Proper Measures of Divergence and Information in Probabilistic Machine Learning

Ferenc Huszár

Computational and Biological Learning Lab
Department of Engineering

Trinity College

University of Cambridge

A thesis submitted for the degree of

Doctor of Philosophy

Yet to be decided

I would like to dedicate this thesis to Jurgen ...

Acknowledgements

Foo bar. Baz.

Abstract

This is where you write your abstract ...

Contents

List of Figures

Chapter 1

Introduction

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus.

Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et vehicula libero dui cursus dui. Mauris tempor ligula sed lacus. Duis cursus enim ut augue. Cras ac magna. Cras nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consectetuer.

Part I

Scoring rules, Divergences and Information

Chapter 2

An introduction to scoring rules

In this section I describe scoring rules, a framework for quantifying the accuracy of probabilistic forecasts. Scoring rules allow one to define useful generalisations of well-known information quantities, such as entropy, divergence and mutual information. Each scoring rule defines a unique geometry over probabilistic models, which can be exploited in a variety of statistical applications. They provide a unifying framework for problems such as parameter estimation, approximate Bayesian inference and Bayesian 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 functional that quantifies this: if the outcome is $X = x$, and our prediction was P we incur a score $S(x, P)$. Mathematically, a scoring rule can be any measurable function that maps an outcome-probability distribution pair onto real numbers: $S : \mathcal{X} \times \mathcal{M}_{\mathcal{X}}^1 \mapsto \mathcal{R} \cup \{\infty\}$. 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 the central quantity of interest in maximum likelihood estimation. The logarithmic scoring rule is a very important example and has several unique characteristics (see section 2.2.1), 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 instead of the logarithmic. In this chapter will give further examples of scoring rules and describe where they have been applied in statistics or machine learning.

2.1 Information quantities

A scoring rule allows us to define useful information quantities, which can be exploited in a variety of applications [see also **?**]: these are generalised notions entropy, divergence and value of information.

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

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

This entropy measures how difficult it is to forecast the outcome on average, when true distribution P of outcomes is known and used as the forecasting model. One 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 2.2.7).

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

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

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

The divergence measures how much worse off one would be using some probability distribution Q, rather than P, to forecast a quantity X, which is indeed sampled from P. It can be interpreted as a measure of dissimilarity between two distributions P and Q. Divergences are normally non-symmetric, that is $d_S[P \| Q] \neq d_S[Q \| P]$.

Since scoring rules measure how accurate a probabilistic forecast is, it is desirable that using the true probability model P should never incur a higher average score than using an incorrect model Q. If that would be the case, the divergence would always be non-negative. However, this is not automatically true for all scoring rules. A scoring rule that has this desirable property is called a *proper scoring rule*.

Definition 3 (Proper scoring rule)**.** $S : \mathcal{X} \times \mathcal{M}_{\mathcal{X}}^1 \mapsto \mathbb{R}$ 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. Ω if equality holds only when $P = Q$.

Strictly proper scoring rules can therefore detect - on average - whether a forecast Q matches the true distribution of the unknown quantity P. This property is exploited in score matching, where a parametric probability model is fitted to i. i. d. observations.

Definition 4 (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:

$$\hat{\theta}_N(x_1, \ldots, x_N) = \underset{\theta\in\Theta}{\operatorname{argmin}} \sum_{n=1}^{N} S(x_n, P_{X|\theta}) \tag{2.5}$$

For most scoring rules the above estimating equation can be formulated in terms of the divergence as follows.

$$\hat{\theta}_N(x_1, \ldots, x_N) = \underset{\theta\in\Theta}{\operatorname{argmin}}\, d_S\left[\frac{1}{N}\sum_{n=1}^{N}\delta(x - x_n)\middle\|P_{X|\theta}\right] \tag{2.6}$$

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 when $x_1, \ldots \sim P_{X|\theta_0}$ i. i. d.

$$\lim_{N\to\infty} \hat{\theta}_N(x_1, \ldots, x_N) = \theta_0 \qquad P_{X|\theta_0}\text{- almost surely} \tag{2.7}$$

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 measures of distance or dissimilarity such as the Eulidean distance or KL divergence.

Definition 5 (Bregman divergence)**.** Let H be a differentiable, strictly concave function on a convex domain Θ. For $P, Q \in \Theta$

$$d_{Bregman,H}[P\|Q] = H(P) - H(Q) + \langle \nabla H(Q), Q - P\rangle \tag{2.8}$$

Statement 1 (Generalised divergences d_S for strictly proper S are Bregman diver-

gences). *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 (sketch). Review the definition of the entropy $\mathbb{H}_S[P]$, using linear algebra notation for the expectation [Amari and Cichocki, 2010]:

$$\mathbb{H}_S[P] = \mathbb{E}_{x\sim P}S(x,P) = \langle P, S(\cdot,P)\rangle \tag{2.9}$$

Using this notation, noting the linearity of scalar products (expectation)

$$\begin{aligned}\nabla\mathbb{H}_S[P] &= \nabla\langle P, S(\cdot,P)\rangle &(2.10)\\ &= S(\cdot,P) + \langle P, \nabla S(\cdot,P)\rangle &(2.11)\end{aligned}$$

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

$$\begin{aligned}d_{Bregman,\mathbb{H}_S}[P\|Q] &= \mathbb{H}_S[Q] + \langle \nabla\mathbb{H}_S[Q], P-Q\rangle - \mathbb{H}_S[P] &(2.12)\\ &= \mathbb{H}_S[Q] + \langle S(\cdot,Q), P-Q\rangle - \mathbb{H}_S[P] &(2.13)\\ &= \langle S(\cdot,Q), P\rangle - \mathbb{H}_S[P] &(2.14)\\ &= d_S[P\|Q] &(2.15)\end{aligned}$$

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

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 2.1.

The information quantities introduced so far only dealt with single random variable X, and comparing probability distributions over the same variable. In the following I will define information quantities that describe the relationship and dependence between more than one variable. A particularly useful quantity is the value of information, which quantifies how much useful information one random variable Y holds about another one X.

Definition 6 (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

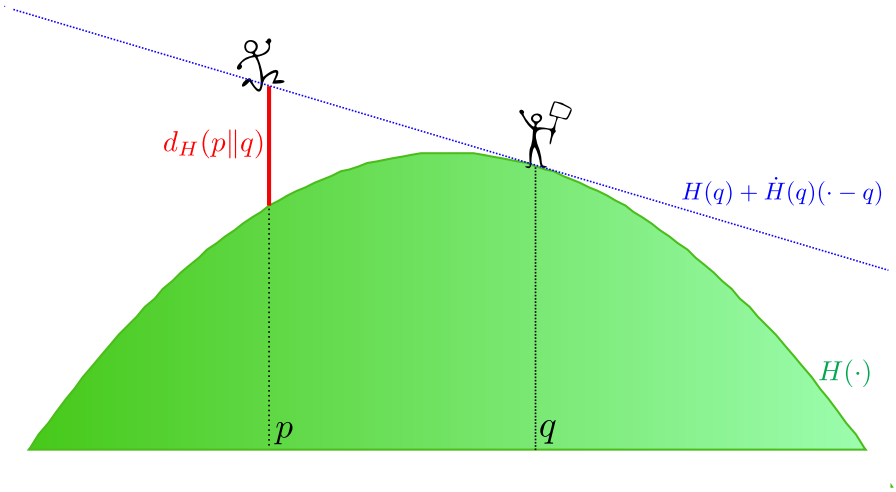

Figure 2.1: Pictorial illustration of Bregman divergences. Peter and Quentin are points who live on a convex hill, whose surface is described by the concave function $H(\cdot)$. Peter lives at $(p, H(p))$, Quentin at $(q, H(q))$. Because the hill is convex and they are both points, they cannot normally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

the scoring rule S as

$$\mathbb{I}_S\left[X \leftarrow Y\right] = \mathbb{E}_{x\sim P_X} S(x, P_X) - \mathbb{E}_{y\sim P_Y}\mathbb{E}_{x\sim P_{X|Y=y}} S(x, P_{X|Y=y}) \tag{2.16}$$

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

$$\mathbb{I}_S\left[X \leftarrow Y\right] = \mathbb{H}_S\left[P_X\right] - \mathbb{E}_{y\sim P_Y}\mathbb{H}_S\left[P_{X|Y=y}\right] \tag{2.17}$$

$$= \mathbb{E}_{y\sim P_Y} d_S\left[P_{X|Y=y}\middle\|P_X\right] \tag{2.18}$$

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

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

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

1. $\mathbb{I}_S\left[X \leftarrow Y\right] = 0$
2. *$X \perp\!\!\!\perp Y$; 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 $\forall y : d_S\left[P_{X|Y=y}\middle\|P_X\right] = 0$, and hence $\mathbb{I}_S\left[X \leftarrow Y\right] = 0$.

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

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

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

Corollary 2 (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 $\mathbb{I}_{S_Y}[Y \leftarrow X] > 0$.*

We can also define a conditional version of this quantity which measures how much additional information Y provides about X given the value of a third variable Z which is also observed.

Definition 7 (Conditional value of information)**.** Let X, Y, Z be random variables with joint distribution $P \in \mathcal{M}_{\mathcal{X}\times\mathcal{Y}\times\mathcal{Z}}^1$. Let $S : \mathcal{X} \times \mathcal{M}_{\mathcal{X}}^1 \mapsto \mathbb{R}$ 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

$$\mathbb{I}_S[X \leftarrow Y|Z = z] = \mathbb{E}_{x\sim P_{X|Z=z}} S(x, P_{X|Z=z}) - \mathbb{E}_{y\sim P_{Y|Z=z}} \mathbb{E}_{x\sim P_{X|Y=y,Z=z}} S(x, P_{X|Y=y,Z=z}) \tag{2.19}$$

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

$$\mathbb{I}_S[X \leftarrow Y|Z = z] = \mathbb{H}_S\left[P_{X|Z=z}\right] - \mathbb{E}_{y\sim P_{Y|Z=z}} \mathbb{H}_S\left[P_{X|Y=y,Z=z}\right] \tag{2.20}$$

$$= \mathbb{E}_{y\sim P_Y} d_S\left[P_{X|Y=y,Z=z} \middle\| P_{X|Z=z}\right] \tag{2.21}$$

Just like in the case of non-conditional value of information, the definition only calls for a scoring rule over X, not over the other variables X or Z. Just like the value of information was related to statistical independence, conditional value of information is related to conditional independence in the following sense.

Statement 2. *Let $S : \mathcal{X} \times \mathcal{M}_{\mathcal{X}}^1 \mapsto \mathbb{R}$ be a strictly proper scoring rule with respect to Borel probability distributions $\mathcal{M}_{\mathcal{X}}^1$, and $P \in \mathcal{M}_{\mathcal{X}\times\mathcal{Y}\times\mathcal{Z}}^1$ the joint probability of variables X and Y and Z. Then the following two statements are equaivalent:*

1. $\mathbb{I}_S[X \leftarrow Y|Z = z] = 0$
2. *$X \perp\!\!\!\perp Y|Z = z$; the variables X and Y are conditionally independent given $Z = z$*

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 rules, 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 [Gretton et al., 2012; Sriperumbudur et al., 2008].

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:

$$S_{log}(x, P) = -\log P(x) \tag{2.22}$$

This score is widely used, most notably in maximum likelihood estimation of parametric models. Maximum likelihood estimation is a special case of score matching as defined in Definition 4:

$$\hat{\theta}_{ML} = \underset{\theta}{\operatorname{argmax}} \sum_{n=1}^{N} \log P(x_i|\theta) \tag{2.23}$$

The associated entropy function is Shannon's entropy, also known as differential entropy for continuous distributions.

$$\mathbb{H}_{Shannon}[P] = -\mathbb{E}_{x\sim P} \log P(x) \tag{2.24}$$

The divergence function is the Kullback-Leibler (KL) divergence, which is very widely used in approximate Bayesian inference, model selection and active learning.

$$d_{KL}[P\|Q] = \mathbb{E}_{x\sim P} \frac{\log P(x)}{\log Q(x)} \tag{2.25}$$

The KL divergence is only well-defined when the distribution Q is absolutely continuous with respect to P. This is a serious limitation 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 express the KL divergence

between, say, an empirical distribution of samples and a continuous distribution, as we did in Definition 4.

A related problem is that Shannon's entropy of atomic distributions or mixed atomic and continuous distributions is either not well defined or depends only on the relative weight of the atoms but ot on their locations. As we will see, information quantities based on other scoring rules remain well defined for wider classes of distributions including atomic ones.

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 scoring rule that is local 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 [Dawid et al., 2012; Parry et al., 2012].

The value of information becomes Shannon's mutual information, a crucial quantity in communication and channel coding [MacKay, 2002; Shannon, 1948]. Shannon's mutual information has several equivalent definitions. Interestingly, it can be rewritten as the KL divergence between the joint distribution $P_{X,Y}$ and the product of its marginals $P_X P_Y$:

$$\begin{aligned}
\mathbb{I}_{Shannon}\left[X \leftarrow Y\right] &= \mathbb{H}_{Shannon}\left[X\right] - \mathbb{E}_{y\sim P_Y}\mathbb{H}_{Shannon}\left[P_{X|Y=y}\right] && (2.26)\\
&= \mathbb{E}_{y\sim P_Y} d_{KL}\left[P_{X|Y=y}\middle\| P_X\right] && (2.27)\\
&= \mathbb{E}_{y\sim P_Y}\left[\mathbb{E}_{x\sim P_{X|Y=y}} \log \frac{P_{X|Y=y}(x)}{P_X(x)}\right] && (2.28)\\
&= \mathbb{E}_{(x,y)\sim P} \log \frac{P(x,y)}{P_X(x)P_Y(y)} && (2.29)\\
&= d_{KL}\left[P(x,y)\|P_X(x)P_Y(y)\right] && (2.30)
\end{aligned}$$

As a consequence, Shannon's information is symmetric. Recall, that the value of information is generally non-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 (2.30), by replacing the KL divergence with a more general

divergence d:

$$\mathbb{J}_d(X,Y) = \mathrm{d}\left[P(x,y) \| P_X(x)P_Y(y)\right] \tag{2.31}$$

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 of parameter estimation in 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.

$$S_{\text{pseudo}}(x,P) = -\sum_{d=1}^{D} \log P(x_d | x_{\neg d}), \tag{2.32}$$

where x_d denotes the d^{th} component of the vector x and$x_{\neg d}$ 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

$$p(x_d | x_{\neg d}) = \frac{P(x)}{\int P(X_d = y, x_{\neg d}) dy} = \frac{C \cdot P(x)}{\int C \cdot P(X_d = y, x_{\neg d}) dy} \tag{2.33}$$

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

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

$$P(x) = \frac{1}{Z} \exp(x^T W x + b^T x), x \in \{0,1\}^D, \tag{2.34}$$

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:

$$
\begin{aligned}
P(x_d | x_{\neg d}, W, b) &= \frac{p(x)}{\int p(x_d = y, x_{\neg d}) dy} && (2.35)\\
&= \frac{\frac{1}{Z} \exp(x^T W x + b^T x)}{\sum_{x_d \in \{0,1\}} \frac{1}{Z} \exp(x^T W x + b^T x)} && (2.36)\\
&= \frac{\exp(x^T W x + b^T x)}{\sum_{x_d \in \{0,1\}} \exp(x^T W x + b^T x)} && (2.37)\\
&= \frac{\exp\left(x_d \left(W_{d,d} + 2 W_{d,\neg d}^T x_{\neg d} + b_d\right)\right)}{\exp(W_{d,d} + 2 W_{d,\neg d}^T x_{\neg d} + b_d) + 1} && (2.38)\\
& && (2.39)
\end{aligned}
$$

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 pseudolikelihood estimation – score matching with the pseudolikelihood score – is consistent for fully visible Boltzmann machines. Besag [1977]; Comets [1992] showed similar results for Markov random fields.

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:

$$
S_{\log}(x, p) = -\log P(x) = -\sum_{d=1}^{D} \log P(x_d | x_{1:d-1}) \tag{2.40}
$$

Here the d^{th} term is a probability conditioned on $d-1$ variables. Computing the d^{th} term therefore would require $D - d$ dimensional integral in the general case. The pseudo-likelihood makes computations more efficient by conditioning on more variables than needed by the chain rule, therefore requiring lower dimensional integrals.

Csiszár and Talata [2004] showed that pseudolikelihood score is strictly proper for strictly postitive distributions. Moreover, for always positive distributions the following generalisation of the pseudolikelihood is also a strictly proper scoring rule [Dawid,

Lauritzen, and Parry, 2012]:

$$S_{\text{DLP12}}(x, P) = -\sum_{d=1}^{D} S_d\left(x_d, P_{X_d|X_{\neg d}=x_{\neg d}}\right), \tag{2.41}$$

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 widely 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 (RMSE) 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 probability distributions, which we define as:

$$\|P\|_2 := \sqrt{\mathbb{E}_{x\sim P} P(x)} \tag{2.42}$$

The above definition, albeit slightly informal, is well defined and finite 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. Similarly, one can define the scalar product between two distributions as follows.

$$\langle P, Q\rangle := \sqrt{\mathbb{E}_{x\sim P} Q(x)} \tag{2.43}$$

Using these definitions we can define the Brier score as follows:

$$\begin{aligned} S_{Brier}(x, P) &= \|P - \delta_x\|_2^2 && (2.44)\\ &= \|P\|_2^2 - 2P(x) + 1 && (2.45)\\ &= \mathbb{E}_{x'\sim P} P(x') - 2P(x) + 1, && (2.46) \end{aligned}$$

where δ_x is the discrete or continuous Dirac measure concentrated at the observed point x.

