteed by the law of large numbers. An appealing property of Monte Carlo methods is that in theory an arbitrarily precise estimate can be obtained by just increasing the number of samples. In this sense, Monte Carlo approximation is non-parametric: the number of parameters that describe the approximate distribution is not fixed ahead of time, and can be arbitrarily large.

When exact sampling from $p_{\mathcal{D}}$ is impossible or impractical, Markov chain Monte Carlo (MCMC) methods are often used. MCMC methods only require knowing the target distribution up to a constant factor. Practically this means that even if the normalisation constant of the posterior is intractable, MCMC techniques can still be used to generate samples from it.

Various variants of MCMC methods can be applied to almost any problem but the convergence rate of the estimate depends on several factors and is hard to estimate [**?**]. Typically, MCMC techniques introduce positive correlation between subsequent samples, and thus are less effective than exact Monte Carlo sampling. For an overview of various Monte Carlo techiques, see [Murray, 2007].

Monte Carlo methods are very general, they guarantee convergence for any measurable integral. Hence, convergence is also guaranteed in the KL divergence sense, and as the posterior risk is expressed as an integral, they also ensure convergence in $d_\ell(\cdot\|\cdot)$ for any loss function ℓ. However, the rate of convergence cannot be fine-tuned to a particular divergence measure. One might hope that if the loss function ℓ is known ahead of time, a faster convergence rate can be achieved, maybe at the cost of slowing down convergence on integrals that are irrelevant to the decision problem.

revisit toy example Let us consider the power plant example from the previous section. To be able to make a decision, the only thing we need to know is the probability of the temperature exceeding the critical temperature. Thus, when the distribution is approximated via Monte Carlo, the only summary statistic we care about is the fraction of samples that are above the critical temperature.

The probability of interest can be written as the expectation of the indicator function that takes value 1 if the temperature exceeds the critical one and 0 otherwise. This indicator function is measurable, therefore an $\mathcal{O}(\frac{1}{\sqrt{N}})$ convergence is guaranteed by exact MCMC sampling. However, it is easy to construct an ideal series of N 'pseudo-samples' where the error is upper bounded by $\frac{1}{N}$. (The problem is equivalent to approximating the probability with a series of rational numbers). This ideal set of N pseudosamples may of course be a terrible general approximation to the full probability distribution $p_{\mathcal{D}}$, but from the perspective of the decision problem it converges much

faster than the random Monte Carlo samples.

TODO: illustrate this on figures: Fig 1: same as in previous section. Fig 2: approximating the probability with random MCMC and with optimal QMC

Quasi monte Carlo approaches The focus of this chapter are quasi-Monte Carlo methods that – instead of sampling randomly – produce a set of pseudo-samples in a deterministic fashion. These methods operate by directly minimising some sort of discrepancy between the empirical distribution of pseudo-samples and the target distribution. Whenever these methods are applicable, they achieve convergence rates superior to the $\mathcal{O}(\frac{1}{\sqrt{N}})$ rate typical of random sampling.

TODOreview existing quasi-Monte-Carlo methods: Sobol sequences, Halton sequence

The quasi-Monte-Carlo methods reviewed here often achieve faster convergence rates than traditional random Monte Carlo, but they are general-purpose sampling tools: they cannot be fine-tuned to particular decision problems we may want to use them for. Here I will introduce a class of quasi-Monte-Carlo methods that I will call loss-calibrated QMC.

Quasi-Monte-Carlo can be interpreted as a special case of approximate inference, where the approximating family is the family of empirical distributions

$$q(x; x_1, \ldots, x_N) = \frac{1}{N} \sum_{n=1}^{N} \delta(x - x_n), \tag{6.1}$$

or weighted empirical distributions

$$q(x; x_1, \ldots, x_N, w_1, \ldots, w_N) = \sum_{n=1}^{N} w_n \delta(x - x_n). \tag{6.2}$$

Finding the optimal loss-calibrated sample set can then be achieved by minimising the loss-calibrated divergence d_ℓ between the target distribution $p_\mathcal{D}$ and the approximation q:

$$\{x_1, \ldots, x_N\}_{n=1}^{N} = \underset{\{x_1,\ldots,x_N\}_{n=1}^{N}}{\operatorname{argmin}} \; d_\ell \left[p_\mathcal{D} \| q(x; x_1, \ldots, x_N)\right] \tag{6.3}$$

It is important to note, that the above procedure does not make sense for general Bregman divergences. For example, the KL divergence $d_{KL}\left[p\|q\right]$ requires the approximate distribution q to be absolutely continuous with respect to the target distribution p, which, unless the target distribution is also discrete, cannot be satisfied if q is atomic.

The minimisation in Equation (6.3) is

Myopic sequential loss-calibrated Quasi Monte Carlo In most cases - just as loss-calibrated approximate inference in general, algorithmic implementations of loss-calibrated QMC requires the ability to evaluate certain integrals over the target distribution easily, therefore practical applications of loss-calibrated QMC in the form presented here are limited. Nevertheles, the framework may provide useful blueprint for designing sampling algotithms that are more tailored to particular decision scenarios.

INTRODUCTION

The problem: Integrals A common problem in statistical machine learning is to compute expectations of functions over probability distributions of the form:

$$Z_{f,p} = \int f(x) p(x) dx \tag{6.4}$$

Examples include computing marginal distributions, making predictions marginalizing over parameters, or computing the Bayes risk in a decision problem. In this paper we assume that the distribution $p(x)$ is known in analytic form, and $f(x)$ can be evaluated at arbitrary locations.

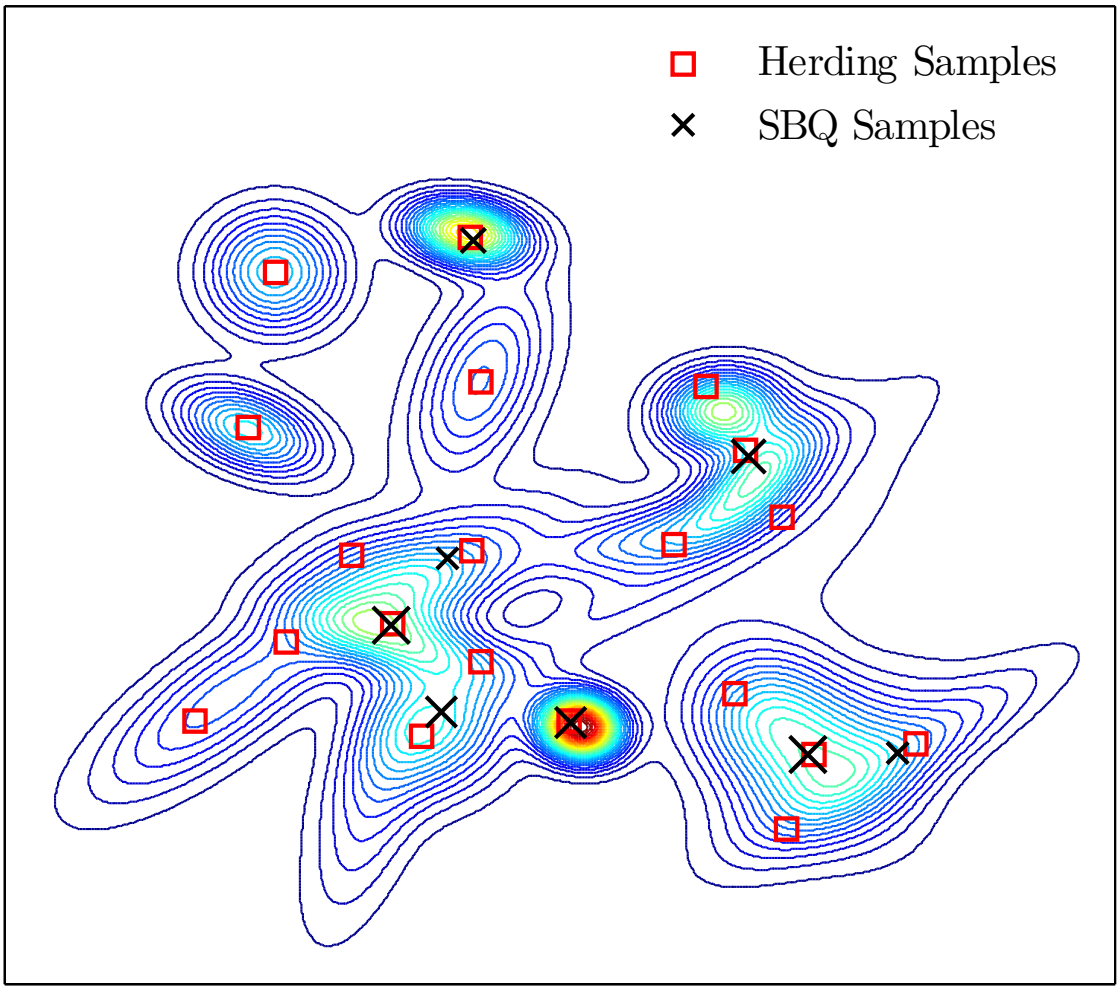

Figure 6.1: The first 8 samples from sequential Bayesian quadrature, versus the first 20 samples from herding. Only 8 weighted SBQ samples are needed to give an estimator with the same maximum mean discrepancy as using 20 herding samples with uniform weights. Relative sizes of samples indicate their relative weights.

Monte Carlo methods produce random samples from the distribution p and then approximate the integral by taking the empirical mean $\hat{Z} = \frac{1}{N}\sum_{n=1}^{N} f_{x_n}$ of the function evaluated at those points. This non-deterministic estimate converges at a rate $\mathcal{O}(\frac{1}{\sqrt{N}})$. When exact sampling from p is impossible or impractical, Markov chain Monte Carlo (MCMC) methods are often used. MCMC methods can be applied to almost any problem but convergence of the estimate depends on several factors and is hard to estimate [**?**]. The focus of this paper is on quasi-Monte Carlo methods that – instead of sampling randomly – produce a set of pseudo-samples in a deterministic fashion. These methods operate by directly minimising some sort of discrepancy between the empirical distribution of pseudo-samples and the target distribution. Whenever these methods are applicable, they achieve convergence rates superior to the $\mathcal{O}(\frac{1}{\sqrt{N}})$ rate typical of random sampling.