The score gives rise to the following entropy function.

$$\mathbb{H}_{Brier}[P] = \mathbb{E}_{x\sim P}\left[\mathbb{E}_{x'\sim P}P(x') - 2P(x) + 1\right] \tag{2.47}$$
$$= 1 - \mathbb{E}_{x\sim P}P(x) \tag{2.48}$$
$$= 1 - \|P\|_2^2 \tag{2.49}$$

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 is always bounded from above.

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

$$d_{Brier}[P\|Q] = \mathbb{E}_{x\sim P}\left[\|Q\|_2^2 - 2Q(x) + 1\right] - \mathbb{H}_{Brier}[P] \tag{2.50}$$
$$= \|Q\|_2^2 - 2\mathbb{E}_{x\sim P}Q(x) + \|P\|_2^2 \tag{2.51}$$
$$= \|Q\|_2^2 - 2\langle P, Q\rangle + \|P\|_2^2 \tag{2.52}$$
$$= \|P - Q\|_2^2 \tag{2.53}$$

Interestingly, the Brier divergence is symmetric, and it is analogous to the squared Euclidean distance.

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

$$\mathbb{I}_{Brier}[X \leftarrow Y] = \mathbb{E}_{y\sim P_Y}\|P_X - P_{X|Y=y}\|_2^2 \tag{2.54}$$
$$\tag{2.55}$$

2.2.4 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:

$$S_{spherical}(x, P) = 1 - \frac{P(x)}{\|P\|_2} \tag{2.56}$$

This gives rise to the following entropy and divergence functionals.

$$\begin{aligned}
\mathbb{H}_{spherical}[P] &= 1 - \mathbb{E}_{x\sim P}\frac{P(x)}{\|P\|_2} && (2.57)\\
&= 1 - \|P\|_2 && (2.58)\\
d_{spherical}[P\|Q] &= -\mathbb{E}_{x\sim P}\frac{Q(x)}{\|Q\|_2} + \|P\|_2 && (2.59)\\
&= \|P\|_2 - \frac{\langle Q, P\rangle}{\|Q\|_2} && (2.60)\\
&= \|P\|_2 \left(1 - \cos(P, Q)\right), && (2.61)
\end{aligned}$$

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

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

The value of information under the spherical score is

$$\mathbb{I}_{spherical}[X \leftarrow Y] = \|P_X\|_2 \mathbb{E}_{y\sim P_Y}\left(1 - \cos(P_X, P_{X|Y=y})\right) \tag{2.62}$$

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:

$$\mathbb{H}_{\gamma,pseudospherical}[P] = -\|P\|_\gamma \tag{2.63}$$

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,

2007; Gneiting and Raftery, 2007; ?].

Independently, 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:

$$d_{\mathcal{F}}\left[P\|Q\right] = \sup_{f\in\mathcal{F}} |\mathbb{E}_{x\sim P} f(x) - \mathbb{E}_{x\sim Q} f(x)| \tag{2.64}$$

Intuitively, if two distributions are close in the integral probability metric sense, then no matter which function f we choose from the function class $\mathcal{F}$, the difference between the expectation of f under P and Q should be small. This class of divergences include the Wasserstein distance [del Barrio et al., 1999], the 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 [Sriperumbudur et al., 2008]:

$$\begin{aligned}
\mathrm{MMD}^2(P,Q) = & \sup_{\substack{f\in\mathcal{H}\\ \|f\|_{\mathcal{H}}=1}} \left(\mathbb{E}_{x\sim P} f(x) - \mathbb{E}_{x\sim Q} f(x)\right)^2 && (2.65)\\
= & \sup_{\substack{f\in\mathcal{H}\\ \|f\|_{\mathcal{H}}=1}} \left|\mathbb{E}_{x\sim P} \langle f, k(\cdot,x)\rangle - \mathbb{E}_{x\sim Q} \langle f, k(\cdot,x)\rangle\right|^2 && (2.66)\\
= & \sup_{\substack{f\in\mathcal{H}\\ \|f\|_{\mathcal{H}}=1}} \left|\langle f, \mathbb{E}_{x\sim P} k(\cdot,x) - \mathbb{E}_{x\sim Q} k(\cdot,x)\rangle\right|^2 && (2.67)\\
= & \sup_{\substack{f\in\mathcal{H}\\ \|f\|_{\mathcal{H}}=1}} \langle f, \mu_P - \mu_Q\rangle^2 && (2.68)\\
= & \|\mu_P - \mu_Q\|^2_{\mathcal{H}} && (2.69)\\
= & \mathbb{E}_{x,x'\sim P} k(x,x') - 2\mathbb{E}_{x\sim P}\mathbb{E}_{x'\sim Q} k(x,x') + \mathbb{E}_{x,x'\sim Q} k(x,x'), && (2.70)
\end{aligned}$$

In the derivation above we exploited the reproducing property of the kernel to arrive at (2.66) and the linearity of expectation to obtain (2.67). Step (2.69) holds because of the Cauchy-Schwartz inequality. $\mu_P(\cdot) = \mathbb{E}_{x\sim P} k(\cdot,x)$ is called the mean element or RKHS-embedding of the probability distribution P [Smola et al., 2007]. The MMD metric is analogous to the Euclidean distance between the mean elements of the two distributions.

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 $d_k[P\|Q] = 0 \iff Q = P$ holds. This is analogous to the strictly proper property of scoring rules and divergences as in Definition 3.

The mean embedding μ_P can be thought of as a generalisation of characteristic functions [see e. g. Ord et al., 1999]. The characteristic function of a probability distribution P with density p over the real line is defined as follows:

$$\phi_p(t) = \mathbb{E}_{x\sim p}\left[e^{itx}\right] = \int e^{itx} p(x)dx, \tag{2.71}$$

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 (2.70) 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 **??**). **TODO: If I do this, I need a technical introduction to kernels - as well as probability distributions** 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 applications were two-sample tests. In two-sample testing one is provided i. i. d. samples from two distributions, and one has to determine whether the two distributions are the same or not. Gretton et al. [2012] developed and analysed efficient empirical methods based on the MMD for this problem.

Herding [Welling, 2009] and its generalisation kernel herding [Chen et al., 2012] have been shown to minimise MMD between a target distribution and the empirical distribution of pseudo-samples. This method is an example of quasi-Monte Carlo methods that are examined in Chapter 4. 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. As we will see shortly, this is a special case of score matching as defined in Definition 4.

The squared MMD in fact conforms to our definition of a generalised divergence in equation (2.2), and corresponds to the following scoring rule:

$$
\begin{aligned}
S_k(x, P) &:= k(x,x) - 2\mathbb{E}_{x'\sim P} k(x,x') + \mathbb{E}_{x',x''\sim P} k(x',x'') && (2.72)\\
&= \sup_{\substack{f\in\mathcal{H}\\ \|f\|_{\mathcal{H}}=1}} \left(f(x) - \mathbb{E}_{x\sim P} f(y)\right)^2 && (2.73)
\end{aligned}
$$

The equivalence can be seen by applying Definition 2 of the generalised divergence:

$$
\begin{aligned}
d_k\left[P\|Q\right] &:= \mathbb{E}_{x\sim P} S_k(x,Q) - \mathbb{E}_{x\sim P} S_k(x,P) && (2.74)\\
&= \mathbb{E}_{x\sim P} k(x,x) - 2\mathbb{E}_{x\sim P}\mathbb{E}_{x'\sim Q} k(x,x') + \mathbb{E}_{x',x''\sim Q} k(x',x'') && (2.75)\\
&\quad - \mathbb{E}_{x\sim P} k(x,x) + 2\mathbb{E}_{x,x'\sim P} k(x,x') - \mathbb{E}_{x',x''\sim P} k(x',x'') && (2.76)\\
&= \mathbb{E}_{x',x''\sim P} k(x',x'') - 2\mathbb{E}_{x\sim P}\mathbb{E}_{x'\sim Q} k(x,x') + \mathbb{E}_{x',x''\sim Q} k(x',x'') && (2.77)\\
&= \mathrm{MMD}^2\left(P,Q\right) && (2.78)
\end{aligned}
$$

This scoring rule in Equation (2.72) is equivalent to the *kernel scoring rule* introduced originally by Eaton [1982]. The term kernel score was later coined by Dawid [2007]. Further references to this scoring rule can be found in [Eaton et al., 1996; Gneiting and Raftery, 2007]. The original definitions differed from the formula by a factor of two, and they did not have the leading $k(x, x)$ term. These differences do not make any 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 [Dawid, 2007]. Also, the statistics community defined the scores in terms of negative definite kernels, rather than positive definite ones which is the common convention in machine learning. It has also been pointed out 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 [Dawid, 2007].

To my knowledge the connection between kernel scores in statistics and maximum mean discrepancy has not been established before. 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.

Depending on the choice of kernel, the kernel score can be strictly proper. [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:

$$\begin{aligned} \mathbb{H}_k[P] &= \mathbb{E}_{x \sim P} S_k(x, P) && (2.79) \\ &= \mathbb{E}_{x \sim P} k(x, x) - 2\mathbb{E}_{x,x' \sim P} k(x, x') + \mathbb{E}_{x',x'' \sim P} k(x', x'') && (2.80) \\ &= \mathbb{E}_{x \sim P} k(x, x) - \mathbb{E}_{x,x' \sim P} k(x, x') && (2.81) \end{aligned}$$

This entropy function is concave for all positive definite kernels k and strictly concave whenever the kernel is characteristic. 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 entropy is also non-negative. Thus, in most practical cases the entropy functional is bounded both from above and below. Irrespective of kernel choice, the entropy is exactly 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 entropy can be computed for any distribution that one can compute expectations over. 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 as discussed in Chapter 4.

Thirdly, the entropy function has the kernel k as free parameter, which is 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 and corresponding divergence to our needs by adjusting parameters, such as the length-scale parameter [Song et al., 2008]. 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.

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

$$\begin{aligned}
\mathbb{I}_k\left[X \leftarrow Y\right] &= \mathbb{E}_{y\sim P_Y} d_k\left[P_X \middle\| P_{X|Y=y}\right] && (2.82)\\
&= \mathbb{E}_{y\sim P_Y} \|\mu_{X|Y=y} - \mu_X\|^2_{\mathcal{H}} && (2.83)\\
&= k(P_X, P_X) - 2 * \mathbb{E}_{y\sim P_Y} k(P_X, P_{X|Y=y}) + \mathbb{E}_{y\sim P_Y} k(P_{X|Y=y}, P_{X|Y=y}) && (2.84)\\
&= \mathbb{E}_{y\sim P_Y} \mathbb{E}_{P_{x_1,x_2\sim P_{X|Y=y}}} k(x_1, x_2) - \mathbb{E}_{x_1,x_2\sim P_X} k(x_1, x_2) && (2.85)
\end{aligned}$$

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 8 (kernel Cross-covariance operator)**.** Let X and Y be two random variables with joint distribution $P \in \mathcal{M}^1_{\mathcal{X}\times\mathcal{Y}}$, and marginals P_X and P_Y. Let $k_{\mathcal{X}} : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{C}$ and $k_{\mathcal{Y}} : \mathcal{Y} \times \mathcal{Y} \mapsto \mathbb{C}$ be positive definite kernels with associated reproducing kernel Hilbert

spaces $\mathcal{H}_\mathcal{X}$ and $\mathcal{H}_\mathcal{Y}$, respectively. Let us define the kernel cross-covariance operator C_{XY} between X and Y so that for all $f \in \mathcal{H}_\mathcal{X}$ and $g \in \mathcal{H}_\mathcal{Y}$

$$\langle f, C_{XY} g \rangle_{\mathcal{H}_\mathcal{X}} = \mathbb{E}_{(x,y)\sim P} \left(f(x) - \mathbb{E}_{x'\sim P_X} f(x') \right) \left(g(y) - \mathbb{E}_{y'\sim P_Y} g(y') \right) \tag{2.86}$$

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

Definition 9 (COCO, see [Gretton et al., 2005b])**.** In the same notation as above let us define define the constrained covariance between X and Y, $COCO_{XY}$, as

$$COCO_{XY} = \sup_{\substack{f\in\mathcal{H}_\mathcal{X}, g\in\mathcal{H}_\mathcal{Y} \\ \|f\|_{\mathcal{H}_\mathcal{X}}=1, \|g\|_{\mathcal{H}_\mathcal{Y}}=1}} \mathbb{C}\mathrm{ov}_{(x,y)\sim P} \left[f(x), g(y) \right] \tag{2.87}$$

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

$$COCO_{XY} = \|C_{XY}\|_2, \tag{2.88}$$

where $\|\cdot\|_2$ denotes the spectral norm, that is the modulus of largest singular value [Gretton et al., 2005b]

A more robust measure of dependence, the Hilbert Schmidt Information Criterion (HSIC) uses the Hilbert-Schmidt norm of the cross-covatiance operator[Gretton et al., 2005a]. Kernel measures of dependence like COCO and HSIC have several useful properties. They are symmetric, and can be effectively estimated from empirical data [Gretton et al., 2005a].

However, as with generalisations of Shannon's information in Eqn. (**??**), 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 (2.83) 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 10 (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 $f, g \in \mathcal{H}$

$$\left\langle f, D_{X|Y} g \right\rangle_{\mathcal{H}} = \mathbb{C}\mathrm{ov}_{y \sim P_Y} \left[\mathbb{E}_{X|Y=y} f, \mathbb{E}_{X|Y=y} g \right] \tag{2.89}$$

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

$$\left\langle f, D_{X|Y} f \right\rangle_{\mathcal{H}} = \mathbb{V}_{y \sim P_Y} \left[\mathbb{E}_{x \sim P_{X|Y=y}} f(x) \right] \tag{2.90}$$

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

Statement 3 (Alternative definition of $D_{X|Y}$)**.** *$D_{X|Y}$ admits the following equivalent definition*

$$D_{X|Y} = \mathbb{E}_{y \sim P_Y} \left(\mu_{X|Y=y} - \mu_X \right) \otimes \left(\mu_{X|Y=y} - \mu_X \right) \tag{2.91}$$

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

$$\begin{aligned}
&\left\langle f, \left(\mathbb{E}_{y \sim P_Y} \left(\mu_{X|Y=y} - \mu_X \right) \otimes \left(\mu_{X|Y=y} - \mu_X \right) \right) g \right\rangle && (2.92)\\
=&\mathbb{E}_{y \sim P_Y} \left\langle f, \left(\left(\mu_{X|Y=y} - \mu_X \right) \otimes \left(\mu_{X|Y=y} - \mu_X \right) \right) g \right\rangle && (2.93)\\
=&\left\langle f, \left(\mu_{X|Y=y} - \mu_X \right) \right\rangle \left\langle g, \left(\mu_{X|Y=y} - \mu_X \right) \right\rangle && (2.94)\\
=&\left\langle f, \left(\mathbb{E}_{x \sim P_{X|Y=y}} k(\cdot, x) - \mathbb{E}_{x \sim P_X} k(\cdot, x) \right) \right\rangle \left\langle g, \left(\mathbb{E}_{x \sim P_{X|Y=y}} k(\cdot, x) - \mathbb{E}_{x \sim P_X} k(\cdot, x) \right) \right\rangle && (2.95)\\
=&\mathbb{E}_{y \sim P_Y} \left(\mathbb{E}_{X|Y=y} f(x) - \mathbb{E}_{x \sim P_X} f(x) \right) \left(\mathbb{E}_{X|Y=y} g(x) - \mathbb{E}_{x \sim P_X} g(x) \right) && (2.96)\\
=&\mathbb{C}\mathrm{ov}_{y \sim P_Y} \left[\mathbb{E}_{X|Y=y} f, \mathbb{E}_{X|Y=y} g \right] && (2.97)\\
=&\left\langle f, D_{X|Y} g \right\rangle_{\mathcal{H}}, && (2.98)
\end{aligned}$$

where we used the linearity of expectation and the reproducing property of the kernel to obtain step (2.96) □

Using this alternative definition it is easy to see that the kernel value of information as defined in Eqn. (2.83) 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):

$$\mathbb{I}_k\left[X \leftarrow Y\right] = \mathbb{E}_{y\sim P_Y}\|\mu_X - \mu_{X|Y=y}\|_2^2 \tag{2.99}$$
$$= \mathbb{E}_{y\sim P_Y}\mathrm{tr}\left\langle \mu_X - \mu_{X|Y=y}, \mu_X - \mu_{X|Y=y}\right\rangle \tag{2.100}$$
$$= \mathbb{E}_{y\sim P_Y}\mathrm{tr}\left(\mu_X - \mu_{X|Y=y}\right)\otimes\left(\mu_X - \mu_{X|Y=y}\right) \tag{2.101}$$
$$= \mathrm{tr}\,\mathbb{E}_{y\sim P_Y}\left(\mu_X - \mu_{X|Y=y}\right)\otimes\left(\mu_X - \mu_{X|Y=y}\right) \tag{2.102}$$
$$= \mathrm{tr}\, I_{X|Y} \tag{2.103}$$
$$= \|I_{X|Y}^{1/2}\|_{HS} \tag{2.104}$$

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 useful 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 Intuitivelyitive divergence. Consider the following scoring rule

$$S_{k,spherical}(x,P) := \|\mu_{\delta_x}\|_{\mathcal{H}} - \frac{\mu_P(x)}{\|\mu_P\|_{\mathcal{H}}} \tag{2.105}$$
$$= \|\mu_{\delta_x}\|_{\mathcal{H}}\left(1 - \cos(\mu_{\delta_x}, \mu_P)\right) \tag{2.106}$$
$$= \sqrt{k(x,x)} - \frac{\mathbb{E}_{x'\sim P}k(x,x')}{\sqrt{\mathbb{E}_{x,x'\sim P}k(x,x')}}, \tag{2.107}$$

The scoring rule gives rise to the following entropy functional:

$$\mathbb{H}_{k,spherical}\left[P\right] = \mathbb{E}_{x\sim P}\|\mu_{\delta_x}\|_{\mathcal{H}} - \|\mu_P\|_{\mathcal{H}} \tag{2.108}$$
$$= \mathbb{E}_{x\sim P}\sqrt{k(x,x)} - \sqrt{\mathbb{E}_{x,x'\sim P}k(x,x')} \tag{2.109}$$

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

$$d_{k,spherical}\left[P\|Q\right] = -\mathbb{E}_{x\sim Q}\frac{\mu_P}{\|\mu_P\|_{\mathcal{H}}} + \|\mu_P\|_{\mathcal{H}} \tag{2.110}$$

$$= \|\mu_P\|_{\mathcal{H}}\left(1 - \cos(\mu_P, \mu_Q)\right) \tag{2.111}$$

$$= \sqrt{\mathbb{E}_{x,x'\sim P}k(x,x')} - \frac{\mathbb{E}_{x\sim P}\mathbb{E}_{x'\sim Q}k(x,x')}{\sqrt{\mathbb{E}_{x,x'\sim Q}k(x,x')}} \tag{2.112}$$

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

$$0 < d_k\left[P\|Q\right] \tag{2.113}$$

$$0 < \|\mu_P\|^2_{\mathcal{H}} + \|\mu_Q\|^2_{\mathcal{H}} - 2\left\langle\mu_P, \mu_Q\right\rangle_{\mathcal{H}} \tag{2.114}$$

$$\left\langle\mu_P, \mu_Q\right\rangle < \frac{1}{2}\left(\|\mu_P\|^2_{\mathcal{H}} + \|\mu_Q\|^2_{\mathcal{H}}\right) \leq \|\mu_P\|_{\mathcal{H}}\|\mu_Q\|_{\mathcal{H}} \tag{2.115}$$

$$\cos(\mu_P, \mu_Q) = \frac{\left\langle\mu_P, \mu_Q\right\rangle_{\mathcal{H}}}{\|\mu_P\|_{\mathcal{H}}\|\mu_Q\|_{\mathcal{H}}} < 1 \tag{2.116}$$

Thus,

$$d_{k,spherical}\left[P\|Q\right] = \|\mu_P\|_{\mathcal{H}}\left(1 - \cos(\mu_P, \mu_Q)\right) > 0 \tag{2.117}$$

□

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:

$$d_{k,spherical}\left[P\|Q\right] = \|\mu_P\|_{\mathcal{H}}\mathbb{P}_{f\sim GP}\left[sign(\mathbb{E}_{x\sim P}f(x)) \neq sign(\mathbb{E}_{x\sim Q}f(x))\right] \tag{2.118}$$

Proof. See Lemma 8 in [Goemans and Williamson, 1995]. □

Thus, the divergence function (2.112) 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. The finite dimensional analogue of Proposition 1 is exploited in sign-random-projection locality sensitive hashing (SRP-LSH) algorithms [Charikar, 2002; Ji et al., 2012]. Similarly, Eqn. (2.118) can be exploited in locality sensitive hashing algorithms for probability distributions, even though practical applications of such algorithms are probably limited.

I am not aware of any previous definition or mention of the spherical kernel scoring rule or the associated divergence functional 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:

$$a^*_P = \operatorname*{argmin}_{a\in\mathcal{A}} \mathbb{E}_{x\sim P}\ell(x, a) \tag{2.119}$$

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 $X = x$.

$$S_\ell(x, P) = \ell(x, a_P^*) \tag{2.120}$$

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:

$$\begin{aligned}
\mathbb{H}_\ell[P] :=& \mathbb{E}_{x\sim P}\ell(x, a_P^*) && (2.121)\\
=& \min_{a\in\mathcal{A}} \mathbb{E}_{x\sim P}\ell(x, a) && (2.122)\\
=& \mathcal{R}_\ell(P) && (2.123)
\end{aligned}$$

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

$$d_\ell[P\|Q] = \mathbb{E}_{x\sim P}\ell(x, a_Q^*) - \min_{a\in\mathcal{A}} \mathbb{E}_{x\sim P}\ell(x, a) \tag{2.124}$$

Because of the definition of $\mathbb{H}_\ell[P]$, the divergence is always non-negative, hence the scoring rule defined this way is always proper. In fact, proper scoring rules and Bayesian decision problems are equivalent, inasmuch as every proper scoring rule can be expressed as Bayesian decision problem as in Eqn. (2.120). Throughout this thesis I will use the decision theoretic notation (ℓ, $\mathcal{A}$, $\mathcal{R}_\ell[\cdot]$) or the scoring rule notation (S, $\mathbb{H}_S[\cdot]$, $d_S[\cdot\|\cdot]$) interchangeably, depending on which one is more natural given the context.

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 Shannon entropy of P:

$$\mathbb{E}_{x\sim P}|f(x)| \geq \frac{1}{\log(2)}\mathbb{H}_{Shannon}\left[P\right], \tag{2.125}$$

where the $1/\log(2)$ is not needed if use base-2 logarithm in the definition of $\mathbb{H}_{Shannon}\left[P\right]$.

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

Assume our task is to estimate value of a set of known functions $f \in \mathcal{F}$ all at the same random point 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 $f \in \mathcal{F}$. We are required to do equally well on all functions, and the loss ℓ we incur is equal to the largest squared error we incur on any of these functions.

$$\ell(x,a) = \sup_{f\in\mathcal{F}} \left(f(x) - a(f)\right)^2 \tag{2.126}$$

Given a probabilistic forecast P over X, the Bayes optimal decision $a^*_P(f)$ is to computes the mean of f under the forecast distribution P:

$$a^*_P(f) = \mathbb{E}_{x\sim P} f(x) \tag{2.127}$$

Thus, we can define the following scoring rule S:

$$S(x,P) = \ell(x,a^*_P) = \sup_{f\in\mathcal{F}} \left(f(x) - \mathbb{E}_{x\sim P} f(x)\right)^2 \tag{2.128}$$

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 (Eqn. (2.72)).

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 proper 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: the kernel scoring rule and maximum mean disrepancy. These rich classes of scoring rules have been used by the machine learning community, where they were derived from different first principles without noting the connection to Bregman divergences.

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 by visualising the Riemannian geometry they induce over probability distributions.

Chapter 3

Information geometry

3.1 Introduction

Strictly proper scoring rules and associated Bregman divergences determine a unique *information geometry* of probability distributions. Different scoring rules are sensitive to different properties of distributions. Consider measuring divergence between Normal distributions, some divergences are more sensitive to small changes in the variance, while others are agnostic to the variance but sensitive to changes in the mean. In this section I aim to develop an understanding of these differences by visualising the geometric structures common scoring rules 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, my 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 one or 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 like 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 2.

3.2 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 4 (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:

$$\frac{\partial}{\partial q_i} d_H[P\|Q] = -\frac{\partial}{\partial q_i} H(Q) + \frac{\partial}{\partial q_i}\langle \nabla_Q H(Q), Q-P\rangle \tag{3.4}$$

$$= -\frac{\partial}{\partial q_i} H(Q) + \frac{\partial}{\partial q_i}\sum_j \frac{\partial}{\partial q_j} H(Q)(q_j - p_j) \tag{3.5}$$

$$= -\frac{\partial}{\partial q_i} H(Q) + \frac{\partial}{\partial q_i} H(Q) + \sum_j \frac{\partial^2}{\partial q_i \partial q_j} H(Q)(q_j - p_j) \tag{3.6}$$

$$= \sum_j \frac{\partial^2}{\partial q_i \partial q_j} H(Q)(q_j - p_j) \tag{3.7}$$

hence by first order Taylor expansion around P:

$$d_H[P\|P+dP] \approx d_H[P\|P+dP] + \frac{1}{2}\left\langle dP, \nabla_Q d_H[P\|Q]|_{Q=P}\right\rangle \tag{3.8}$$

$$= \frac{1}{2}\sum_i dp_i \sum_j \frac{\partial^2 H(P)}{\partial p_i \partial p_j} dp_j \tag{3.9}$$

$$= \frac{1}{2}\sum_{i,j}\frac{\partial^2 H(P)}{\partial p_i \partial p_j} dp_i dp_j \tag{3.10}$$

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 square-root 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 11 (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.

3.3 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.

3.3.1 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.

3.3.2 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:

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

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.

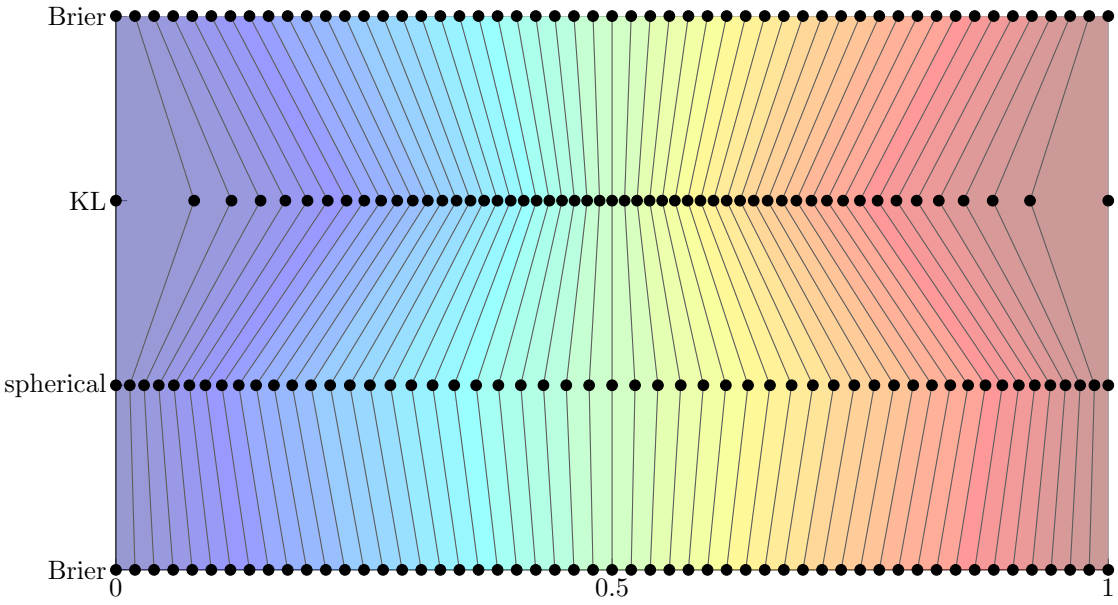

Figure 3.1: Illustration of the differences between the Brier, spherical, and KL divergences between single parameter Bernoulli distributions. Each horizontal line of dots shows the embedding Bernoulli distributions corresponding to an uniform grid of parameter values between $0 + \epsilon$ and $1 - \epsilon$ 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.

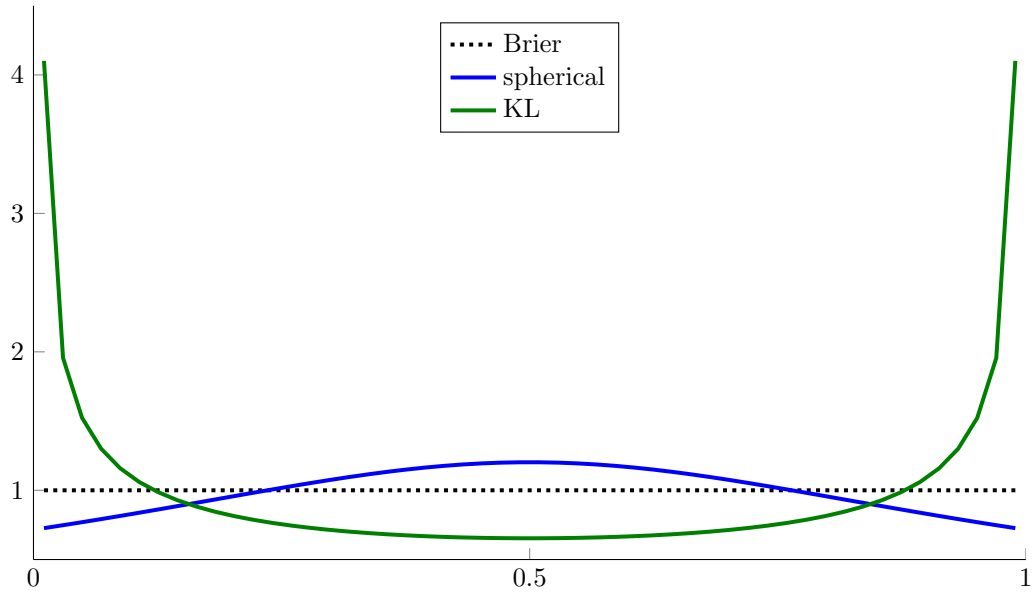

Figure 3.2: Illustration of the differences between local distances on statistical manifolds of Bernoulli distributions. Each line shows the magnitude of the local distance on the manifold relative to the Euclidean distance as a function of the parameter value. Distance on the Brier manifold is equivalent 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 **??**.

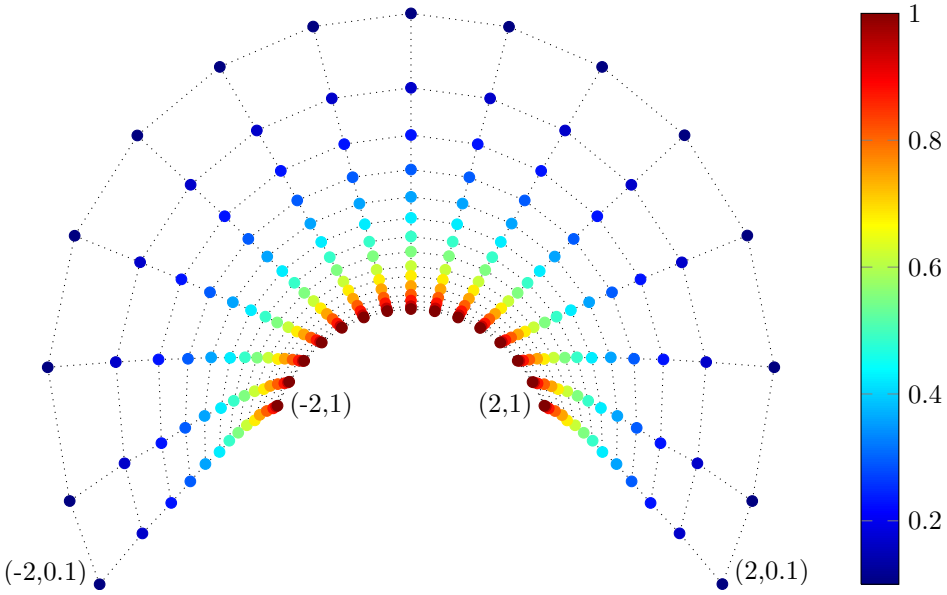

Figure 3.3: Map of Normal distributions on the statistical manifold induced by the logarithmic score and KL divergence. Distributions are chosen from a uniform grid in parameter space, with mean ranging between -2 and 2 *(left to right)*, and standard deviation between 0.1 and 1 *(from outside inwards)*. 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.