In this paper we highlight and explore the connections between two deterministic sampling and integration methods: Bayesian quadrature (BQ) [**??**] (also known as Bayesian Monte Carlo) and kernel herding [**?**]. Bayesian quadrature estimates integral (6.4) by inferring a posterior distribution over f conditioned on the observed evaluations f_{x_n}, and then computing the posterior

expectation of $Z_{f,p}$. The points where the function should be evaluated can be found via Bayesian experimental design, providing a deterministic procedure for selecting sample locations.

Herding, propositionosed recently by **?**, produces pseudosamples by minimising the discrepancy of moments between the sample set and the target distribution. Similarly to traditional Monte Carlo, an estimate is formed by taking the empirical mean over samples $\hat{Z} = \frac{1}{N}\sum_{n=1}^{N} f_{x_n}$. Under certain assumptions, herding has provably fast, $\mathcal{O}(\frac{1}{N})$ convergence rates in the parametric case, and has demonstrated strong empirical performance in a variety of tasks.

Summary of contributions In this paper, we make two main contributions. First, we show that the Maximum Mean Discrepancy (MMD) criterion used to choose samples in kernel herding is identical to the expected error in the estimate of the integral $Z_{f,p}$ under a Gaussian process prior for f. This expected error is the criterion being minimized when choosing samples for Bayesian quadrature. Because Bayesian quadrature assigns different weights to each of the observed function values $f(x)$, we can view Bayesian quadrature as a weighted version of kernel herding. We show that these weights are optimal in a minimax sense over all functions in the Hilbert space defined by our kernel. This implies that Bayesian quadrature dominates uniformly-weighted kernel herding and other non-optimally weighted herding in rate of convergence.

Second, we show that minimising the MMD, when using BQ weights is closely related to the sparse dictionary selection problem studied in [**?**], and therefore is approximately submodular with respect to the samples chosen. This allows us to reason about the performance of greedy forward selection algorithms for Bayesian Quadrature. We call this greedy method Sequential Bayesian Quadrature (SBQ).

We then demonstrate empirically the relative performance of herding, i.i.d random sampling, and SBQ, and demonstrate that SBQ attains a rate of convergence faster than $\mathcal{O}(1/N)$.

HERDING

Herding was introduced by **?** as a method for generating pseudo-samples from a distribution in such a way that certain nonlinear moments of the sample set closely match those of the target distribution. The empirical mean $\frac{1}{N}\sum_{n=1}^{N} f_{x_n}$ over these pseudosamples is then used to estimate integral 6.4.

Maximum Mean Discrepancy

For selecting pseudosamples, herding relies on an objective based on the maximum mean discrepancy [MMD; **?**]. MMD measures the divergence between two distributions, p and q with respect to a class of integrand functions $\mathcal{F}$ as follows:

$$\div\mathcal{F}pq = \sup_{f\in\mathcal{F}} \left| \int f(x)p(x)dx - \int f(x)q(x)dx \right| \tag{6.5}$$

Intuitively, if two distributions are close in the MMD sense, then no matter which function f we choose from $\mathcal{F}$, the difference in its integral over p or q should be small. 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 [**?**]:

$$MMD^2_{\mathcal{H}}(p, q) = \sup_{\substack{f \in \mathcal{H} \\ \|f\|_{\mathcal{H}}=1}} \left| \int f_x p(x) dx - \int f_x q(x) dx \right|^2 \tag{6.6}$$

$$= \|\mu_p - \mu_q\|^2_{\mathcal{H}} \tag{6.7}$$

$$= \iint k(x, y) p(x) p(y) dx dy$$

$$-2 \iint k(x, y) p(x) q(y) dx dy$$

$$+ \iint k(x, y) q(x) q(y) dx dy, \tag{6.8}$$

where in the above formula $\mu_p = \int \phi(x) p(x) dx \in \mathcal{H}$ denotes the *mean element* associated with the distribution p. For characteristic kernels, such as the Gaussian kernel, the mapping between a distribution and its mean element is bijective. As a consequence $\text{MMD}_{\mathcal{H}}(p, q) = 0$ if and only if $p = q$, making it a powerful measure of divergence.

Herding uses maximum mean discrepancy to evaluate of how well the sample set $\{x_1, \ldots, x_N\}$ represents the target distribution p:

$$\epsilon_{herding}\left(\{x_1, \ldots, x_N\}\right) = \text{MMD}_{\mathcal{H}}\left(p, \frac{1}{N} \sum_{n=1}^{N} \delta_{x_n}\right) \tag{6.9}$$

$$= \iint k(x, y) p(x) p(y) dx dy$$

$$-2 \frac{1}{N} \sum_{n=1}^{N} \int k(x, x_n) p(x) dx + \frac{1}{N^2} \sum_{n,m=1}^{N} k(x_n, x_m) \tag{6.10}$$

The herding procedure greedily minimizes its objective $\epsilon_{herding}\left(\{x_1, \ldots, x_N\}\right)$, adding pseudosamples x_n one at a time. When selecting the $n+1$-st pseudosample:

$$x_{n+1} \leftarrow \operatorname*{argmin}_{x \in \mathcal{X}} \epsilon_{herding}\left(\{x_1, \ldots, x_n, x\}\right) \tag{6.11}$$

$$= \operatorname*{argmax}_{x \in \mathcal{X}} 2\mathbb{E}_{x' \sim p} k(x, x') - \frac{1}{n+1} \sum_{m=1}^{n} k(x, x_m),$$

assuming $k(x, x) = \text{const}$. The formula (6.11) admits an intuitive interpretation: the first term encourages sampling in areas with high mass under the target distribution $p(x)$. The second term discourages sampling at points close to existing samples.

Evaluating (6.11) requires us to compute $\mathbb{E}_{x' \sim p} k(x, x')$, that is to integrate the kernel against the target distribution. Throughout the paper we will assume that these integrals can be computed in closed form. Whilst the integration can indeed be carried out analytically in several cases [Song et al., 2008, **?**], this requirement is the most pertinent limitation on applications of kernel herding, Bayesian quadrature and related algorithms.

Complexity and Convergence Rates

Criterion (6.11) can be evaluated in only $\mathcal{O}(n)$ time. Adding these up for all subsequent samples, and assuming that optimisation in each step has $\mathcal{O}(1)$ complexity, producing N pseudosamples via kernel herding costs $\mathcal{O}(N^2)$ operations in total.

In finite dimensional Hilbert spaces, the herding algorithm has been shown to reduce MMD at a rate $\mathcal{O}(\frac{1}{N})$, which compares favourably with the $\mathcal{O}(\frac{1}{\sqrt{N}})$ rate obtained by non-deterministic Monte Carlo samplers. However, as pointed out by **?**, this fast convergence is not guaranteed in infinite dimensional Hilbert spaces, such as the RKHS corresponding to the Gaussian kernel.

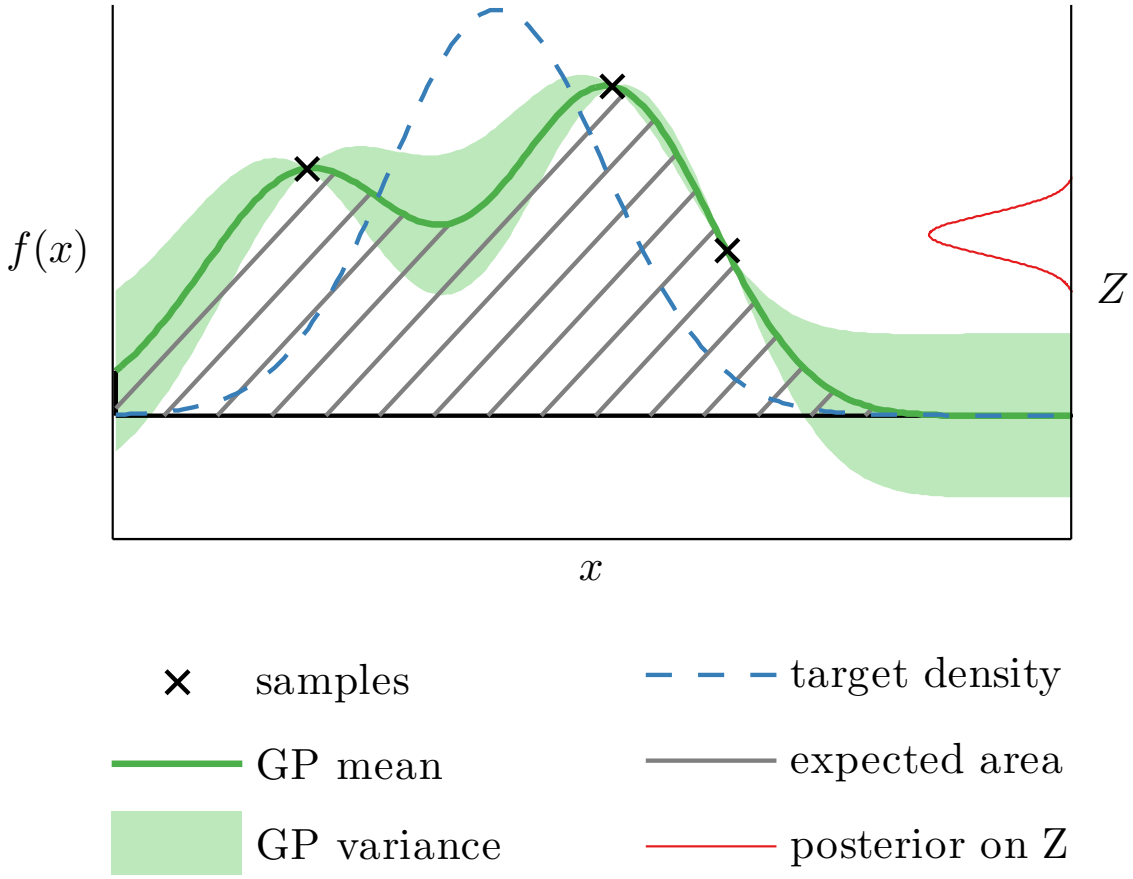

Figure 6.2: An illustration of Bayesian Quadrature. The function $f(x)$ is sampled at a set of input locations. This induces a Gaussian process posterior distribution on f, which is integrated in closed form against the target density, $p(x)$. Since the amount of volume under f is uncertain, this gives rise to a (Gaussian) posterior distribution over $Z_{f,p}$.

BAYESIAN QUADRATURE

So far, we have only considered integration methods in which the integral (6.4) is approximated by the empirical mean of the function evaluated at some set of samples, or pseudo-samples. Equivalently, we can say that Monte Carlo and herding both assign an equal $\frac{1}{N}$ weight to each of the samples.

In [**?**], an alternate method is propositionosed: Bayesian Monte Carlo, or Bayesian quadrature (BQ). BQ puts a prior distribution on f, then estimates integral (6.4) by inferring a posterior distribution over the function f, conditioned on the observations $f(x_n)$ at some query points x_n. The posterior distribution over f then implies a distribution over $Z_{f,p}$. This method allows us to choose sample locations x_n in any desired manner. See Figure 6.2 for an illustration of Bayesian Quadrature.

BQ Estimator

Here we derive the BQ estimate of (6.4), after conditioning on function evaluations $f(x_1) \dots f(x_N)$, denoted as $f(X)$. The Bayesian solution implies a distribution over $Z_{f,p}$. The mean of this distribution, $\mathbb{E}Z$ is the optimal Bayesian estimator for a squared loss.

For simplicity, f is assigned a Gaussian process prior with kernel function k and mean 0. This assumption is very similar to the one made by kernel herding in Eqn. (6.10).

After conditioning on f_x, we obtain a closed-form posterior over f:

$$p(f(x_\star)|f(X)) = \mathcal{N} f_{x_\star} \bar{f}(x_\star) \mathbb{C}\mathrm{ov}_([x, \star] , x'_\star) \tag{6.12}$$

where

$$\bar{f}(x_\star) = k(x_\star, X) K^{-1} f(X) \tag{6.13}$$

$$\mathbb{C}\mathrm{ov}_([x, \star] , x'_\star) = k(x_\star, x_\star) - k(x_\star, X) K^{-1} k(X, x_\star) \tag{6.14}$$

and $K = k(X, X)$. Conveniently, the GP posterior allows us to compute the expectation of (6.4) in closed form:

$$\mathbb{E}_{\mathrm{GP}} Z = \mathbb{E}_{\mathrm{GP}} \int f(x) p(x) dx \tag{6.15}$$

$$= \iint f(x) p(f(x)|f(X)) p(x) dx df \tag{6.16}$$

$$= \int \bar{f}(x) p(x) dx \tag{6.17}$$

$$= \left[\int k(x, X) p(x) dx \right] K^{-1} f(X) \tag{6.18}$$

$$= \mathbf{z}^T K^{-1} f(X) \tag{6.19}$$

where

$$z_n = \int k(x, x_n) p(x) dx = \mathbb{E}_{x' \sim p} k(x_n, x'). \tag{6.20}$$

Conveniently, as in kernel herding, the desired expectation of $Z_{f,p}$ is simply a linear combination of observed function values $f(x)$:

$$\mathbb{E}_{\mathrm{GP}} Z = \mathbf{z}^T K^{-1} f(X) \tag{6.21}$$

$$= \sum_n w_{\mathrm{BQ}}^{(n)} f_{x_n}) \tag{6.22}$$

where

$$w_{\mathrm{BQ}}^{(n)} = \sum_m \mathbf{z}_j^T K_{nm}^{-1} \tag{6.23}$$

Thus, we can view the BQ estimate as a weighted version of the herding estimate. Interestingly, the weights w_{BQ} do not need to sum to 1, and are not even necessarily positive.

Non-normalized and Negative Weights

When weighting samples, it is often assumed, or enforced [as in ?Song et al., 2008], that the weights w form a probability distribution. However, there is no technical reason for this requirement, and in fact, the optimal weights do not have this propositionerty. Figure 6.3 shows a representative set of 100 BQ weights chosen on samples representing the distribution in figure 6.1. There are several negative weights, and the sum of all weights is 0.93.

Figure 6.4 demonstrates that, in general, the sum of the Bayesian weights exhibits shrinkage when the number of samples is small.

Optimal sampling for BQ

Bayesian quadrature provides not only a mean estimate of $Z_{f,p}$, but a full Gaussian posterior distribution. The variance of this distribution $\mathbb{V} Z_{f,p} | f_{x_1}, \ldots, f_{x_N}$ quantifies our uncertainty in the

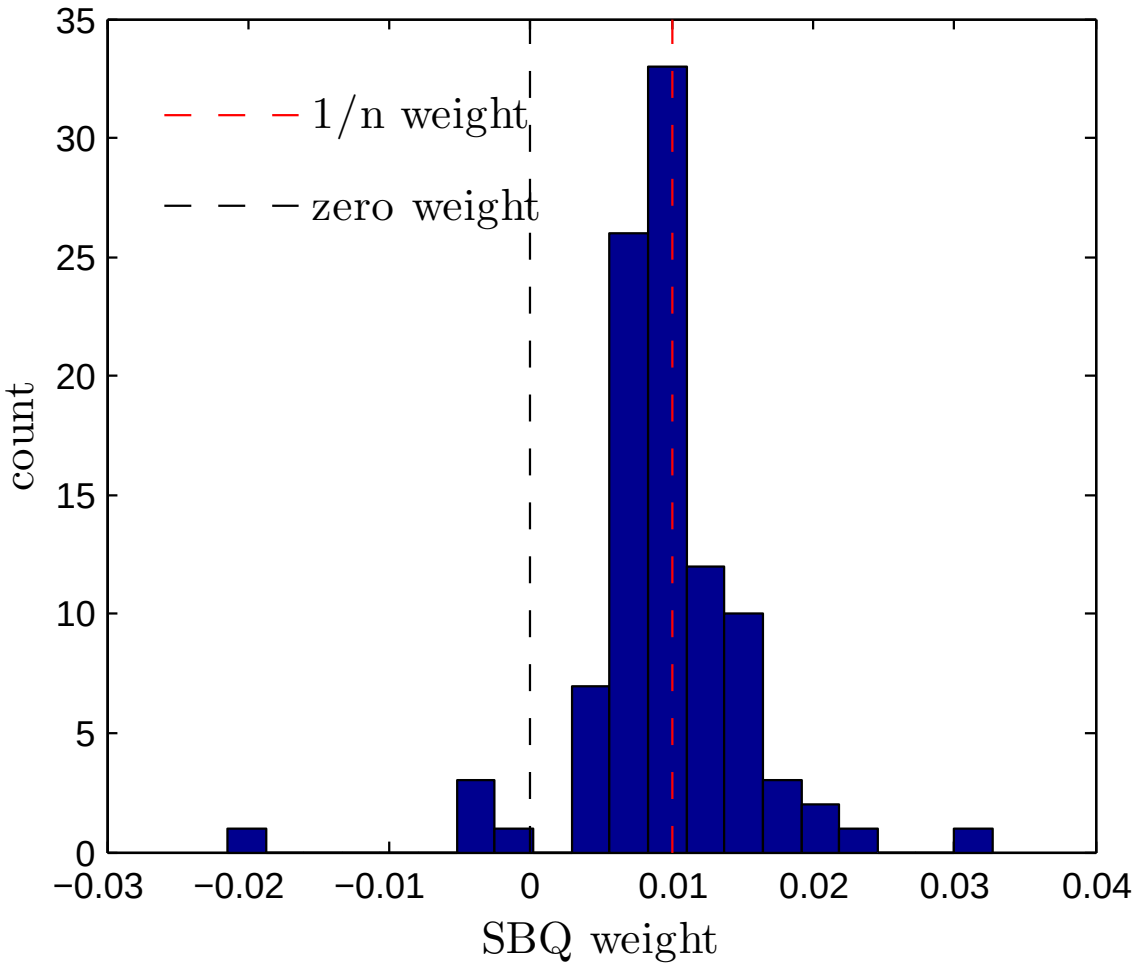

Figure 6.3: A set of optimal weights given by BQ, after 100 SBQ samples were selected on the distribution shown in Figure 6.1. Note that the optimal weights are spread away from the uniform weight ($\frac{1}{N}$), and that some weights are even negative. The sum of these weights is 0.93.