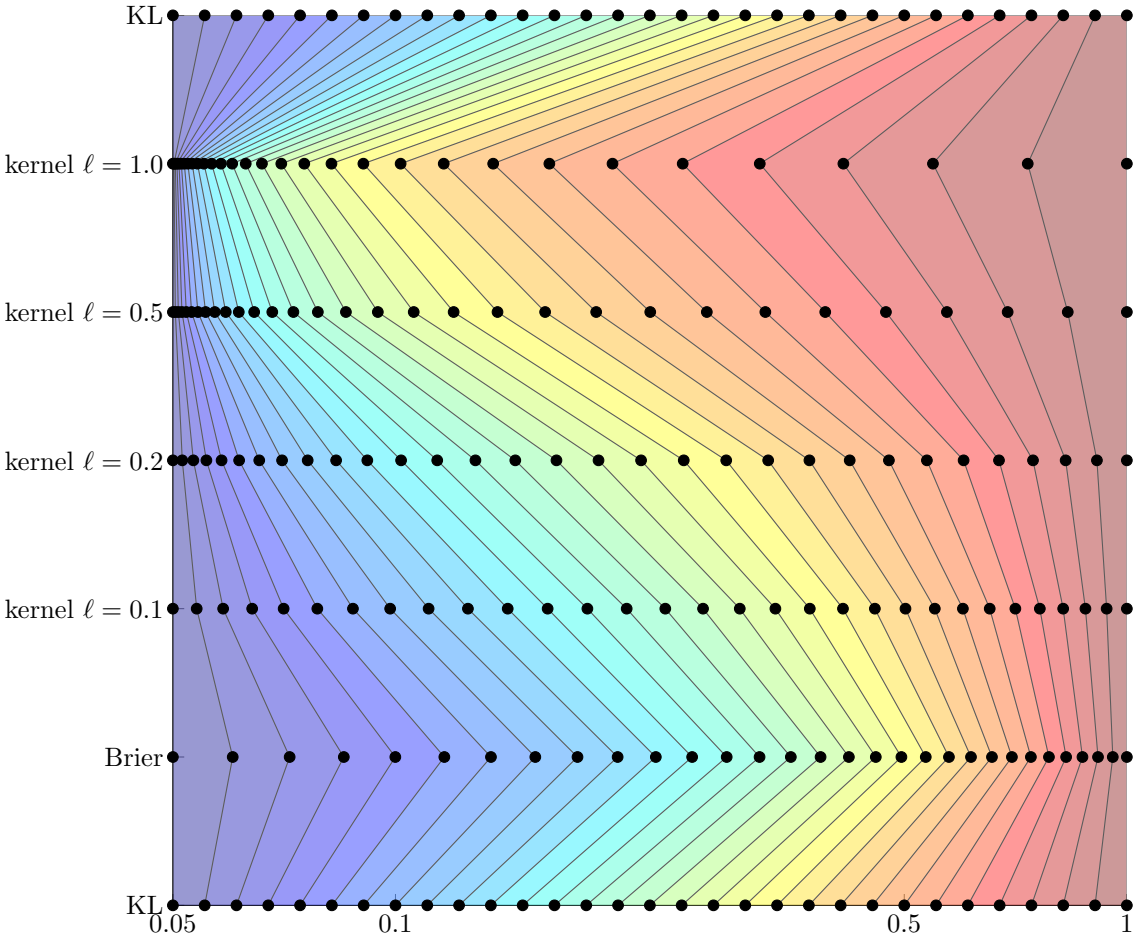

Figure 3.4: Illustration of the differences between the statistical manifolds of Normal distributions induced by the KL, Brier and kernel divergences. Each horizontal line of points shows the one-dimensional manifold of zero-mean Gaussians. The dots correspond to distributions with logarithmically spaced 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.

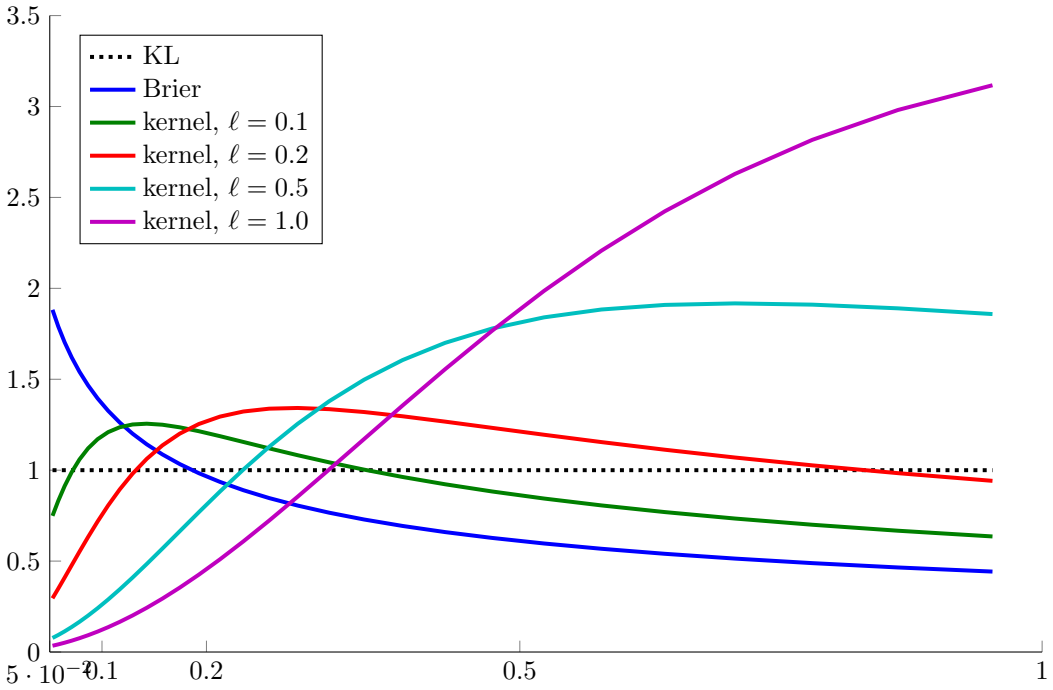

Figure 3.5: Illustration of the differences between local distances induced by various scoring rules on the statistical manifold of zero-mean Normal distributions. Each line shows the magnitude of the local distance on each manifold relative to that induced by the KL divergence as a function of variance. 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.

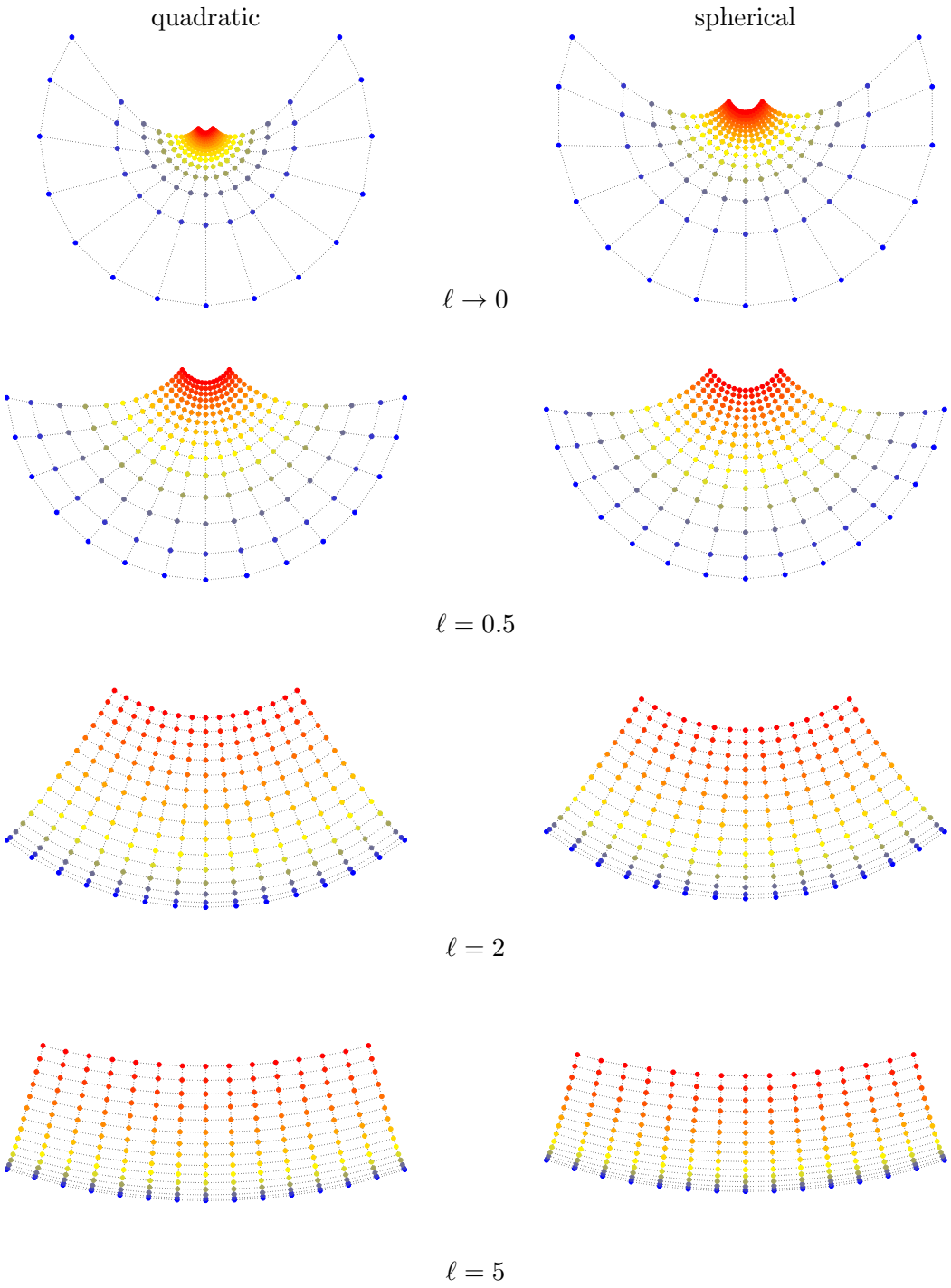

Figure 3.6: Maps of the statistical manifold induced by the (quadratic) kernel score and the spherical kernel score over Gaussian distributions for different setting of the kernel bandwidth parameter. The two panels in the top row $\ell \to 0$ correspond to the limiting cases of the Brier score and the 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.

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:

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

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

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 2 this flexibility also poses a challenge in applications where there is no principled way of choosing kernel hyperparameters.

3.3.3 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 probability density function of Gamma distributions is as follows:

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

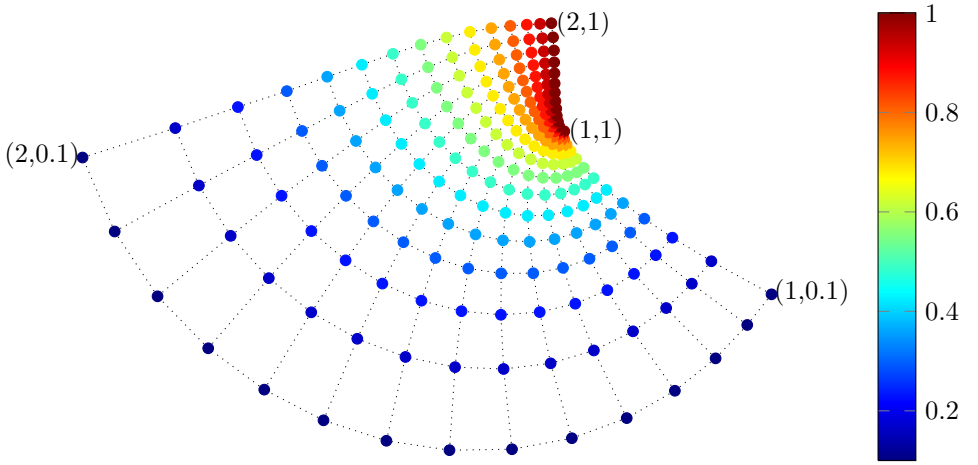

Figure 3.7: Map of Gamma distributions on the statistical manifold induced by the logarithmic score and KL divergence. To be comparable to the manifold of Normal distributions in Figure 3.3, the distributions are parametrised by their mean and standard deviation. Distributions are chosen from a uniform 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.

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.

Part II

Approximate Bayesian inference

Chapter 4

Loss calibrated approximate inference

Summary of contributions: The loss-calibrated approximate inference framework presented in this chapter is joint work with Simon Lacoste-Julien and Zoubin Ghahramani, and has been published as part of [Lacoste-Julien et al., 2011]. FH and SLJ contributed equally to the development of the framework. The loss-calibrated quasi-Monte Carlo framework and the interpretation of kernel herding in this framework is original contribution by FH. The equivalence of optimally weighted kernel herding and Bayesian quadrature is joint work with David Duvenaud and has been published [Huszár and Duvenaud, 2012]. FH and DD contributed equally to designing research and interpreting results. The theoretical analysis of approximate submodularity is original contribution by FH. The method was implemented and experiments were carried out by DD. Some figures are taken from [Huszár and Duvenaud, 2012] with the permission of the co-author.

4.1 Introduction

Bayesian methods model observed data $\mathcal{D}$ by introducing latent parameters $\boldsymbol{\theta}$, and inferring their value via Bayes' rule

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}} \tag{4.1}$$

The likelihood $p(\mathcal{D}|\boldsymbol{\theta})$ describes how data is related to the parameters $\boldsymbol{\theta}$, and $p(\boldsymbol{\theta})$ is a prior distribution which capture's one a priori expectations about what the value of $\boldsymbol{\theta}$ may be. The posterior distribution $p_{\mathcal{D}} = p(\boldsymbol{\theta}|\mathcal{D})$ captures all statistically relevant

information that the data $\mathcal{D}$ provides about $\boldsymbol{\theta}$, and it is therefore of central importance. The marginal likelihood, also called the model evidence $Z = \int p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}$, is often used to quantify how well a Bayesian model – the combination of likelihood and prior – fit the data.

In practically interesting Bayesian models, the posterior distribution and model evidence are often computationally intractable to obtain and therefore one has to resort to approximations. The most popular methods for Bayesian approximate inference are variational inference and Markov chain Monte Carlo.

Variational inference replaces the posterior by a computationally convenient distribution q. It operates by minimising an objective function that expresses divergence between the true posterior $p_{\mathcal{D}}$ and the approximation q. The divergence is often chosen to be a variant of Kullback-Leibler divergence (Eqn. (**??**)), as it allows easy rearrangement of terms and makes local message-passing style computations possible.

In section **??** I 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. In [Lacoste-Julien et al., 2011] we devised a framework we termed loss-calibrated approximate inference, which is related to traditional variational approaches and is based on minimising scoring-rule-based Bregman divergences. In this chapter I will illustrate on a simple example how loss-calibrated approximate works.

Monte Carlo methods produce random samples (approximately) drawn from the posterior, which in turn allow for approximating relevant integrals and making predictinos. Monte Carlo techniques have been successfully applied to a wide range of Bayesian inference problems. However, just as most variational approaches, Monte Carlo techniques are 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 that kernel herding, a recent heuristic algorithm proposed by Chen et al. [2012] can be considered a special case of loss-calibrated quasi-Monte Carlo, and point out connections between this method and Bayesian Quadrature.

4.2 The goals of approximate inference

In many practically relevant cases computing the Bayesian posterior is not analytically tractable. This is predominantly due to the fact that the integral defining the marginal

likelihood $\int p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}$ cannot be computed analytically, and therefore the posterior is only known up to a multiplicative constant. It is usual practice to approximate the intractable posterior by something simpler, a computationally convenient approximate distribution q. The problem of finding an approximate posterior q is referred to as approximate Bayesian 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, mean-field estimation, 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 the objective functions they minimise, which measure discrepancy between the target posterior distribution $p_{\mathcal{D}}$ and the approximation q.

4.2.1 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.

$$\log p(\mathcal{D}) = \log \int p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta} \tag{4.2}$$

$$= \log \int \frac{p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{q(\boldsymbol{\theta})} q(\boldsymbol{\theta})d\boldsymbol{\theta} \tag{4.3}$$

$$\geq \int \log \frac{p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{q(\boldsymbol{\theta})} q(\boldsymbol{\theta})d\boldsymbol{\theta} \tag{4.4}$$

$$= \log p(\mathcal{D}) - d_{KL}\left[q \| p_{\mathcal{D}}\right] \tag{4.5}$$

Maximising the lower bound amounts to minimising the Kullback-Leibler divergence $d_{KL}\left[q\|p_{\mathcal{D}}\right]$ (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 $\mathcal{Q}$. In addition, q is often such that it factorises over multivariate quantities. When these assumptions are made, the optimal solution maximising the lower bound in Eqn. (4.5) can usually be expressed in closed form. If analytical solution is not available, efficient iterative algorithms based on message passing exist for finding a locally optimal solution numerically [Winn and Bishop, 2006].

The KL divergence is non-symmetric, therefore the order of its arguments matter.

Variational methods minimise $d_{KL}[q\|p_{\mathcal{D}}]$, that is with the approximate distribution being the first argument. Computationally, 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 2, in the scoring rule interpretation suggests that the *right*, theoretically well motivated way to use the divergence would be $d_{KL}[p_{\mathcal{D}}\|q]$. This has been pointed out previously in [Csató and Opper, 2002; Minka, 2001b] and by several other authors. This does not mean that variational inference does not work, it just means that when performing variational inference, we loose the convenient 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 technically challenging, as computing the divergence $d_S[p_{\mathcal{D}}\|q]$ always involves computing an integral over the posterior $p_{\mathcal{D}}$, which is normally intractable.

Assumed density filtering, and its generalisation, expectation propagation (EP) try to approximate the ideal method of minimising $d_{KL}[p_{\mathcal{D}}\|q]$ in the following way [Minka, 2001a]. EP assumes the posterior can be written as a product of factors as such:

$$p_{\mathcal{D}}(\boldsymbol{\theta}) = \frac{1}{Z}\prod t_i(\boldsymbol{\theta}) \tag{4.6}$$

The terms t_i are assumed simple, and in most cases depend only on a few components of the multivariate parameter vector $\boldsymbol{\theta}$. 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

$$q(\boldsymbol{\theta}) = \prod \tilde{t}_i(\boldsymbol{\theta}) \tag{4.7}$$

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

$$\tilde{t}_i^{new} = \underset{t\in\mathbb{Q}}{\operatorname{argmin}}\, d_{KL}\left[\underbrace{\frac{1}{\int q(\boldsymbol{\theta})\frac{t_i(\boldsymbol{\theta})}{\tilde{t}_i(\boldsymbol{\theta})}d\boldsymbol{\theta}}q(\boldsymbol{\theta})\frac{t_i(\boldsymbol{\theta})}{\tilde{t}_i(\boldsymbol{\theta})}}_{\tilde{q}}\middle\| q(\boldsymbol{\theta})\frac{t(\boldsymbol{\theta})}{t_i(\boldsymbol{\theta})}\right] \tag{4.8}$$

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, instead of minimising the divergence from the true posterior, EP uses the moving target $\tilde{q}$. Expectation propagation is known to perform well in a variety of graphical models, most famously in Gaussian process classification [Nickisch and Rasmussen, 2008]. It exploits convenient analytic properties of the logarithmic loss and the KL divergence, which make computing the expressions in Eqn. (4.8) possible. Therefore the method does not readily generalise to other scoring rules or divergences.