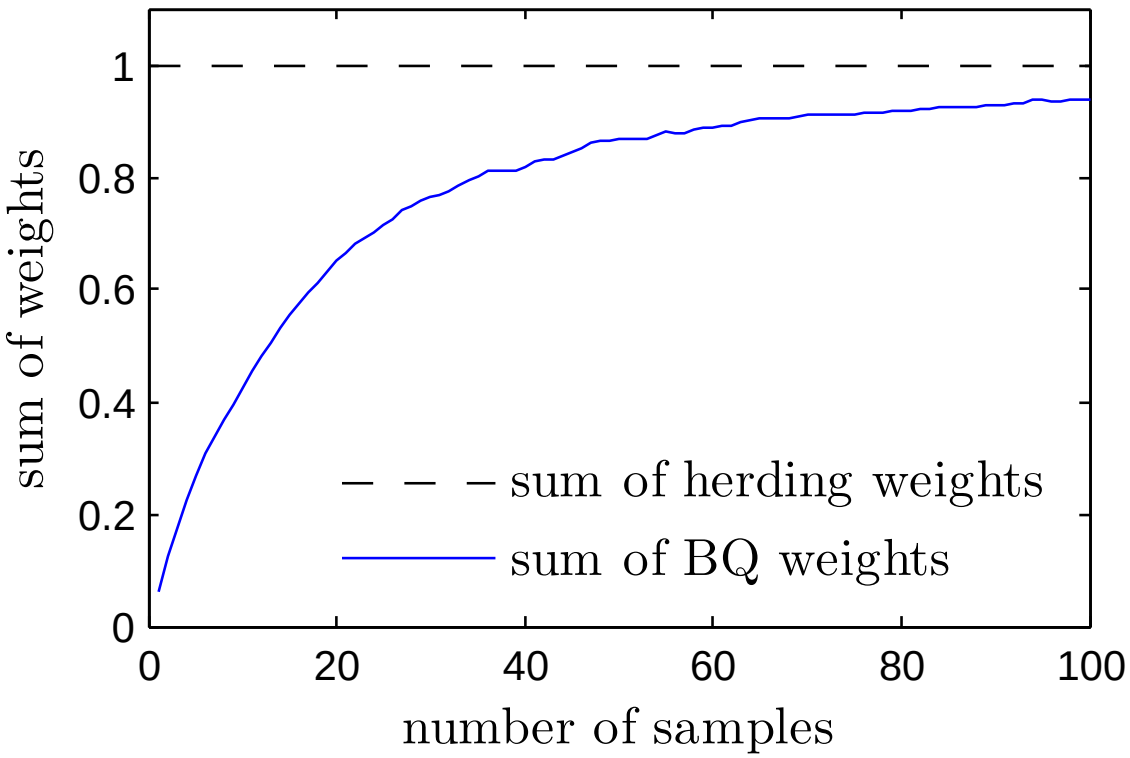

Figure 6.4: An example of Bayesian shrinkage in the sample weights. In this example, the kernel width is approximately $1/20$ the width of the distribution being considered. Because the prior over functions is zero mean, in the small sample case the weights are shrunk towards zero. The weights given by simple Monte Carlo and herding do not exhibit shrinkage.

estimate. When selecting locations to evaluate the function f, minimising the posterior variance is a sensible strategy. Below, we give a closed form formula for the posterior variance of $Z_{f,p}$, conditioned on the observations $f_{x_1} \dots f_{x_N}$, which we will denote by ϵ^2_{BQ}. For a longer derivation, see **?**.

$$\epsilon^2_{\text{BQ}}(x_1, \dots, x_N) = \mathbb{V} Z_{f,p} | f_{x_1}, \dots, f_{x_N} \tag{6.24}$$

$$= \mathbb{E}_{x,x' \sim p} k(x, x') - \mathbf{z}^T K^{-1} \mathbf{z}, \tag{6.25}$$

where $\mathbf{z}_n = \mathbb{E}_{x' \sim p} k(x_n, x')$ as before. Perhaps surprisingly, the posterior variance of $Z_{f,p}$ does not depend on the observed function values, only on the location x_n of samples. A similar independence is observed in other optimal experimental design problems involving Gaussian processes [**?**]. This allows the optimal samples to be computed ahead of time, before observing any values of f at all [**?**].

We can contrast the BQ objective ϵ^2_{BQ} in (6.25) to the objective being minimized in herding, $\epsilon^2_{herding}$ of equation (6.10). Just like $\epsilon^2_{herding}$, ϵ^2_{BQ} expresses a trade-off between accuracy and diversity of samples. On the one hand, as samples get close to high density regions under p, the values in $\mathbf{z}$ increase, which results in decreasing variance. On the other hand, as samples get closer to each other, eigenvalues of K increase, resulting in an increase in variance.

In a similar fashion to herding, we may use a greedy method to minimise ϵ^2_{BQ}, adding one sample at a time. We will call this algorithm *Sequential Bayesian Quadrature* (SBQ):

$$x_{n+1} \leftarrow \operatorname*{argmin}_{x \in \mathcal{X}} \epsilon_{\text{BQ}} \left(\{x_1, \dots, x_n, x\} \right) \tag{6.26}$$

Using incremental updates to the Cholesky factor, the criterion can be evaluated in $\mathcal{O}(n^2)$ time. Iteratively selecting N samples thus takes $\mathcal{O}(N^3)$ time, assuming optimisation can be done on $\mathcal{O}(1)$ time.

RELATING $\mathbb{V} Z_{f,p}$ TO mmd

The similarity in the behaviour of $\epsilon^2_{herding}$ and ϵ^2_{BQ} is not a coincidence, the two quantities are closely related to each other, and to MMD.

Proposition 2. *The expected variance in the Bayesian quadrature ϵ^2_{BQ} is the maximum mean discrepancy between the target distribution p and $q_{\text{BQ}}(x) = \sum_{n=1}^N w^{(n)}_{\text{BQ}} \delta_{x_n}(x)$*

Proof. The proof involves invoking the representer theorem, using bilinearity of scalar products and the fact that if f is a standard Gaussian process then $\forall g \in \mathcal{H} : \langle f, g \rangle \sim \mathcal{N}(0, \|g\|_{\mathcal{H}})$:

$$\mathbb{V} Z_{f,p} | f_{x_1}, \dots, f_{x_N} = \tag{6.27}$$

$$= \mathbb{E}_{f \sim GP} \left(\int f(x) p(x) dx - \sum_{n=1}^N w^{(n)}_{\text{BQ}} f(x_n) \right)^2 \tag{6.28}$$

$$= \mathbb{E}_{f \sim GP} \left(\int \langle f, \phi(x) \rangle \, p(x) dx - \sum_{n=1}^N w^{(n)}_{\text{BQ}} \langle f, \phi(x_n) \rangle \right)^2 \tag{6.29}$$

$$= \mathbb{E}_{f \sim GP} \left\langle f, \int \phi(x) p(x) dx - \sum_{n=1}^N w^{(n)}_{\text{BQ}} \phi(x_n) \right\rangle^2 \tag{6.30}$$

$$= \| \mu_p - \mu_{q_{\text{BQ}}} \|^2_{\mathcal{H}} \tag{6.31}$$

$$= \text{MMD}^2(p, q_{\text{BQ}}) \tag{6.32}$$

□

We know that the the posterior mean $\mathbb{E}_{\mathrm{GP}} Z_{f,p} | f_1, \ldots, f_N$ is a Bayes estimator and has therefore the minimal expected squared error amongst all estimators. This allows us to further rewrite ϵ^2_{BQ} into the following minimax forms:

$$\epsilon^2_{\mathrm{BQ}} = \sup_{\substack{f \in \mathcal{H} \\ \|f\|_{\mathcal{H}} \mathcal{H} = 1}} \left| \int f_x p(x) dx - \sum_{n=1}^{N} w^{(n)}_{\mathrm{BQ}} f_{x_n} \right|^2 \tag{6.33}$$

$$= \inf_{\hat{Z}: \mathcal{X}^N \mapsto \mathbb{R}} \sup_{\substack{f \in \mathcal{H} \\ \|f\|_{\mathcal{H}} = 1}} \left| Z - \hat{Z}\left(f_{x_1}, \ldots, f_{x_N}\right) \right|^2 \tag{6.34}$$

$$= \inf_{\boldsymbol{w} \in \mathbb{R}^N} \sup_{\substack{f \in \mathcal{H} \\ \|f\|_{\mathcal{H}} = 1}} \left| \int f_x p(x) dx - \sum_{n=1}^{N} w_n f_{x_n} \right|^2 \tag{6.35}$$

Looking at ϵ^2_{BQ} this way, we may discover the deep similarity to the criterion $\epsilon^2_{herding}$ that kernel herding minimises. Optimal sampling for Bayesian quadrature minimises the same objective as kernel herding, but with the uniform $\frac{1}{N}$ weights replaced by the optimal weights. As a corollary

$$\epsilon^2_{\mathrm{BQ}}(x_1, \ldots, x_N) \leq \epsilon^2_{KH}(x_1, \ldots, x_N) \tag{6.36}$$

It is interesting that ϵ^2_{BQ} has both a Bayesian interpretation as posterior variance under a Gaussian process prior, and a frequentist interpretation as a minimax bound on estimation error with respect to an RKHS.

SUBMODULARITY

In this section, we use the concept of approximate submodularity [**?**], in order to study convergence propositionerties of SBQ.

A set function $s : 2^{\mathcal{X}} \mapsto \mathbb{R}$ is *submodular* if, for all $A \subseteq B \subseteq \mathcal{X}$ and $\forall x \in \mathcal{X}$

$$s(A \cup \{x\}) - s(A) \geq s(B \cup \{x\}) - s(B) \tag{6.37}$$

Intuitively, submodularity is a diminishing returns propositionerty: adding an element to a smaller set has larger relative effect than adding it to a larger set. A key result [see e. g. **?**, and references therein] is that greedily maximising a submodular function is guaranteed not to differ from the optimal strategy by more than a constant factor of $(1 - \frac{1}{e})$.

Herding and SBQ are examples of greedy algorithms optimising set functions: they add each pseudosample in such a way as to minimize the instantaneous reduction in MMD. So it is intuitive to check whether the objective functions these methods minimise are submodular. Unfortunately, neither $\epsilon_{herding}$, not ϵ_{BQ} satisfies all conditions for submodularity. However, noting that SBQ is identical to the sparse dictionary selection problem studied in detail by **?**, we can conclude that SBQ satisfies a weaker condition called *approximate submodularity*.

A set function $s : 2^{\mathcal{X}} \mapsto \mathbb{R}$ is *approximately submodular* with constant $\epsilon > 0$, if for all $A \subseteq B \subseteq \mathcal{X}$ and $\forall x \in \mathcal{X}$

$$s(A \cup \{x\}) - s(A) \geq s(B \cup \{x\}) - s(B) - \epsilon \tag{6.38}$$

Proposition 3. *$\epsilon^2_{\mathrm{BQ}}(\emptyset) - \epsilon^2_{\mathrm{BQ}}(\cdot)$ is weakly a weakly submodular set function with constant $\epsilon < 4r$, where r is the incoherency*

$$r = \max_{x, x' \in \mathcal{P} \subseteq \mathcal{X}} \frac{k(x, x')}{\sqrt{k(x, x) k(x', x')}} \tag{6.39}$$

Proof. By the definition of MMD we can see that $-\epsilon^2_{\mathrm{BQ}} = \inf_{w \in \mathbb{R}^N} \|\mu_p - \sum_{n=1}^{N} w^{(n)}_{\mathrm{BQ}} k(\cdot, x_n)\|^2_{\mathcal{H}}$ is the negative squared distance between the mean element μ_p and its projection onto the subspace spanned by the elements $k(\cdot, x_n)$. Substituting $k = 1$ into Theorem 1 of **?** concludes the proof. □

Unfortunately, weak submodularity does not provide the strong near-optimality guarantees as submodularity does . If $s : 2^{\mathcal{X}} \mapsto \mathbb{R}$ is a weakly submodular function with constant ϵ, and $|\mathcal{A}_n| = n$ is the result of greedy optimisation of s, then

$$s(\mathcal{A}_n) \geq \left(1 - \frac{1}{e}\right) \max_{|\mathcal{A}| \leq n} s(\mathcal{A}) - n\epsilon \tag{6.40}$$

As pointed out by **?**, this guarantee is very weak as in our case the objective function $\epsilon^2_{\text{BQ}}(\emptyset) - \epsilon^2_{\text{BQ}}(\cdot)$ is upper bounded by a constant. However, establishing a connection between SBQ and sparse dictionary selection problem opens up interesting directions for future research, and it may be possible to apply algorithms and theory developed for sparse dictionary selection to kernel-based quasi-Monte Carlo methods.

EXPERIMENTS

In this section, we examine empirically the rates of convergence of sequential Bayesian quadrature and herding. We examine both the expected error rates, and the empirical error rates.

In all experiments, the target distribution p is chosen a 2D mixture of 20 Gaussians, whose equiprobability contours are shown in Figure 6.1. To ensure a comparison fair to herding, the target distribution, and the kernel used by both methods, correspond exactly to the one used in [**?**, Fig. 1]. For experimental simplicity, each of the sequential sampling algorithms minimizes the next sample location from a pool of 10000 locations randomly drawn from the base distribution. In practice, one would run a local optimizer from each of these candidate locations, however in our experiments we found that this did not make a significant difference in the sample locations chosen.

Matching a distribution

We first extend an experiment from [**?**] designed to illustrate the mode-seeking behavior of herding in comparison to random samples. In that experiment, it is shown that a small number of i. i. d. samples drawn from a multimodal distribution will tend to, by chance, assign too many samples to some modes, and too few to some other modes. In contrast, herding places 'super-samples' in such a way as to avoid regions already well-represented, and seeks modes that are under-represented.

We demonstrate that although herding improves upon i. i. d. sampling, the uniform weighting of super-samples leads to sub-optimal performance. Figure 6.1 shows the first 20 samples chosen by kernel herding, in comparison with the first 8 samples chosen by SBQ. By weighting the 8 SBQ samples by the quadrature weights in (6.23), we can obtain the same expected loss as by using the 20 uniformly-weighted herding samples. Figure 6.5 shows MMD versus the number of samples added, on the distribution shown in Figure 6.1. We can see that in all cases, SBQ dominates herding. It appears that SBQ converges at a faster rate than $\mathcal{O}(1/N)$, although the form of this rate is unknown.

There are two differences between herding and SBQ: SBQ chooses samples according to a different criterion, and also weights those samples differently. We may ask whether the sample locations or the weights are contributing more to the faster convergence of SBQ. Indeed, in Figure 6.1 we observe that the samples selected by SBQ are quite similar to the samples selected by kernel herding. To answer this question, we also plot in Figure 6.5 the performance of a fourth method, which selects samples using herding, but later re-weights the herding samples with BQ weights. Initially, this method attains similar performance to SBQ, but as the number of samples increases, SBQ attains a better rate of convergence. This result indicates that the different sample locations chosen by SBQ, and not only the optimal weights, are responsible for the increased convergence rate of SBQ.

Estimating Integrals

We then examined the empirical performance of the different estimators at estimating integrals of real functions. To begin with, we looked at performance on 100 randomly drawn functions, of the

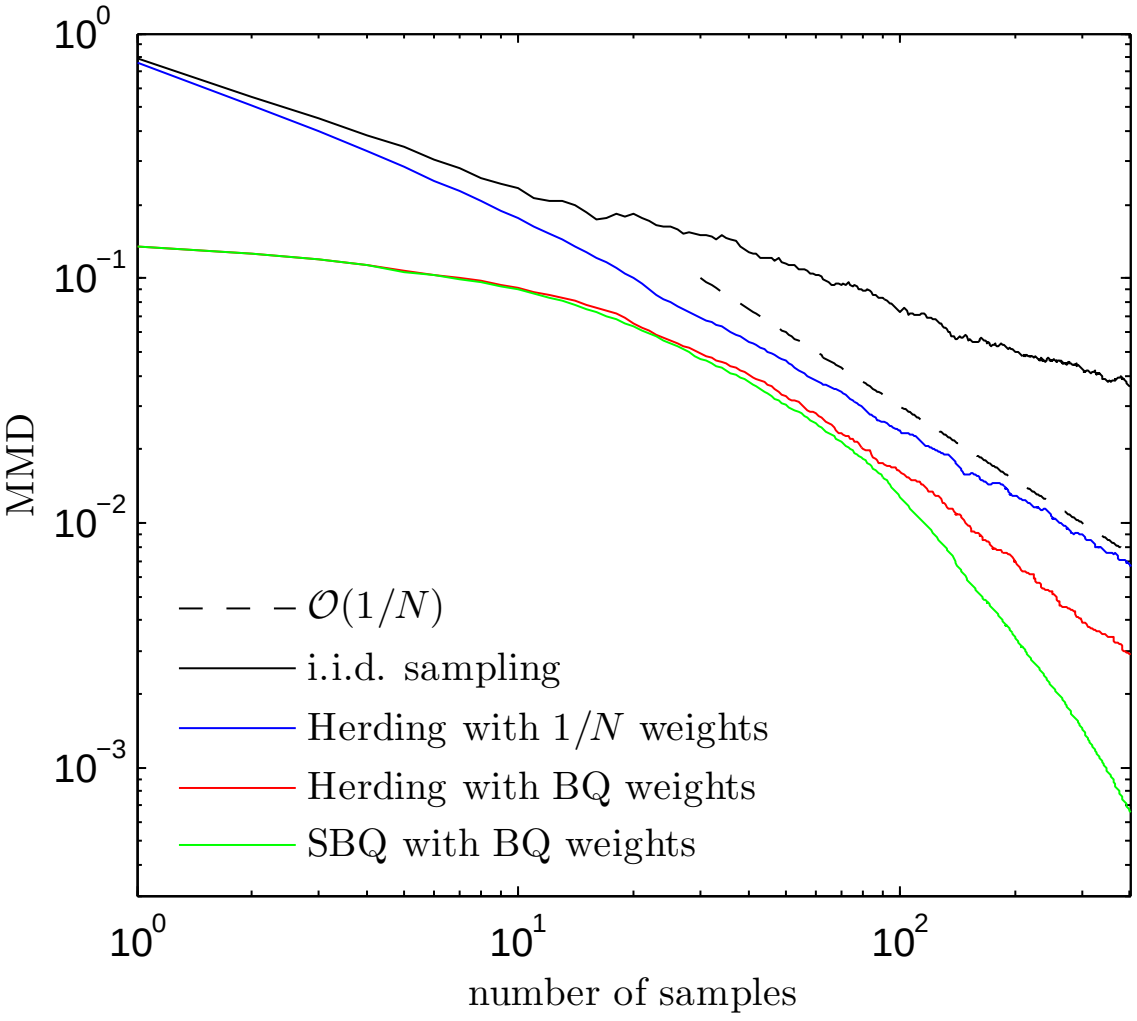

Figure 6.5: The maximum mean discrepancy, or expected error of several different quadrature methods. Herding appears to approach a rate close to $\mathcal{O}(1/N)$. SBQ appears to attain a faster, but unknown rate.