4.2.2 Loss-calibrated approximate inference

Although often overlooked, the main theoretical motivations for the Bayesian paradigm are rooted in Bayesian decision theory [Berger, 1985], which provides a well-defined theoretical framework for rational decision making under uncertainty about a hidden parameter $\boldsymbol{\theta}$. 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 [Robert, 2001] or Ch. 1 of [Berger, 1985] for example):

- a loss $\ell(\boldsymbol{\theta}, a)$ which quantifies the cost of taking action $a \in \mathcal{A}$ when the world state is $\boldsymbol{\theta} \in \Theta$;
- an observation model $p(\mathcal{D}|\boldsymbol{\theta})$ which gives the probability of observing some data or dataset $\mathcal{D} \in \mathcal{O}$ assuming that the world state is $\boldsymbol{\theta}$;
- a prior belief $p(\boldsymbol{\theta})$ 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* [Schervish, 1995])

$$\mathcal{R}_{p_\mathcal{D}}(a) \doteq \int_\Theta \ell(\boldsymbol{\theta}, a)\, p(\boldsymbol{\theta}|\mathcal{D}) d\boldsymbol{\theta} \tag{4.9}$$

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 $\boldsymbol{\theta}$ 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 $L(\boldsymbol{\theta} > T_{\text{crit}}, \text{'on'})$. 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 $p_\mathcal{D}(\boldsymbol{\theta})$ (**??**, 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.

4.2.3 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*:

$$\mathcal{R}_q(h) \doteq \int_{\Theta} q(\boldsymbol{\theta}) L(\boldsymbol{\theta}, h) d\boldsymbol{\theta}, \tag{4.10}$$

obtaining a *q-optimal* action h_q:

$$h_q \doteq \operatorname{argmin} h \in \mathcal{H} \mathcal{R}_q(h). \tag{4.11}$$

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:

$$q_{\text{opt}} = \operatorname{argmin} q \in \mathcal{Q} \mathcal{R}_{p_{\mathcal{D}}}(h_q), \tag{4.12}$$

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:

$$d_L(p\|q) \doteq \mathcal{R}_p(h_q) - \mathcal{R}_p(h_p). \tag{4.13}$$

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 $L(\boldsymbol{\theta}, h) = -\log h(\boldsymbol{\theta})$. The q-risk $R_q(h)$ then becomes the cross-entropy $H(q, h) = -\int_\Theta q(\boldsymbol{\theta}) \log(h(\boldsymbol{\theta}))d\boldsymbol{\theta}$, 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.

4.3 Loss-calibrated quasi-Monte Carlo

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 [Cowles and Carlin, 1996]. 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

$$q(x; x_1, \ldots, x_N) = \frac{1}{N} \sum_{n=1}^{N} \delta(x - x_n), \tag{4.14}$$

or weighted empirical distributions

$$q(x; x_1, \ldots, x_N, w_1, \ldots, w_N) = \sum_{n=1}^{N} w_n \delta(x - x_n). \tag{4.15}$$

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:

$$\{x_1, \ldots, x_N\}_{n=1}^{N} = \underset{\{x_1,\ldots,x_N\}_{n=1}^{N}}{\operatorname{argmin}} \; d_\ell \left[p_\mathcal{D} \| q(x; x_1, \ldots, x_N)\right] \tag{4.16}$$

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 (4.16) 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.

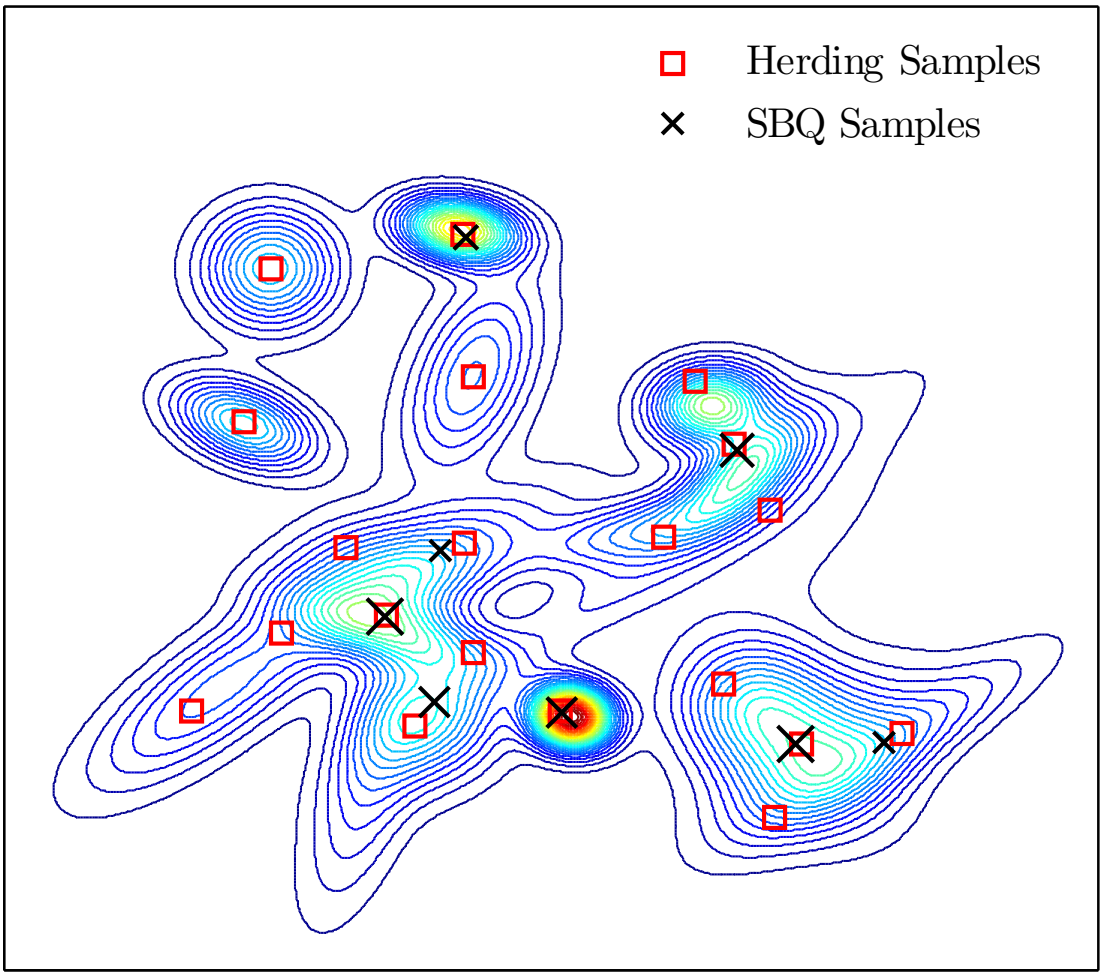

Figure 4.1: The first 8 samples from sequential Bayesian quadrature, versus the first 20 samples from herding. Only 8 weighted SBQ samples are needed to give an estimator with the same maximum mean discrepancy as using 20 herding samples with uniform weights. Relative sizes of samples indicate their relative weights.

4.4 Bayesian herding

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

$$Z_{f,p} = \int f(x)p(x)dx \tag{4.17}$$

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.

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 [Cowles and Carlin, 1996]. 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) [O'Hagan, 1991; Rasmussen and Ghahramani, 2003] (also known as Bayesian Monte Carlo) and kernel herding [Chen et al., 2010]. Bayesian quadrature estimates integral (4.17) 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 [Chen et al., 2010], 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 [Krause and Cevher, 2010], 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 $\mathcal{O}(1/N)$.

4.4.1 Herding

Herding was introduced by [Welling, 2009] 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 4.17.

For selecting pseudosamples, herding relies on an objective based on the maximum mean discrepancy [MMD; Sriperumbudur et al., 2008]. MMD measures the divergence between two distributions, p and q with respect to a class of integrand functions $\mathcal{F}$ as follows:

$$\div\mathcal{F}pq = \sup_{f\in\mathcal{F}} \left| \int f(x)p(x)dx - \int f(x)q(x)dx \right| \tag{4.18}$$

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 [Sriperumbudur et al., 2008]:

$$MMD^2_{\mathcal{H}}(p, q) = \sup_{\substack{f \in \mathcal{H} \\ \|f\|_{\mathcal{H}}=1}} \left| \int f_x p(x) dx - \int f_x q(x) dx \right|^2 \tag{4.19}$$

$$= \|\mu_p - \mu_q\|^2_{\mathcal{H}} \tag{4.20}$$

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

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

$$+ \iint k(x, y) q(x) q(y) dx dy, \tag{4.21}$$

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:

$$\epsilon_{herding}\left(\{x_1, \ldots, x_N\}\right) = \text{MMD}_{\mathcal{H}}\left(p, \frac{1}{N}\sum_{n=1}^{N} \delta_{x_n}\right) \tag{4.22}$$

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

$$-2\frac{1}{N}\sum_{n=1}^{N} \int k(x, x_n) p(x) dx + \frac{1}{N^2}\sum_{n,m=1}^{N} k(x_n, x_m) \tag{4.23}$$

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:

$$x_{n+1} \leftarrow \operatorname*{argmin}_{x \in \mathcal{X}} \epsilon_{herding}\left(\{x_1, \ldots, x_n, x\}\right) \tag{4.24}$$

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

assuming $k(x, x) = \text{const}$. The formula (4.24) 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 (4.24) 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 [Chen et al., 2010; 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 (4.24) 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 [Bach et al., 2012], this fast convergence is not guaranteed in infinite dimensional Hilbert spaces, such as the RKHS corresponding to the Gaussian kernel.

4.4.2 Bayesian quadrature

So far, we have only considered integration methods in which the integral (4.17) 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 [Rasmussen and Ghahramani, 2003], an alternate method is propositionosed: Bayesian Monte Carlo, or Bayesian quadrature (BQ). BQ puts a prior distribution on f, then estimates integral (4.17) 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 4.2 for an illustration of Bayesian Quadrature.

Here we derive the BQ estimate of (4.17), 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.

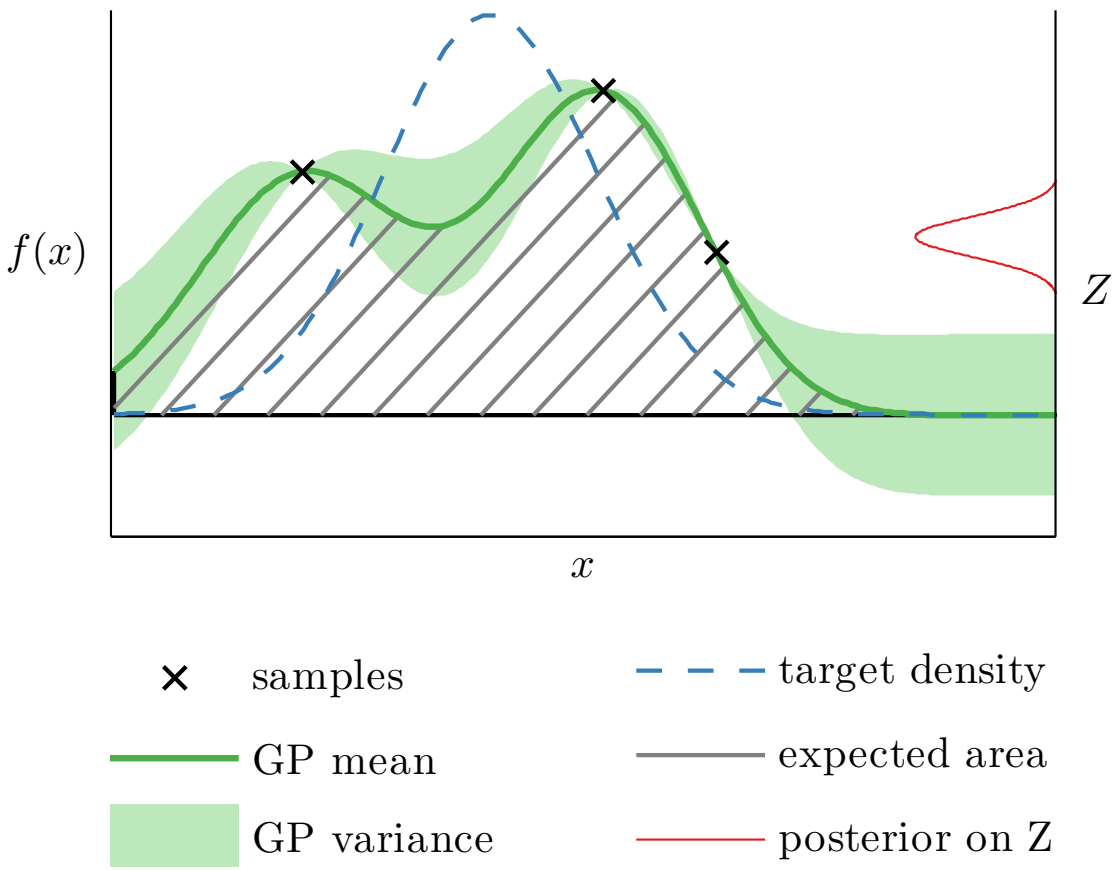

Figure 4.2: An illustration of Bayesian Quadrature. The function $f(x)$ is sampled at a set of input locations. This induces a Gaussian process posterior distribution on f, which is integrated in closed form against the target density, $p(x)$. Since the amount of volume under f is uncertain, this gives rise to a (Gaussian) posterior distribution over $Z_{f,p}$.

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. (4.23).

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

$$p(f(x_\star)|f(X)) = \mathcal{N} f_{x_\star} \bar{f}(x_\star) \mathbb{C}\mathrm{ov}_([x, \star], x'_\star) \tag{4.25}$$

where

$$\bar{f}(x_\star) = k(x_\star, X) K^{-1} f(X) \tag{4.26}$$

$$\mathbb{C}\mathrm{ov}_([x, \star], x'_\star) = k(x_\star, x_\star) - k(x_\star, X) K^{-1} k(X, x_\star) \tag{4.27}$$

and $K = k(X, X)$. Conveniently, the GP posterior allows us to compute the expectation of (4.17) in closed form:

$$\mathbb{E}_{\mathrm{GP}} Z = \mathbb{E}_{\mathrm{GP}} \int f(x) p(x) dx \tag{4.28}$$

$$= \iint f(x) p(f(x)|f(X)) p(x) dx df \tag{4.29}$$

$$= \int \bar{f}(x) p(x) dx \tag{4.30}$$

$$= \left[\int k(x, X) p(x) dx \right] K^{-1} f(X) \tag{4.31}$$

$$= \mathbf{z}^T K^{-1} f(X) \tag{4.32}$$

where

$$z_n = \int k(x, x_n) p(x) dx = \mathbb{E}_{x' \sim p} k(x_n, x'). \tag{4.33}$$

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

$$\mathbb{E}_{\mathrm{GP}} Z = \mathbf{z}^T K^{-1} f(X) \tag{4.34}$$

$$= \sum_n w_{\mathrm{BQ}}^{(n)} f_{x_n}) \tag{4.35}$$

where

$$w_{\mathrm{BQ}}^{(n)} = \sum_m \mathbf{z}_j^T K_{nm}^{-1} \tag{4.36}$$

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.

Weights in Bayesian quadrature

When weighting samples, it is often assumed, or enforced [as in Bach et al., 2012; 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 4.3 shows a representative set of 100 BQ weights chosen on samples representing the distribution in figure 4.1. There are several negative weights, and the sum of all weights is 0.93.

Figure 4.4 demonstrates that, in general, the sum of the Bayesian weights exhibits shrinkage when the number of samples is small.

4.4.3 Sequential 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 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 $f_{x_1} \ldots f_{x_N}$, which we will denote by ϵ_{BQ}^2. For a longer derivation, see [Rasmussen and Ghahramani, 2003].

$$\epsilon_{\mathrm{BQ}}^2(x_1, \ldots, x_N) = \mathbb{V}Z_{f,p}|f_{x_1}, \ldots, f_{x_N} \tag{4.37}$$

$$= \mathbb{E}_{x,x' \sim p} k(x, x') - \mathbf{z}^T K^{-1} \mathbf{z}, \tag{4.38}$$

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 [Krause et al., 2006b]. This allows the optimal samples to be computed ahead of time, before observing any values of f at all [Minka, 2000].