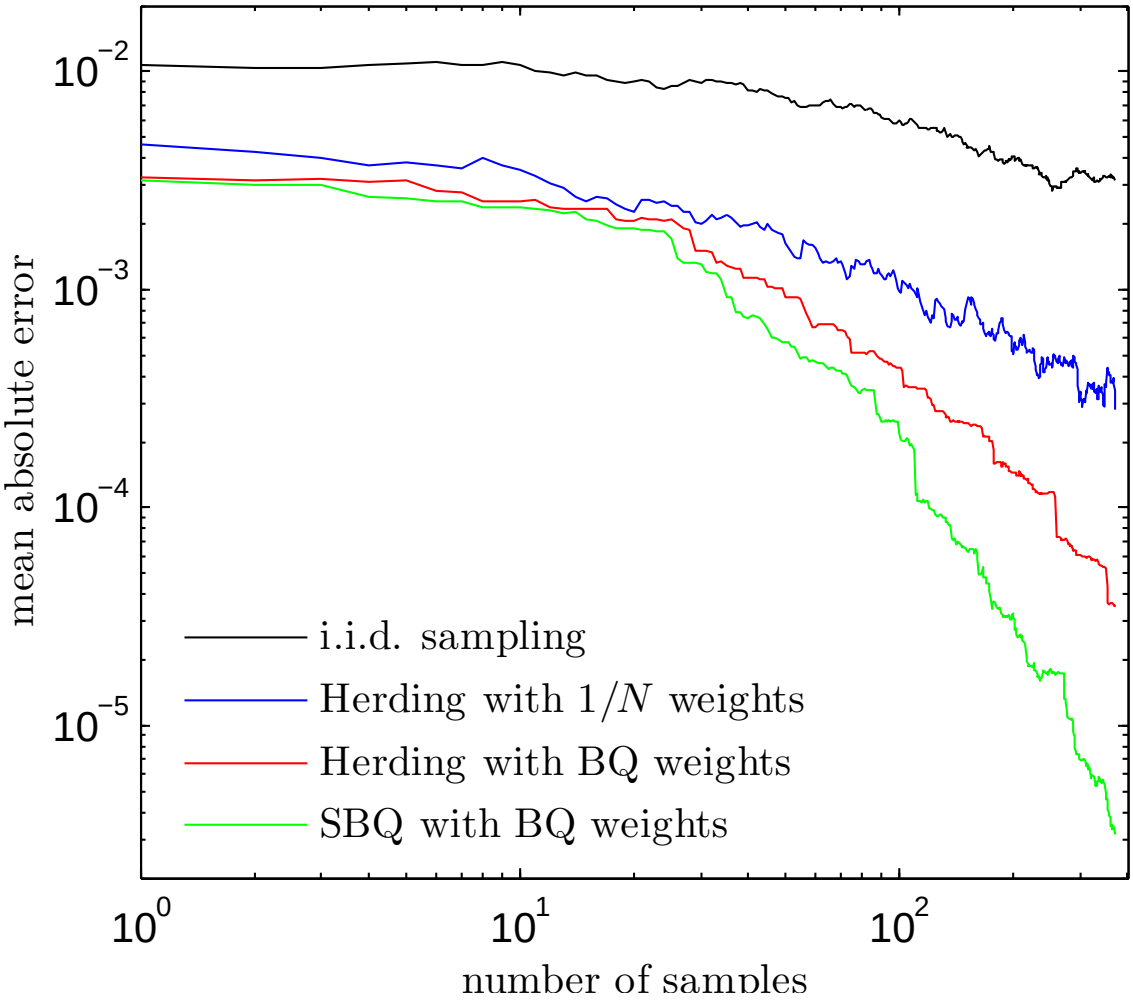

Figure 6.6: Within-model error: The empirical error rate in estimating $Z_{f,p}$, for several different sampling methods, averaged over 250 functions randomly drawn from the RKHS corresponding to the kernel used.

form:

$$f(x) = \sum_{i=1}^{10} \alpha_i k(x, c_i) \tag{6.41}$$

where

$$\|f\|_{\mathcal{H}}^2 = \sum_{i=1}^{10} \sum_{j=1}^{10} \alpha_i \alpha_j k(c_i, c_j) = 1 \tag{6.42}$$

That is, these functions belonged exactly to the unit ball of the RKHS defined by the kernel $k(x, x')$ used to model them. Figure 6.6 shows the empirical error versus the number of samples, on the distribution shown in Figure 6.1. The empirical rates attained by the method appear to be similar to the MMD rates in Figure 6.5.

By definition, MMD provides a upper bound on the estimation error in the integral of any function in the unit ball of the RKHS (Eqn. (6.6)), including the Bayesian estimator, SBQ. Figure 6.7 demonstrates this quickly decreasing bound on the SBQ empirical error.

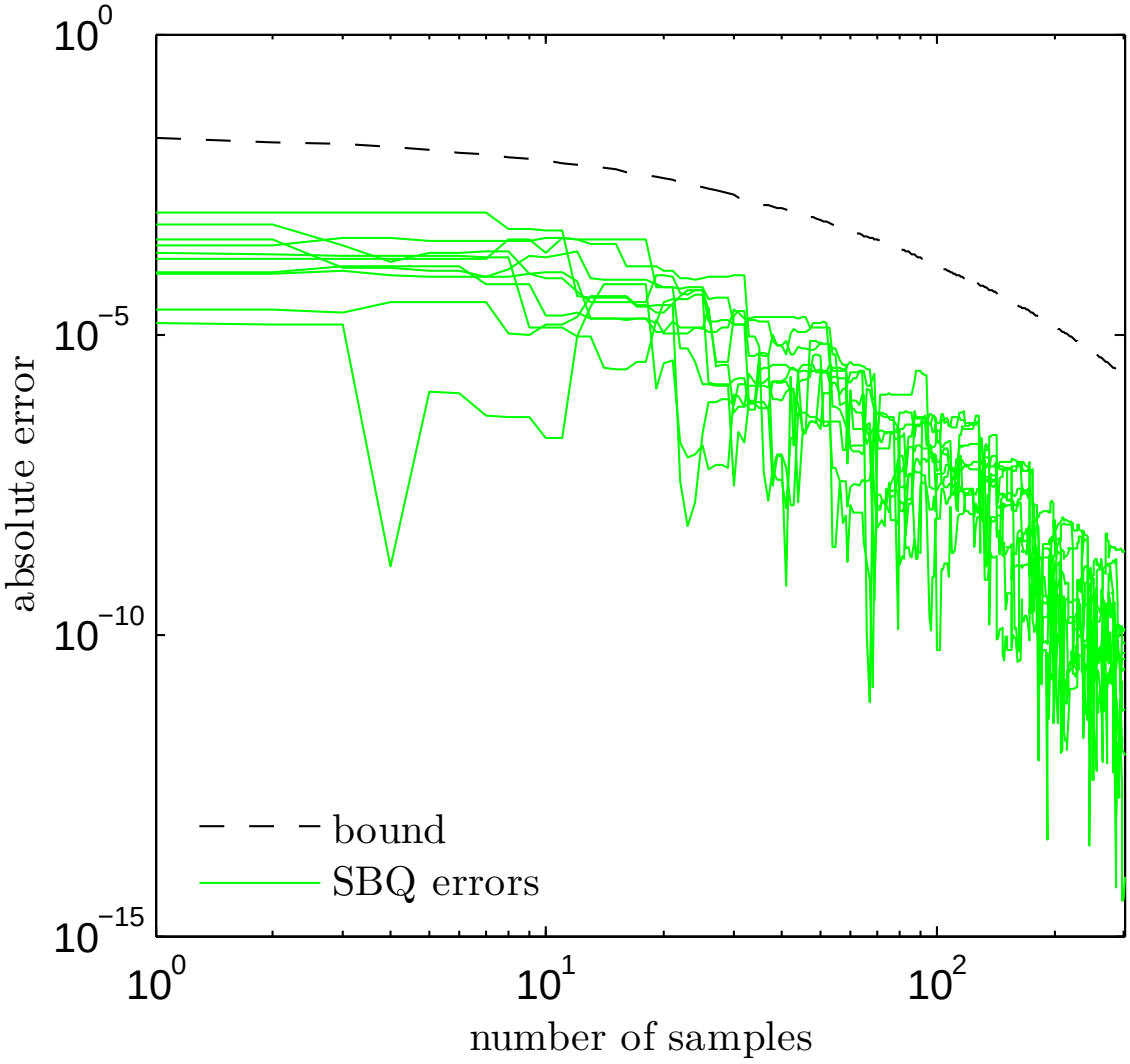

Figure 6.7: The empirical error rate in estimating $Z_{f,p}$, for the SBQ estimator, on 10 random functions drawn from the RKHS corresponding to the kernel used. Also shown is the upper bound on the error rate implied by the MMD.