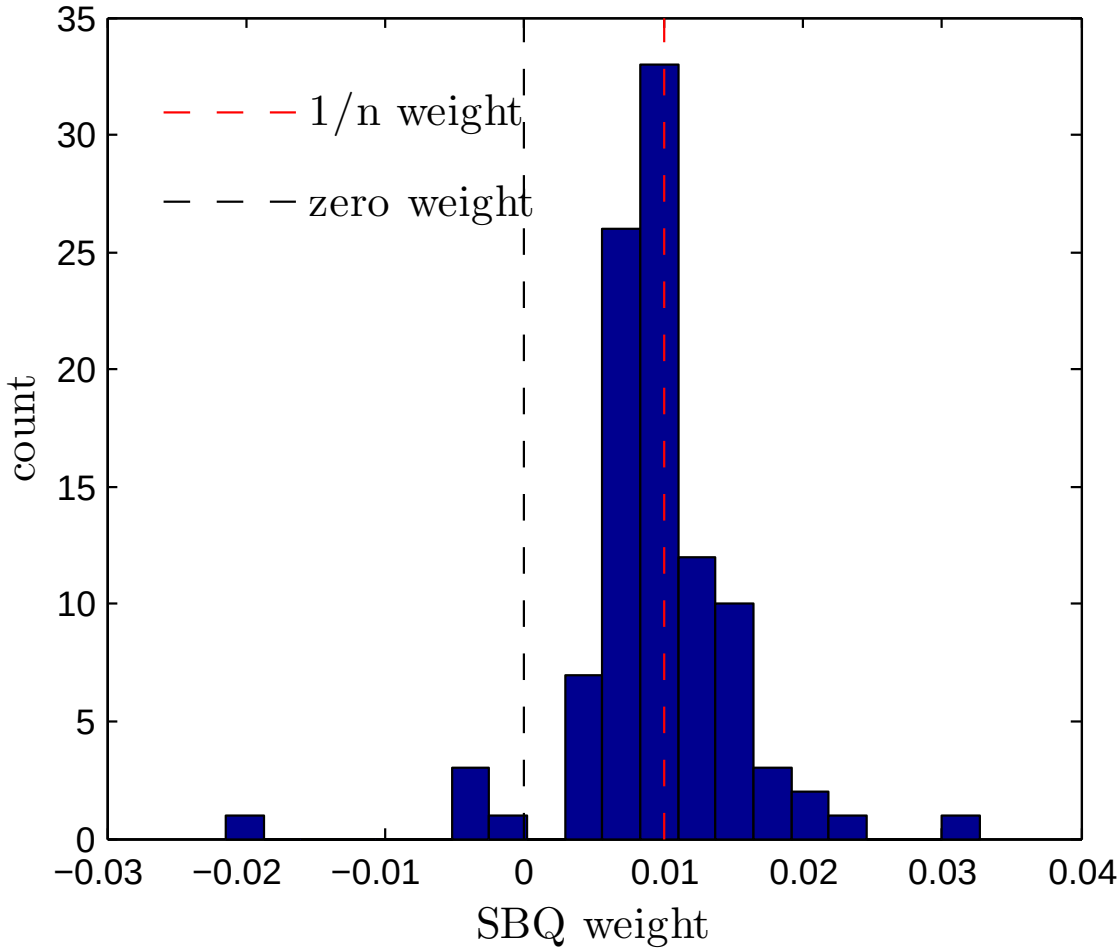

Figure 4.3: A set of optimal weights given by BQ, after 100 SBQ samples were selected on the distribution shown in Figure 4.1. Note that the optimal weights are spread away from the uniform weight ($\frac{1}{N}$), and that some weights are even negative. The sum of these weights is 0.93.

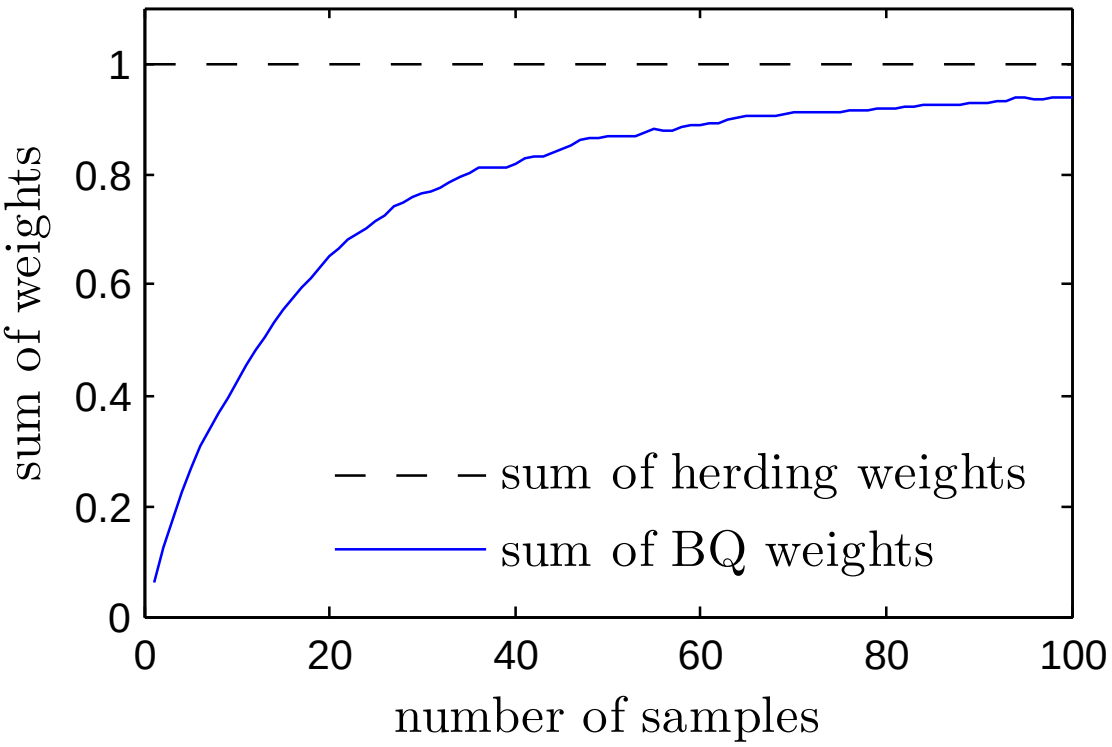

Figure 4.4: An example of Bayesian shrinkage in the sample weights. In this example, the kernel width is approximately $1/20$ the width of the distribution being considered. Because the prior over functions is zero mean, in the small sample case the weights are shrunk towards zero. The weights given by simple Monte Carlo and herding do not exhibit shrinkage.

We can contrast the BQ objective ϵ^2_{BQ} in (4.38) to the objective being minimized in herding, $\epsilon^2_{herding}$ of equation (4.23). 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):

$$x_{n+1} \leftarrow \operatorname*{argmin}_{x \in \mathcal{X}} \epsilon_{\text{BQ}} \left(\{x_1, \ldots, x_n, x\} \right) \tag{4.39}$$

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 $\forall g \in \mathcal{H} : \langle f, g \rangle \sim \mathcal{N}(0, \|g\|_{\mathcal{H}})$:

$$\mathbb{V}Z_{f,p} | f_{x_1}, \ldots, f_{x_N} = \tag{4.40}$$

$$= \mathbb{E}_{f \sim GP} \left(\int f(x) p(x) dx - \sum_{n=1}^{N} w^{(n)}_{\text{BQ}} f(x_n) \right)^2 \tag{4.41}$$

$$= \mathbb{E}_{f \sim GP} \left(\int \langle f, \phi(x) \rangle \, p(x) dx - \sum_{n=1}^{N} w^{(n)}_{\text{BQ}} \langle f, \phi(x_n) \rangle \right)^2 \tag{4.42}$$

$$= \mathbb{E}_{f \sim GP} \left\langle f, \int \phi(x) p(x) dx - \sum_{n=1}^{N} w^{(n)}_{\text{BQ}} \phi(x_n) \right\rangle^2 \tag{4.43}$$

$$= \| \mu_p - \mu_{q_{\text{BQ}}} \|^2_{\mathcal{H}} \tag{4.44}$$

$$= \text{MMD}^2(p, q_{\text{BQ}}) \tag{4.45}$$

□

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:

$$\epsilon^2_{\mathrm{BQ}} = \sup_{\substack{f \in \mathcal{H} \\ \|f\|_{\mathcal{H}} \mathcal{H} = 1}} \left| \int f_x p(x) dx - \sum_{n=1}^{N} w^{(n)}_{\mathrm{BQ}} f_{x_n} \right|^2 \tag{4.46}$$

$$= \inf_{\hat{Z}: \mathcal{X}^N \mapsto \mathbb{R}} \sup_{\substack{f \in \mathcal{H} \\ \|f\|_{\mathcal{H}} = 1}} \left| Z - \hat{Z}\left(f_{x_1}, \ldots, f_{x_N}\right) \right|^2 \tag{4.47}$$

$$= \inf_{\boldsymbol{w} \in \mathbb{R}^N} \sup_{\substack{f \in \mathcal{H} \\ \|f\|_{\mathcal{H}} = 1}} \left| \int f_x p(x) dx - \sum_{n=1}^{N} w_n f_{x_n} \right|^2 \tag{4.48}$$

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

$$\epsilon^2_{\mathrm{BQ}}(x_1, \ldots, x_N) \leq \epsilon^2_{KH}(x_1, \ldots, x_N) \tag{4.49}$$

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.

4.4.4 Approximate submodularity

In this section, we use the concept of approximate submodularity [Krause and Cevher, 2010], 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 $\forall x \in \mathcal{X}$

$$s(A \cup \{x\}) - s(A) \geq s(B \cup \{x\}) - s(B) \tag{4.50}$$

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. Krause and Cevher, 2010, 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 Krause and Cevher [2010], 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 $\forall x \in \mathcal{X}$

$$s(A \cup \{x\}) - s(A) \geq s(B \cup \{x\}) - s(B) - \epsilon \tag{4.51}$$

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

$$r = \max_{x,x' \in \mathcal{P} \subseteq \mathcal{X}} \frac{k(x,x')}{\sqrt{k(x,x)k(x',x')}} \tag{4.52}$$

Proof. By the definition of MMD we can see that $-\epsilon^2_{\text{BQ}} = \inf_{w \in \mathbb{R}^N} \|\mu_p - \sum_{n=1}^N w^{(n)}_{\text{BQ}} k(\cdot, x_n)\|^2_{\mathcal{H}}$ 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 Krause and Cevher [2010] 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

$$s(\mathcal{A}_n) \geq \left(1 - \frac{1}{e}\right) \max_{|\mathcal{A}| \leq n} s(\mathcal{A}) - n\epsilon \tag{4.53}$$

As pointed out by Krause and Cevher [2010], 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.

4.4.5 Experimental evaluation

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 4.1. To ensure a comparison fair to herding, the target distribution, and the kernel used by both methods, correspond exactly to the one used in [Chen et al., 2010, 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 [Chen et al., 2010] 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 4.1 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 (4.36), we can obtain the same expected loss as by using the 20 uniformly-weighted herding samples. Figure 4.5 shows MMD versus the number of samples added, on the distribution shown in Figure 4.1. 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 4.1 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 4.5 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

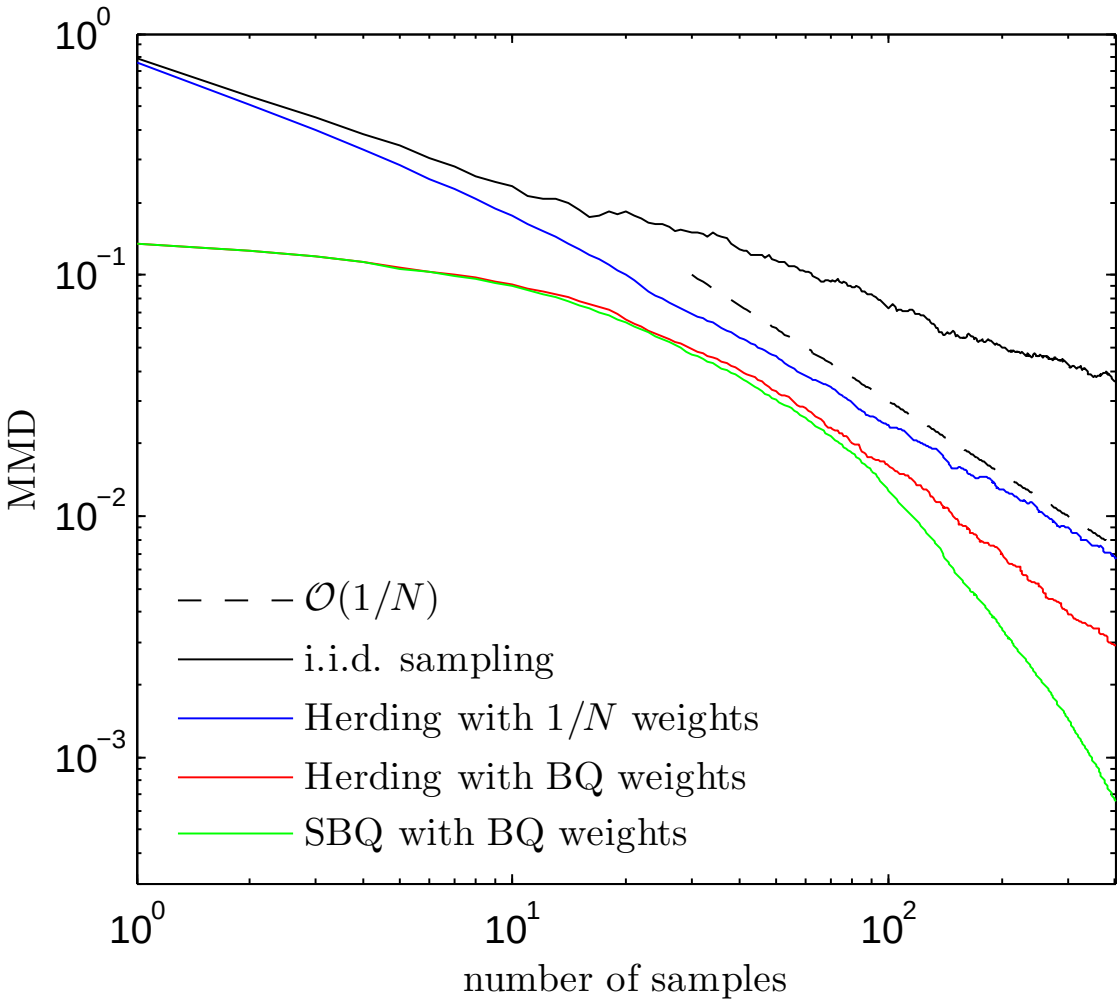

Figure 4.5: The maximum mean discrepancy, or expected error of several different quadrature methods. Herding appears to approach a rate close to $\mathcal{O}(1/N)$. SBQ appears to attain a faster, but unknown rate.

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 form:

$$f(x) = \sum_{i=1}^{10} \alpha_i k(x, c_i) \tag{4.54}$$

where

$$\|f\|_{\mathcal{H}}^2 = \sum_{i=1}^{10}\sum_{j=1}^{10} \alpha_i \alpha_j k(c_i, c_j) = 1 \tag{4.55}$$

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 4.6 shows the empirical error versus the number of samples, on the distribution shown in Figure 4.1. The empirical rates attained by the method appear to be similar to the MMD rates in Figure 4.5.

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. (4.19)), including the Bayesian estimator, SBQ. Figure 4.7 demonstrates this quickly decreasing bound on the SBQ empirical error.

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:

$$f(x) = \sum_{i=1}^{10} \alpha_i \exp(-\frac{1}{2}(x - c_i)^T \Sigma_i^{-1}(x - c_i) \tag{4.56}$$

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 mod-

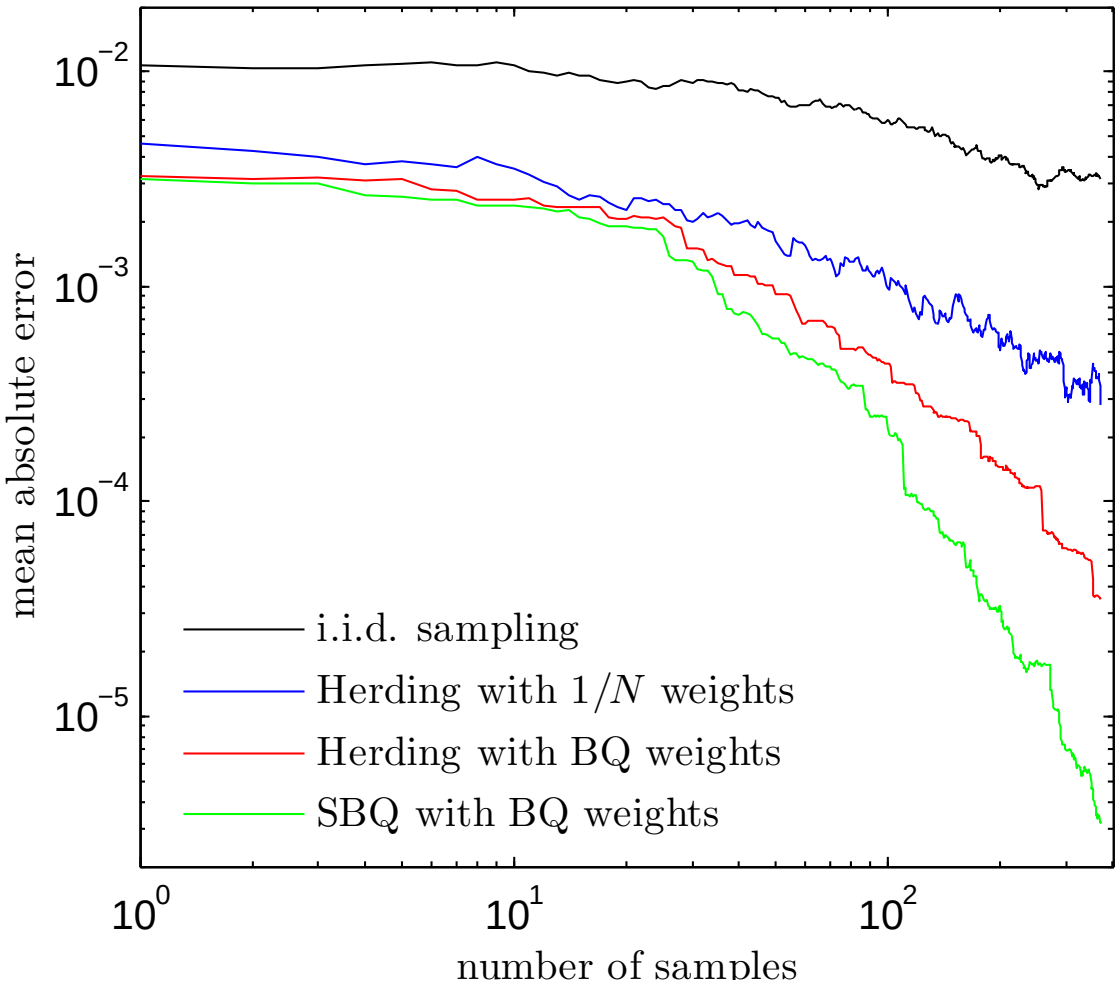

Figure 4.6: Within-model error: The empirical error rate in estimating $Z_{f,p}$, for several different sampling methods, averaged over 250 functions randomly drawn from the RKHS corresponding to the kernel used.

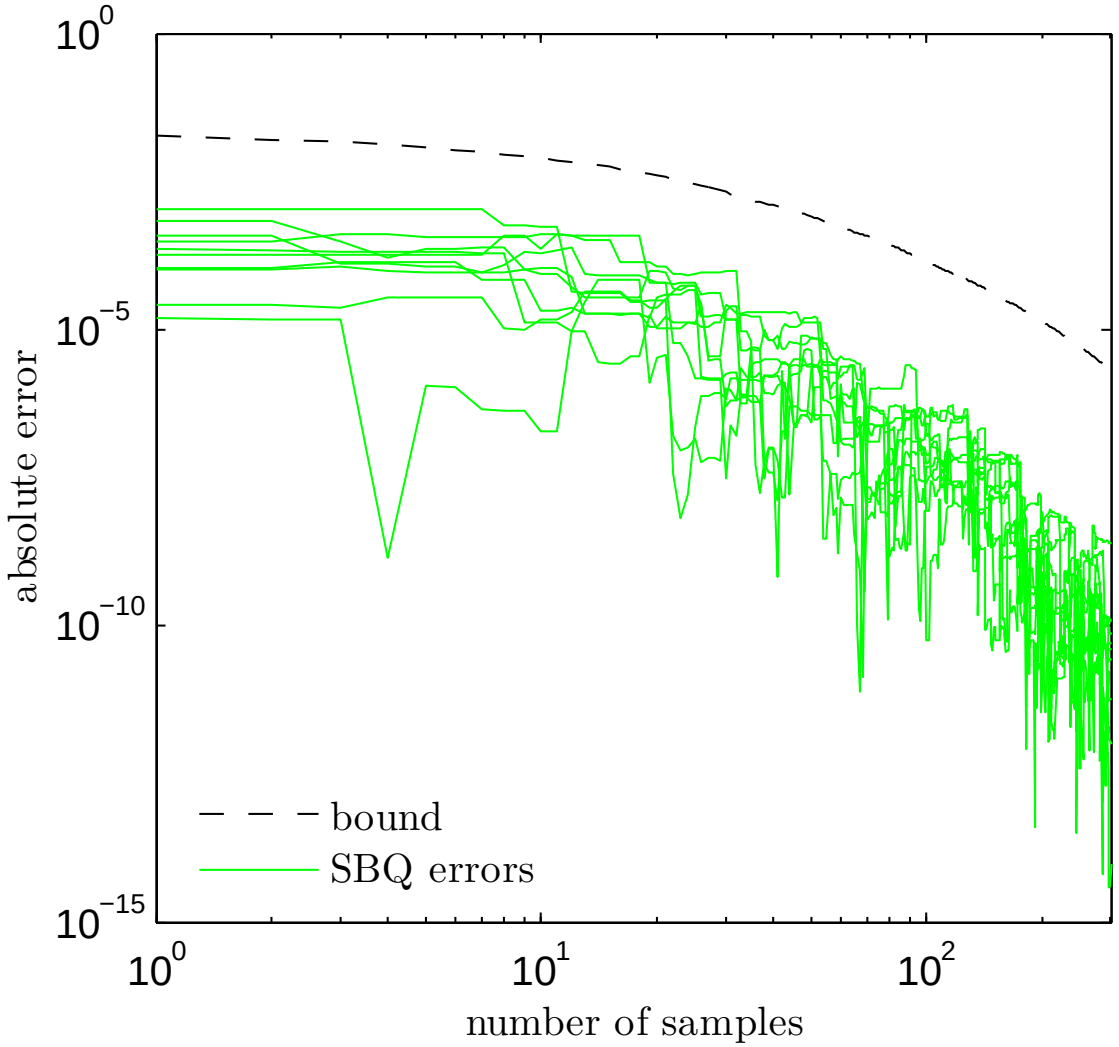

Figure 4.7: The empirical error rate in estimating $Z_{f,p}$, for the SBQ estimator, on 10 random functions drawn from the RKHS corresponding to the kernel used. Also shown is the upper bound on the error rate implied by the MMD.

elled by functions belonging to the isotropic kernel defined by k. Figure 4.8 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/`

4.4.6 Summary and Discussions

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 [Krause and Cevher, 2010].

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.

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 [Sriperumbudur et al., 2008], 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 [Rasmussen and Williams, 2006b].

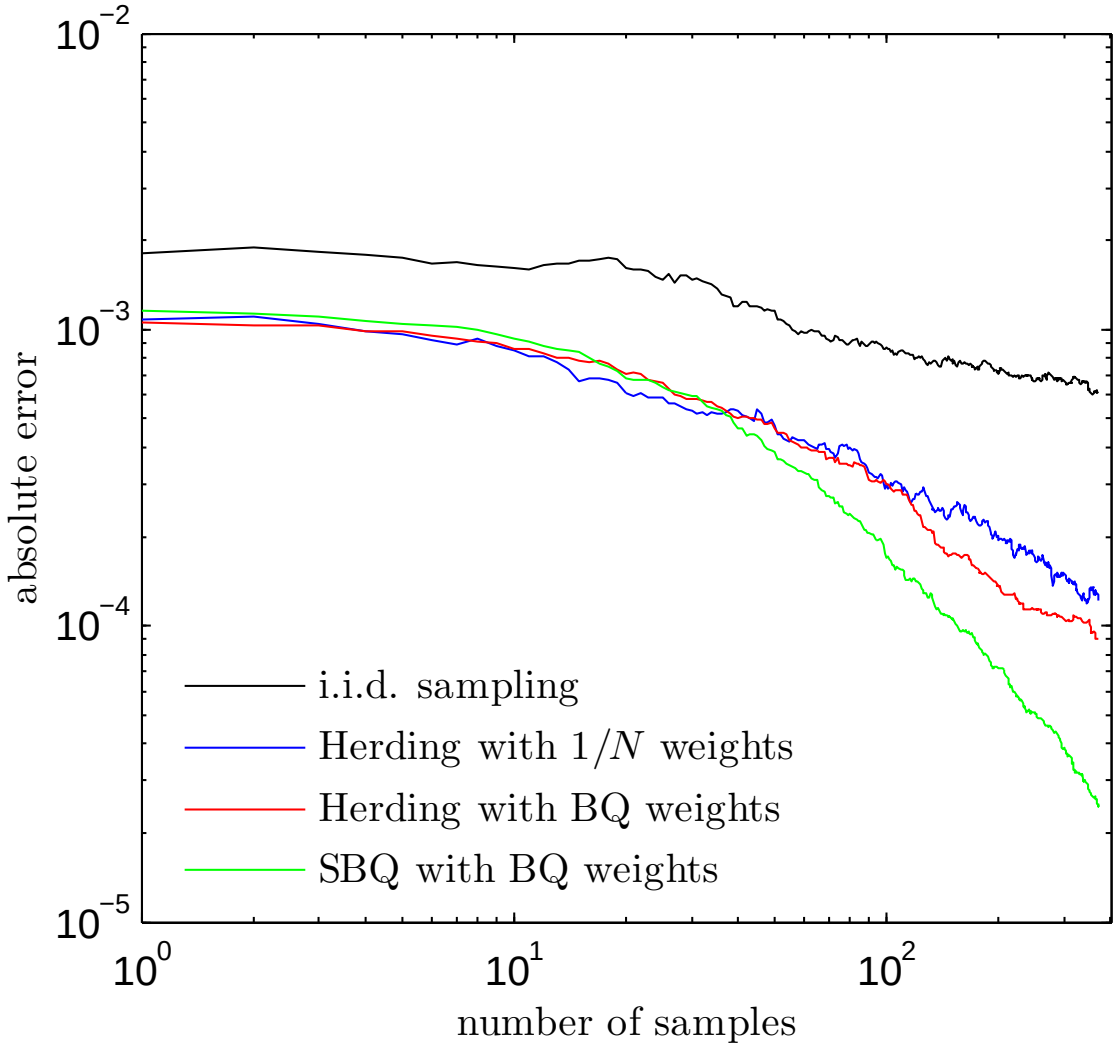

Figure 4.8: Out-of-model error: The empirical error rates in estimating $Z_{f,p}$, for several different sampling methods, averaged over 250 functions drawn from outside the RKHS corresponding to the kernels used.

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	$\mathcal{O}(N^2)$	$\frac{1}{\sqrt{N}} \geq \cdot \geq \frac{1}{N}$	[Bach et al., 2012; Chen et al., 2010]
SBQ	$\mathcal{O}(N^3)$	unknown	approximate submodularity

Table 4.1: A comparison of the rates of convergence and computational complexity of several integration methods.

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 4.1 summarizes the rates of convergence of all the methods considered here.

Future Work

In section 4.4.4, 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 4.4.5, specifically Figure 4.5, 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.

Part III

Optimal Experiment Design

Chapter 5

A Bayesian Framework for Active Learning

Summary of contributions: The unifying framework presented in this chapter unpublished contribution by Ferenc Huszár. A similar framework was previously discussed by Dawid [1994] in an unpublished technical report.

5.1 Introduction

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 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, with foundations laid down by Lindley [1956] and Jaynes [1957], it is still an active area of research.

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. 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 2 can be used to formalise the problem of active learning and optimal experiment design. I devise a framework flexible enough to accomodate and unify a wide range of existing techniques. I will provide a unifying review of Shannon-information-based active learning [Houlsby et al., 2011; Krause et al., 2006a; MacKay, 1992] decision theoretic active learning [Kapoor et al., 2007; Zhu et al., 2003b], Bayesian optimisation [Hennig and Kiefel, 2012; Hennig and Schuler, 2012] and Bayesian quadrature [O'Hagan, 1991; Rasmussen and Ghahramani, 2003].

Building on the theoretical foundations laid down in this chapter, in subsequent chapters I present practical methods and applications. In Chapter 6 derive a computationally convenient algorithm, called Bayesian Active Learning by Disagreement (BALD) and present multiple applications to binary classification, multi-user preference learning and quantum physics (Chapter 7).

5.2 A general framework for Bayesian experiment design

In active learning the goal is to learn about dependence of a *target 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 future queries such that the observed outcomes provide us with the most information about relevant properties of the system. There are many approaches to active learning that carry out predictions and quantify the value of information in different ways. Here we take a Bayesian discriminative approach, that assumes the existence of some latent parameters $\boldsymbol{\theta}$, which directly 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

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

In this chapter I will assume that inference is possible without approximations, and the posterior distribution $p(\boldsymbol{\theta}|\mathcal{D})$ is available in closed form. In practice this is rarely the case, and in subsequent chapters I will apply the framework to cases where only

approximations to the posterior are available.

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 $\boldsymbol{\theta}$.

The posterior $p_{\mathcal{D}}$ is a probabilistic estimate of $\boldsymbol{\theta}$, hence its 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 actual 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 their best estimate of this score is minimised. The best estimate, that is, the Bayes estimator to the score $S(\boldsymbol{\theta}, p_{\mathcal{D}})$ is the expected posterior score or generalised entropy of the posterior $p_{\mathcal{D}}$.

$$\mathbb{H}_S\left[p_{\mathcal{D}}\right] = \mathbb{E}_{\boldsymbol{\theta}\sim p_{\mathcal{D}}} S(\boldsymbol{\theta}, p_{\mathcal{D}}) \tag{5.2}$$

Hence, in the Bayesian framework, the (generalised) entropy of the posterior can be used as a measure of information in data, and the active learner's goal should be to minimise the entropy of the posterior. Recall from section 2.2.7 that generalised entropy is intimately related to Bayes Risk, so in that context the goal of the learner is to collect data so as to reduce the risk of decisions they have to make in the future.

This framework of using a strictly proper scoring rule-based objective in Bayesian active learning has been previously proposed by Dawid [1994], but I am unaware of any subsequent mention or application of it in the statistics or machine learning literature.

Having defined a quantitative measure of the information contained in data, I now turn to discussing how this criterion can be exploited to proactively select measurement x to speed up the process of learning. In this thesis I only consider myopic strategies to active learning, whereby the learner optimises the immediate value of information that the next observation provides, as if the next one was the last measurement to perform. This shortsighted strategy is known to lead to suboptimal performance if the learner is allowed to query multiple measurements in a sequence. However, the optimisation problem for non-myopic strategies get quickly intractable because of the combinatorial nature of the problem [Krause and Guestrin, 2007].

In myopic active learning, the learner solves the same problem in each step. Having

observed some data $\mathcal{D}$, what is the next best measurement x one should make to minimise the entropy of the posterior? This optimisation problem can be formalised as follows:

$$\underset{x}{\text{argmax}} \ \left[\mathbb{H}_S\left[p(\boldsymbol{\theta}|\mathcal{D})\right] - \mathbb{E}_{y\sim p(y|x,\mathcal{D})}\mathbb{H}_S\left[p(\boldsymbol{\theta}|x,y,\mathcal{D})\right]\right] \tag{5.3}$$

The first term is the entropy of the posterior the learner currently has. This is a function of data $\mathcal{D}$ that is already observed, hence it is constant with respect to x and could be ignored. The second term is the entropy of the posterior after making measurement x and observing outcome y. Because at the point of choosing the next measurement the outcome y is unknown, the learner has to rely on an approximation again and take an expectation with respect to its current predictive model of the outcome $p(y|x,\mathcal{D})$. Note that here, the latent parameter $\boldsymbol{\theta}$ has been integrated out:

$$p(y|x,\mathcal{D}) = \int p(y|x,\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathcal{D})d\boldsymbol{\theta} \tag{5.4}$$

Recall Definition 7 of the generalised conditional value of information. Using this definition we can also say that the learner's goal is to maximise the conditional value of information of the measurement x with respect to $\boldsymbol{\theta}$ given $\mathcal{D}$.

$$\underset{x}{\text{argmax}} \ \mathbb{I}_S\left[\boldsymbol{\theta} \leftarrow x|\mathcal{D}\right] \tag{5.5}$$

Another equivalent formulation of the criterion expresses the learner's goal as maximising the expected divergence between the current and new posterior after observing the outcome of the next measurement:

$$\underset{x}{\text{argmax}} \ \mathbb{E}_{y\sim p(y|x,\mathcal{D})}d_S\left[P_{\mathcal{D}\cup\{x,y\}}\middle\|P_{\mathcal{D}}\right] \tag{5.6}$$

The equivalence between these three criteria has been noted in the context of the Shannon's information in [MacKay, 1992]. The framework presented here is a generalisation of MacKay's work, and allows one to specify the goal of Bayesian active learning as a scoring rule for the posterior distribution. In the following section I will review a wide range of Bayesian methods in the literature that all fall within the scope of this approach.

5.3 Examples and special cases

5.3.1 Shannon's entropy

The most commonly used special case of the scoring-rule-based Bayesian active learning framework uses the logarithmic score and hence, Shannon's entropy.

Sannon's entropy is used in a wide range of Bayesian active learning work [Houlsby et al., 2011; Huszár and Houlsby, 2012; Ji et al., 2008; Krause et al., 2006a; Lawrence and Platt, 2004; MacKay, 1992; Settles, 2010].

As discussed in section 2.2.1, Shannon's mutual information is unique in that it is symmetric in its arguments. In Chapter 6 I explore how this property can be exploited in practical implementations of active learning.