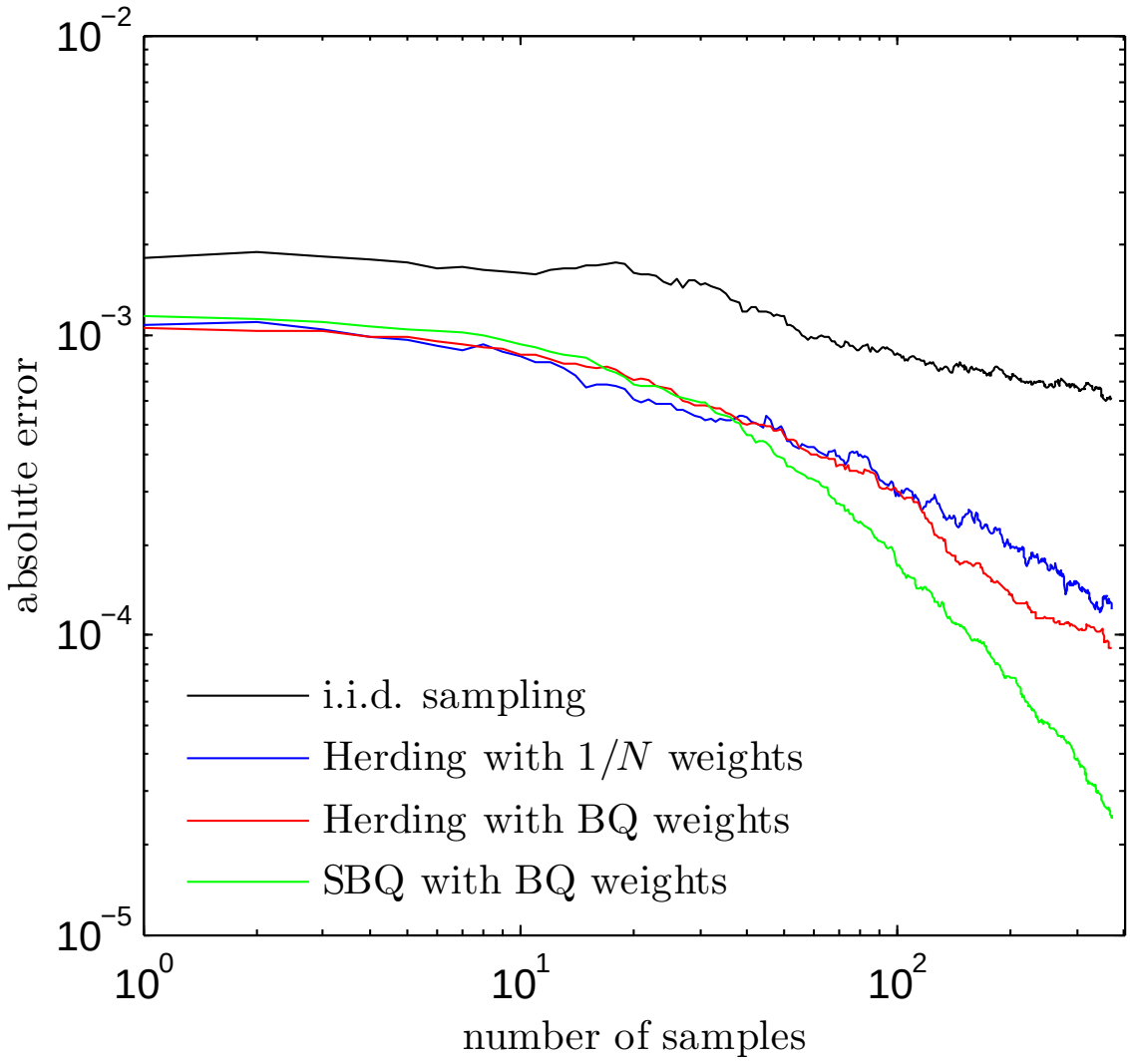

Figure 6.8: Out-of-model error: The empirical error rates in estimating $Z_{f,p}$, for several different sampling methods, averaged over 250 functions drawn from outside the RKHS corresponding to the kernels used.

Out-of-model performance

A central assumption underlying SBQ is that the integrand function belongs to the RKHS specified by the kernel. To see how performance is effected if this assumption is violated, we performed empirical tests with functions chosen from outside the RKHS. We drew 100 functions of the form:

$$f(x) = \sum_{i=1}^{10} \alpha_i \exp(-\frac{1}{2}(x - c_i)^T \Sigma_i^{-1}(x - c_i) \tag{6.43}$$

where each α_i c_i Σ_i were drawn from broad distributions. This ensured that the drawn functions had features such as narrow bumps and ridges which would not be well modelled by functions belonging to the isotropic kernel defined by k. Figure 6.8 shows that, on functions drawn from outside the assumed RKHS, relative performance of all methods remains similar.

Code to reproduce all results is available at `http://mlg.eng.cam.ac.uk/duvenaud/`

method	complexity	rate	guarantee
MCMC	$\mathcal{O}(N)$	variable	ergodic theorem
i.i.d. MC	$\mathcal{O}(N)$	$\frac{1}{\sqrt{N}}$	law of large numbers
herding	$\mathcal{O}(N^2)$	$\frac{1}{\sqrt{N}} \geq \cdot \geq \frac{1}{N}$	[**??**]
SBQ	$\mathcal{O}(N^3)$	unknown	approximate submodularity

Table 6.1: A comparison of the rates of convergence and computational complexity of several integration methods.

DISCUSSION

Choice of Kernel

Using herding techniques, we are able to achieve fast convergence on a Hilbert space of *well-behaved* functions, but this fast convergence is at the expense of the estimate not necessarily converging for functions outside this space. If we use a characteristic kernel [**?**], such as the exponentiated-quadratic or Laplacian kernels, then convergence in MMD implies weak convergence of q_N to the target distribution. This means that the estimate converges for any bounded measurable function f. The speed of convergence, however, may not be as fast.

Therefore it is crucial that the kernel we choose is representative of the function or functions f we will integrate. For example, in our experiments, the convergence of herding was sensitive to the width of the Gaussian kernel. One of the major weaknesses of kernel methods in general is the difficulty of setting kernel parameters. A key benefit of the Bayesian interpretation of herding and MMD presented in this paper is that it provides a recipe for adapting the Hilbert space to the observations $f(x_n)$. To be precise, we can fit the kernel parameters by maximizing the marginal likelihood of Gaussian process conditioned on the observations. Details can be found in [**?**].

Computational Complexity

While we have shown that Bayesian Quadrature provides the optimal re-weighting of samples, computing the optimal weights comes at an increased computational cost relative to herding. The computational complexity of computing Bayesian quadrature weights for N samples is $\mathcal{O}(N^3)$, due to the necessity of inverting the Gram matrix $K(x, x)$. Using the Woodbury identity, the cost of adding a new sample to an existing set is $\mathcal{O}(N^2)$. For herding, the computational complexity of evaluating a new sample is only $\mathcal{O}(N)$, making the cost of choosing N herding samples $\mathcal{O}(N^2)$. For Monte Carlo, the cost of adding an i.i.d. sample from the target distribution is only $\mathcal{O}(1)$.

The relative computational cost of computing samples and weights using BQ, herding, and sampling must be weighed against the cost of evaluating f at the sample locations. Depending on this trade-off, the three sampling methods form a Pareto frontier over computational speed and estimator accuracy. When computing f is cheap, we may wish to use Monte Carlo methods. In cases where f is computationally costly, we would expect to choose the SBQ method. When f is relatively expensive, but a very large number of samples are required, we may choose to use kernel herding instead. However, because the rate of convergence of SBQ is faster, there may be situations in which the $\mathcal{O}(N^3)$ cost is relatively inexpensive, due to the smaller N required by SBQ to achieve the same accuracy as compared to using other methods.

There also exists the possibility to switch to a less costly sampling algorithm as the number of samples increases. Table 6.1 summarizes the rates of convergence of all the methods considered here.

6.1 CONCLUSIONS

In this paper, we have shown two main results: First, we proved that the loss minimized by kernel herding is closely related to the loss minimized by Bayesian quadrature, when selecting sample locations. This implies that sequential Bayesian quadrature can viewed as an optimally-weighted version of kernel herding.

Second, we showed that the loss minimized by the Bayesian method is approximately submodular with respect to the samples chosen, and established connections to the submodular dictionary selection problem studied in [**?**].

Finally, we empirically demonstrated a superior rate of convergence of SBQ over herding, and demonstrated a bound on the empirical error of the Bayesian quadrature estimate.

Future Work

In section 6, we showed that SBQ is approximately submodular, which provides only weak sub-optimality guarantees of its performance. It would be of interest to further explore the connection between Bayesian Quadrature and the dictionary selection problem to see if algorithms developed for dictionary selection can provide further practical or theoretical developments. The results in section 6, specifically Figure 6.5, suggest that the convergence rate of SBQ is faster than $\mathcal{O}(1/N)$. However, we are not aware of any work showing what the theoretically optimal rate is. It would be of great interest to determine this optimal rate of convergence for particular classes of kernels.

Acknowledgements

Part III

Optimal Experiment Design

Chapter 7

A Bayesian Framework for Experiment Design

In most machine learning applications, a learner passively observes data with which it can make inferences about its environment. It is generally true that as more data becomes available the inferences become more accurate. However, not each and every datapoint is equally useful. Some datapoints will be critically informative, while many more will become redundant given the context and information already learnt from other examples.

It is intuitive to think that, by actively seeking out measurements to be used in inference, the learner can significantly improve the quality of inference using smaller quantities of data. Amongst machine learning researchers this process of choosing which measurements to take is known as active learning; the same problem is called optimal experimental design in the statistics literature. Although active learning has been studied for several decades [**??**], it is still an active area of research and no general solution exists.

The active learning paradigm is as pertinent now as it has ever been. With the advent and rapid expansion of the Internet, very large amounts of unlabelled data have become available; however, it is relatively costly to obtain labels. Therefore, one must seek the most informative data points. Experimental scientists work with ever growing volumes of data, carrying out experiments or labeling datapoints is a time-consuming and costly process for them. Carefully pre-selecting only the most informative experiments can result in substantial improvements in terms of faster processing or reduced costs. Searching for the most useful data in vast spaces of measurements calls for powerful active learning algorithms.

In this chapter I explain how scoring-rule based information quantities described in Chapter **??** can be used to formalise the problem of active learning and experiment design. I devise a framework that is flexible enough to accomodate and connect a wide range of existing techniques. Examples include decision theoretic active learning, Bayesian optimisation and Bayesian quadrature.

In the second half of this chapter I focus on a special case of Bayesian active learning that attempts to maximise Shannon's information. I derive a computationally convenient method, called Bayesian Active Learning by Disagreement (BALD) and present multiple applications to binary classification, multi-user preference learning and quantum physics.

7.1 A general framework for Bayesian experiment design

In active learning the goal is to learn about dependence of some variable $y \in \mathcal{Y}$ on the input variable $x \in \mathcal{X}$ by interactively querying the system with inputs $x_i \in \mathcal{X}$ and observing the system's response y_i. Ultimately, having observed data $\mathcal{D} = \{(x_i, y_i)\}$, our goal is to choose queries such that the observed outcomes provide us with the most information about relevant properties of the system. Different approaches quantify information in different ways. Here we take a Bayesian

approach, that assumes the existence of some latent parameters $\boldsymbol{\theta}$, that control the dependence between inputs and outputs, $p(y|x, \boldsymbol{\theta})$.

Our goal is to infer the value of $\boldsymbol{\theta}$ from the observed data $\mathcal{D} = \{(x_i, y_i)\}$, which is possible via Bayes' rule

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}} \tag{7.1}$$

Here I will assume that inference is possible without approximations, and the posterior $p(\boldsymbol{\theta}|\mathcal{D})$ is available in closed form. In practice this is rarely the case, but it simplifies the analysis of active learning.

A core problem in active learning is to describe how informative data is. In the Bayesian inference framework the posterior $p_{\mathcal{D}}(\boldsymbol{\theta}) := p(\boldsymbol{\theta}|\mathcal{D})$ captures and summarises everything there is to know about the parameter $\boldsymbol{\theta}$ based on the data $\mathcal{D}$. It therefore makes sense to assess the quality of data $\mathcal{D}$ in terms of the quality of the posterior $p_{\mathcal{D}}$. Informative data should allow one to construct an accurate prediction $p_{\mathcal{D}}$ of the parameters θ. If the data is redundant, our estimate $p_{\mathcal{D}}$ is going to be poor.

The posterior $p_{\mathcal{D}}$ is a probabilistic estimate, hence it's accuracy can be quantified using a scoring rule $S(\boldsymbol{\theta}, p_{\mathcal{D}})$. The goal of the active learner should be to gather data $\mathcal{D}$ so that $p_{\mathcal{D}}$ minimises this quantity. However, during the process of active learning, the parameters $\boldsymbol{\theta}$ or indeed the score $S(\boldsymbol{\theta}, p_{\mathcal{D}})$ are never explicitly revealed to the learner, otherwise there was no point in learning. The best strategy the learner can follow is to collect data so that the zn estimate of this score is minimised. The Bayes estimator to the score $S(\boldsymbol{\theta}, p_{\mathcal{D}})$ is the expected score or generalised entropy of the posterior $p_{\mathcal{D}}$.

Hence, in the scoring rule framework, the information in data is quantified by the (negative) generalised entropy of the posterior. When selecting the next measurement x this is the quantity we should therefore aim to decrease.

$$\operatorname*{argmax}_x \left[\mathbb{H}_S\left[p(\boldsymbol{\theta}|\mathcal{D})\right] - \mathbb{E}_{y\sim p(y|x,\mathcal{D})}\mathbb{H}_S\left[p(\boldsymbol{\theta}|x, y, \mathcal{D})\right]\right] \tag{7.2}$$

In the scoring rule framework, the goal of active learning is to choose measurements x whose value of information $formula$ with regards to the latent parameters $\boldsymbol{\theta}$ is maximal.

7.2 Examples and special cases

7.2.1 Shannon's entropy

7.2.2 Decision theoretic active classification

Often, applying active learning in a general supervised learning, the performance is assessed in terms of average prediction error on a held-out test dataset. If this is the case, the scoring rule used to define information in the framework should be chosen to reflect this.

$$S_\ell(\boldsymbol{\theta}, p_{\mathcal{D}}) = \iint \ell(\hat{y}(x, p_{\mathcal{D}}), y)p(y|x, \boldsymbol{\theta})dy\ p_{test}(x)dx, \tag{7.3}$$

where p_{test} is the distribution of test examples, $\ell(y, y')$ is the loss incurred for prediction y when the true output is y', and $\hat{y}(x, p)$ is the predicted label for input x given the posterior $p_{\mathcal{D}}$ over $\boldsymbol{\theta}$. $\hat{y}(x, p)$ can be the Bayes decision, but this is not a requirement, as long as it is consistent with how decisions are computed in the final evaluation.

Exact decesion theoretic active learning is complicated to achieve in practice, however approximation strategies exist. apply **?** use a similar objective function to perform graph-based active learning. A related approach minimises the Shannon mutual information between the label of the chosen point and the unseen labels of a pool of unlabelled points in a transductive setting.

7.2.3 Bayesian optimisation

7.2.4 Bayesian quadrature

7.3 Bayesian active learning by Disagreement (BALD)

Chapter 8

Active Learning in Gaussian Process Models

Chapter 9

Adaptive Bayesian Quantum Tomography

9.1 Introduction

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

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

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

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

importance.

In this part of the thesis I will formally introduce the problem of quantum state tomography, provide Bayesian analysis of the problem and then propose a

9.2 Overview of quantum statistics

quantum states

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

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

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

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

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

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

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

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

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

$$e = mc^2 \tag{9.1}$$

In both cases the probability of observing 0 and 1 is the same for both sources, and is a function of the measurement. We can therefore conclude that no matter what measurements we perform, there is no way to tell apart the two sources on the basis of observations. We therefore may call these two sources *observationally equivalent*. In more general terms, classical and quantum

uncertainty cannot be disambiguated by observing a system. We can therefore define a equivalence classes of systems, and parametrise them via the so called density matrix ρ.

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

Born rule

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

9.2.1 Inference in quantum tomography

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

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

9.2.2 optimal experiment design and active tomography

9.3 Adaptive Bayesian Quantum Tomography

9.4 Results

Bibliography

S. Amari and A. Cichocki. Information geometry of divergence functions. *Bulletin of the Polish Academy of Sciences: Technical Sciences*, 58(1):183–195, 2010.

J. Besag. Efficiency of pseudolikelihood estimation for simple gaussian fields. *Biometrika*, pages 616–618, 1977.

G.W. Brier. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3, 1950.

Y. Chen, M. Welling, and A. Smola. Super-samples from kernel herding. *arXiv preprint arXiv:1203.3472*, 2012.

F. Comets. On consistency of a class of estimators for exponential families of markov random fields on the lattice. *The Annals of Statistics*, pages 455–468, 1992.

L. Csató and M. Opper. Sparse on-line gaussian processes. *Neural Computation*, 14(3):641–668, 2002.

I. Csiszár and Z. Talata. Consistent estimation of the basic neighborhood of markov random fields. In *Information Theory, 2004. ISIT 2004. Proceedings. International Symposium on*, page 170. IEEE, 2004.

A.P. Dawid. The geometry of proper scoring rules. *Annals of the Institute of Statistical Mathematics*, 59(1):77–93, 2007.

A.P. Dawid and P. Sebastiani. Coherent dispersion criteria for optimal experimental design. *Annals of Statistics*, pages 65–81, 1999.

A.P. Dawid, S. Lauritzen, and M. Parry. Proper local scoring rules on discrete sample spaces. *The Annals of Statistics*, 40(1):593–608, 2012.

E. del Barrio, J.A. Cuesta-Albertos, C. Matrán, J. Rodríguez-Rodríguez, et al. Tests of goodness of fit based on the l_2-wasserstein distance. *The Annals of Statistics*, 27(4):1230–1239, 1999.

R.M. Dudley. Metric entropy of some classes of sets with differentiable boundaries. *Journal of Approximation Theory*, 10(3):227–236, 1974.

M.L. Eaton, A. Giovagnoli, and P. Sebastiani. A predictive approach to the bayesian design problem with application to normal regression models. *Biometrika*, 83(1):111–125, 1996.

C.A.T. Ferro. Comparing probabilistic forecasting systems with the brier score. *Weather and Forecasting*, 22(5):1076–1088, 2007.

K. Fukumizu, L. Song, and A. Gretton. Kernel bayes' rule. *arXiv preprint arXiv:1009.5736*, 2010.

T. Gneiting and A.E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

Irving John Good. discussion of proper scores for probability forecasters by hendrickson and buehler. discussion paper, 1971.

A. Gretton, K. Fukumizu, Z. Harchaoui, and B.K. Sriperumbudur. A fast, consistent kernel two-sample test. *Advances in neural information processing systems*, 22:673–681, 2009.

A. Gretton, K.M. Borgwardt, M.J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13:723–773, 2012.

A. Hyvärinen. Consistency of pseudolikelihood estimation of fully visible boltzmann machines. *Neural Computation*, 18(10):2283–2292, 2006.

V.R.R. Jose, R.F. Nau, and R.L. Winkler. Scoring rules, generalized entropy, and utility maximization. *Operations Research*, 56(5):1146–1157, 2008.

David J. C. MacKay. *Information Theory, Inference & Learning Algorithms*. Cambridge University Press, New York, NY, USA, 2002. ISBN 0521642981.

T.P. Minka. Expectation propagation for approximate bayesian inference. In *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, pages 362–369. Morgan Kaufmann Publishers Inc., 2001.

Iain Murray. *Advances in Markov chain Monte Carlo methods*. PhD thesis, Gatsby computational neuroscience unit, University College London, 2007.

J.K. Ord, S.F. Arnold, A. O'Hagan, and J. Forster. *Kendall's advanced theory of statistics*. A. Arnold, 1999.

Matthew Parry, A. Philip Dawid, and Steffen Lauritzen. Proper local scoring rules. *Annals of Statistics*, 40(1):561–592, 2012.

B. Póczos and J. Schneider. On the estimation of α-divergences. In *Proc. 14th Int. Conf. AI and Stat.(Fort Lauderdale, FL, 11–13 April 2011,)*, pages 609–17, 2011.

D.A. Redelmeier, D.A. Bloch, and D.H. Hickam. Assessing predictive accuracy: how to compare brier scores. *Journal of Clinical Epidemiology*, 44(11):1141–1146, 1991.

David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Neurocomputing: foundations of research. chapter Learning representations by back-propagating errors, pages 696–699. MIT Press, Cambridge, MA, USA, 1988. ISBN 0-262-01097-6. URL `http://dl.acm.org/citation.cfm?id=65669.104451`.

Claude E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 1948.

A. Smola, A. Gretton, L. Song, and B. Schölkopf. A hilbert space embedding for distributions. In *Algorithmic Learning Theory*, pages 13–31. Springer, 2007.

L. Song, X. Zhang, A. Smola, A. Gretton, and B. Schölkopf. Tailoring density estimation via reproducing kernel moment matching. In *Proceedings of the 25th international conference on Machine learning*, pages 992–999. ACM, 2008.

D.J. Spiegelhalter. Probabilistic prediction in patient management and clinical trials. *Statistics in Medicine*, 5(5):421–433, 2006.

B.K. Sriperumbudur, A. Gretton, K. Fukumizu, G. Lanckriet, and B. Schölkopf. Injective hilbert space embeddings of probability measures. 2008.

B.K. Sriperumbudur, K. Fukumizu, A. Gretton, B. Schölkopf, and G.R.G. Lanckriet. On integral probability metrics,\ phi-divergences and binary classification. *arXiv preprint arXiv:0901.2698*, 2009.

M. Welling. Herding dynamical weights to learn. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1121–1128. ACM, 2009.