5.3.2 Transductive active learning

The goal of supervised learning is to build a model that accurately predicts the outcome y for a test input x. Supervised learning can be divided into inductive and transductive approaches, based on the precise goals that one tries to achieve. In induction, the learner's goal is to induce general rules from the data, that can later be used in a variety of decision tasks. Information theoretic active learning based on Shannon's information is a good example of induction.

In contrast, in transductive machine learning the learner's goal is to predict, from the observed data, the outcomes y for a specific, pre-defined set of test examples. Generalising the solution to outside this training set is not important. Transduction was introduced to machine learning by Vladimir Vapnik [see e. g. Gammerman et al., 1998], who argued that when solving a specific problem of predicting labels on a test set, one should not try to solve a more complicated problem first.

Transductive learning can be expressed in a Bayesian framework [Graepel et al., 1999], and so can transductive active learning as a special case of the scoring-rule based active learning framework. Consider the following scoring rule.

$$S_{\text{trans}}(\boldsymbol{\theta}, p_{\mathcal{D}}) = \mathbb{E}_{x\sim p_{\text{test}}} \mathbb{E}_{p(y|x,\boldsymbol{\theta})} S(y, p(y|x, \mathcal{D})) \tag{5.7}$$

where p_{test} is the distribution of test examples, $p(y|x, \mathcal{D}) = \int p_{\mathcal{D}}(\boldsymbol{\theta}) p(y|x, \boldsymbol{\theta}) d\boldsymbol{\theta}$ is the posterior predictive distribution and S is an arbitrary scoring rule over $\mathcal{Y}$. In transduction, the test examples are often known ahead of time, which can be modeled by setting p_{test} to be uniform mixture of point masses at those test locations

$p_{\text{test}} = {}^1/_{N_{\text{test}}} \sum_{n=1}^{N_{\text{test}}} \delta(x - x_n)$, in which case the criterion can be written as follows.

$$S_{\text{trans}}(\boldsymbol{\theta}, p_{\mathcal{D}}) = \frac{1}{N_{\text{test}}} \sum_{n=1}^{N_{\text{test}}} \mathbb{E}_{p(y|x_n,\boldsymbol{\theta})} S(y, p(y|x_n, \mathcal{D})) \tag{5.8}$$

When S is chosen to be the logarithmic score, the active learning criterion (5.3) becomes the average mutual information between the newly unveiled label and the unknown labels of test examples. This objective function was considered in prior work by MacKay [1992] and is used in various other applications [see e. g. Ertin et al., 2003; Fuhrmann, 2003].

One may also consider a loss-based scoring rule from section 2.2.7

$$S_{\text{trans},\ell}(\boldsymbol{\theta}, p_{\mathcal{D}}) = \mathbb{E}_{x \sim p_{\text{test}}} \left[\min_{\hat{y}} \left\{ \mathbb{E}_{y \sim p(y|x,\boldsymbol{\theta})} \ell(\hat{y}, y) \right\} \right] \tag{5.9}$$

where $\ell(\hat{y}, y)$ is the loss incurred for predicting $\hat{y}$ when the true outcome is y.

Transductive active learning with the loss-based criterion (5.9) is often referred to as decision theoretic active learning. Exact decesion theoretic active learning is computationally demanding to achieve in practice. Kapoor et al. [2007] provide details of such an algorithm in the context of Gaussian process models. Zhu et al. [2003b] use a similar objective function to perform graph-based transductive active learning based using harmonic functions.

5.3.3 Bayesian optimisation

Numerical optimisation techniques are an important tool for many applications in engineering and computational science. The goal is to find the minimum of an objective function f by probing the function at a sequence of locations x_n. At each step, the function value $f(x_n)$, and often local gradients are revealed. Many of these search and optimisation algorithms can be interpreted as active learning algorithms that try to learn about the objective function f, which takes the role of $\boldsymbol{\theta}$ in this framework. This interpretation allowed for the development of novel class of probabilistic optimisation approaches which explicitly model the objective function, and perform probabilistic inference as part of the optimisation procedure. Many of these modern methods use Gaussian processes to model complex surfaces.

The first generation of Bayesian optimisation algorithms were typically based on the

concept of improvement: the aim is to evaluate the objective function at a sequence of points, such that that subsequent function values become lower and lower. Several heuristics have been developed around this idea, including the popular expected improvement [Frean and Boyle, 2008; Jones et al., 1998; Mockus, 1982], probability of improvement [Jones, 2001; Lizotte, 2008] and upper confidence bounds [Srinivas et al., 2009]. These algorithms are prone to getting stuck in local minima because they attempt to collect low function values, rather than to learn about the location of the optimum [Hennig and Schuler, 2012].

The newest class of algorithms [Hennig and Schuler, 2012], called information-greedy algorithms separate the problem of learning about the function and providing an estimate to the minimum. The sequence of function evaluations does not necessarily converge, or indeed it does not generally decrease.

Entropy search, and information-greedy Bayeisan optimisation algorithm presented in [Hennig and Schuler, 2012] is a special case of the scoring-rule-based active learning framework. It employs a scoring rule of the following form:

$$S_{\text{argmin}}(f, p_{\mathcal{D}}) = S(\operatorname*{argmin}_x f, p_{\text{argmin}}), \tag{5.10}$$

where f is the objective function which takes the role of $\boldsymbol{\theta}$ as the parameter of interest. $p_{\mathcal{D}}$ is a posterior measure over objective functions – in most recent work this is a Gaussian process measure. $p_{\text{argmin}} \in \mathcal{M}_{\mathcal{X}}^1$ is the probability distribution that $p_{\mathcal{D}}$ implies over the location of the minimum $\operatorname{argmin}_x f$. Note that in [Hennig and Schuler, 2012] p_{argmin} is called p_{min}, I use a different notation for consistency. The scoring rule S could be any strictly proper scoring rule, Hennig and Schuler use the logarithmic loss which they derive approximations for.

Equation (5.11) implies that the goal of optimisation is to find the exact location of the global minimum. In many applications however this is not required, instead, it is sufficient to learn about the value at the minimum. In this case, one could use a scoring rule of the following form.

$$S_{\text{min}}(f, p_{\mathcal{D}}) = S(\min_x f, p_{\text{min}}), \tag{5.11}$$

where $\min_x f$ is the minimal value of the objective function, p_{min} the measure $p_{\mathcal{D}}$ induces over the minimal value of f.

Lastly, it is possible that neither the location nor the exact value of the minimum is important. The goal is simply to find a location x^* where the function $f(x^*)$ is as small as possible. In most modern applications of numerical optimisation this is a more

reasonable requirement than finding the exact location of the global optimum. This goal can be expressed in the scoring rule framework using the following scoring rule.

$$S_{\text{best}}(f, p_{\mathcal{D}}) = f(\operatorname*{argmin}_x \mathbb{E}_{g\sim p_{\mathcal{D}}} g(x)) \tag{5.12}$$

Note that S_{best} is also a special case of a Bayesian decision score S_ℓ (section 2.2.7, Equation (2.120)) with actions $\mathcal{A} = \mathcal{X}$ and loss function $\ell(f, a) = f(a)$.

Information-greedy Bayesian optimisation methods are still in their infancy. Interpreting them in the general scoring-rule-based framework allows one to understand the underlying goals, and propose straightforward extensions. I am not aware of either (5.11) or (5.12) used in previous work in the context of Bayesian optimisation.

5.3.4 Sequential Bayesian quadrature

Sequential Bayesian quadrature, described in detail in chapter 4.4 is also an example of scoring-rule-based active learning. In that case, one seeks to approximate the expectation of a function f under a probability distribution p:

$$Z_{f,p} = \mathbb{E}_{x\sim p} f(x) = \int f(x)p(x)dx \tag{5.13}$$

Since the integral is intractable to calculate in closed form, the integral is approximated by weighted sum of function evaluations $y_n = f(x_n)$ at particular sampling locations x_n.

$$\hat{Z}_D = \frac{1}{N}\sum_{n=1}^{N} w_n y_n \tag{5.14}$$

The active learning problem involves finding the next sample point x_{N+1} given the points already querried, in such a way that the expected error of the estimate is minimised:

$$S_{SBQ}(f, p_{\mathcal{D}}) = \left(Z_{p,f} - \hat{Z}_{\mathcal{D}}\right)^2 \tag{5.15}$$

Using this scoring rule in the Bayesian active learning framework is equivalent to minimising the expected reduction in posterior variance of the quadrature estimate as in Eqn. (**??**). As we have shown in section 4.4, this optimisation problem corersponds to minimising MMD between the empirical distribution of test locations and the target distribution p.

Instead of the quadratic, one could use a different scoring rule to assess the accuracy

of the forecast of the integral $Z_{p,f}$:

$$S_{quadrature}(f, p_{\mathcal{D}}) = S(Z_{p,f}, p_{Z|\mathcal{D}}), \tag{5.16}$$

where $p_{Z|\mathcal{D}}$ is the distribution that $p_{\mathcal{D}}$ induces over the integral Z. If $p_{\mathcal{D}}$ is a Gaussian process, $p_{Z|\mathcal{D}}$ is always a Normal distribution.

5.4 Summary

Chapter 6

Bayesian Active Learning by Disagreement

Summary of contributions: The work presented in this chapter has been carried out in collaboration with Neil M. T. Houlsby, Jose Miguel Hernandez-Lobato, Zoubin Ghahramani and Máté Lengyel. This work forms the basis of the technical report [Houlsby et al., 2011] and the peer-reviewed conference paper [Houlsby et al., 2012]. In addition, elements of this work have also been presented by Ferenc Huszár and Neil Houlsby at the NIPS 2011 workshops "Preference Learning" and "Bayesian Optimization and Active Learning". All authors contributed equally to the design of the research and to the development of statistical models. The derivation of the preference kernel in Section (6.6.1) and the approximation to the BALD formula in Experimentalqn. (6.23) are original contributions by Ferenc Huszár. Some experiments reported in Figures 6.2 and 6.3 were carried out by Neil Houlsby.

6.1 Introduction

In the previous chapter I have introduced a general framework for Bayesian active learning based on scoring rules. On one hand, the framework is very satisfying in that it highlights the connection between various techniques and optimisation problems that have cropped up in the machine learning literature. On the other hand, the information quantities the framework is built on are intractable to calculate and optimise in general, and the framework provides no practical guidance as to how the quantities can be effectively calculated in practical active learning applications.

In this section I focus on the special case of Bayesian active learning using the Shannon's mutual information (see Section 5.3.1). I present a practical method that

makes active learning possible even in complicated Bayesian models by exploiting the symmetry of Shannon's information (Eqn (2.30)). The method bears resemblance to the principle of maximal disagreement introduced by Seung et al. [1992], hence we call the method Bayesian active learning by Disagreement (BALD).

In the second half of the chapter I derive a Bayesian active learning algorithm for binary classification based on the Gaussian process classification model. I will discuss how this simple model can be extended and applied to binary preference learning and multi-user preference elicitation problems as well. Experimental results presented in this chapter demonstrate that BALD is a very competitive method for active classification and that it often provides best-in-class performance compared to other myopic active learning methods.

6.2 Bayesian Active Learning by Disagreement

Recall the objective function for information theoretic active learning, first proposed in [Lindley, 1956] is to seek the a data point x that satisfies:

$$\underset{x}{\operatorname{argmax}}\, \mathbb{H}\left[\boldsymbol{\theta}|\mathcal{D}\right] - \mathbb{E}_{y\sim p(y|x\mathcal{D})}\left[\mathbb{H}\left[\boldsymbol{\theta}|y,x,\mathcal{D}\right]\right] \tag{6.1}$$

Recall from Chapter 5, that expectation over the unseen output y is required because this will not be revealed until the selection of the next input x is made.

Computationally, Eqn. (6.1) poses two difficulties: Firstly, to consider k different potential queries x, when the output y may take on l possible values, one has to apply Bayes' rule and update the posterior $k \cdot l$ times to evaluate the objective for each x–y pair in question. Because updating the posterior is often the computationally most intensive part of Bayesian active learning, performing this step multiple times can be prohibitive.

Secondly, calculating entropies in parameter space may be hard, or intractable. Often we can only approximate the posterior via samples, from which estimating entropy is notoriously difficult [Panzeri and Petersen, 2007]. Worse still, for non-parametric processes parameter space is infinite dimensional so Eqn. (6.1) becomes poorly defined and non-trivial to compute. [Krishnapuram et al., 2004; Lawrence and Platt, 2004; MacKay, 1992] use this objective but must make approximations to the complicated entropy term.

If we use the logarithmic score to define the entropy in Eqn. (6.1), both of these problems can be overcome. As discussed in Chapter 2, Shannon's mutual information, unlike the value of information functional in general, is symmetric. The information

variable X provides about Y is the same as the information Y provides about X. In this chapter we expoloit this unique symmetry property and rearrange Eqn. (6.1) in a form that expresses $\mathbb{I}_\theta\left[y \leftarrow \mathcal{D}; x\right]$ in terms of entropies in the output space $\mathcal{Y}$:

$$\underset{x}{\operatorname{argmax}}\, H[\boldsymbol{\theta}|\mathcal{D}] - \mathbb{E}_{y\sim p(y|x\mathcal{D})}\left[H[\boldsymbol{\theta}|y,x,\mathcal{D}]\right] = \tag{6.2}$$

$$\underset{x}{\operatorname{argmax}}\, \mathbb{I}_{\text{Shannon}}\left[\theta \leftarrow y|\mathcal{D}; x\right] = \tag{6.3}$$

$$\underset{x}{\operatorname{argmax}}\, \mathbb{I}_{\text{Shannon}}\left[y \leftarrow \boldsymbol{\theta}|\mathcal{D}; x\right] = \tag{6.4}$$

$$\underset{x}{\operatorname{argmax}}\, H[y|x,\mathcal{D}] - \mathbb{E}_{\boldsymbol{\theta}\sim p(\boldsymbol{\theta}|\mathcal{D})}\left[H[y|x,\boldsymbol{\theta}]\right] \tag{6.5}$$

Eqn. (6.5) overcomes the aforementioned challenges. Entropies are now calculated in the – usually low dimensional – output space. Also $\boldsymbol{\theta}$ is now conditioned only on $\mathcal{D}$, never on x and y, so we do not need to update the posterior for every possible outcome, saving a factor of $k \cdot l$ posterior updates.

Equation (6.5) provides us with an intuition about the objective; we seek the x for which the model is marginally most uncertain about y (high $H[y|x,\mathcal{D}]$), but for which individual setting of the parameters are confident (low $\mathbb{E}_{\boldsymbol{\theta}\sim p(\boldsymbol{\theta}|\mathcal{D})}\left[H[y|x,\boldsymbol{\theta}]\right]$). This can be interpreted as seeking the x for which the parameters under the posterior disagree about the outcome the most. Indeed, the objective function can further be written as:

$$\underset{x}{\operatorname{argmax}}\, H[y|x,\mathcal{D}] - \mathbb{E}_{\boldsymbol{\theta}\sim p(\boldsymbol{\theta}|\mathcal{D})}\left[H[y|x,\boldsymbol{\theta}]\right] = \tag{6.6}$$

$$\underset{x}{\operatorname{argmax}}\, \mathbb{E}_{\boldsymbol{\theta}\sim p(\boldsymbol{\theta}|\mathcal{D})} d_{KL}\left[p(y|\boldsymbol{\theta};x)\|p(y|\mathcal{D};x)\right] \tag{6.7}$$

So by maximising the expected reduction in posterior entropy, the learner seeks a measurement, such that each probable parameter θ predicts something different for what the outcome y will be. In the statistics iterature this is known as the *principle of maximal disagreement* and provides the theoretical motivation for *query by committee* methods [Freund et al., 1997]. We will call the approach of using the rearranged objective in Eqn. (6.5) as Bayesian Active Learning by Disagreement or BALD [Houlsby et al., 2011, 2012; Huszár and Houlsby, 2012]. Note that these rearrangements were only possible because of the symmetry of Shannon's mutual information, hence this interpretation does not hold in general for Bayesian active learning when using a scoring rule other than the logarithmic.

6.3 Related Techniques

In this section we briefly review some of the very many related algorithms that are applicable to active learning and relate them to the BALD information theoretic objective (6.5).

Other work that uses rearrangement to observation space (Eqn. (6.5)) include Maximum Entropy Sampling (MES) [Sebastiani and Wynn, 2000]. MES was proposed for regression models with input-independent observation noise. The MES objective function is similar to Eqn. (6.5), but with its second term ignored. This is theoretically well motivated if the output noise is indeed independent of the input, that is when $H[y|\boldsymbol{\theta}, x]$ is constant in x.

For heteroscedastic regression or other applications such as classification, this assumption is violated, thus MES is inappropriate; it fails to differentiate between model uncertainty and observation uncertainty (about which our model may be confident). Many toy demonstrations show the 'information based' active learning criterion performing pathologically in classification by repeatedly querying points close the decision boundary or in regions of high observation uncertainty, e. g. those presented in [Dasgupta and Hsu, 2008; Huang et al., 2010]. This is because MES is inappropriate in those circumstances. Using the full BALD objective distinguishes between observation and model uncertainty and eliminates these problems.

The Informative Vector Machine [IVM, **?**] algorithm was designed for sub-sampling a data set to be used to train a Gaussian process regression or classification model. It may not fall under the term 'active learning' because all y values are revealed to the algorithm a priori. Their objective is based on Shannon's entropy as in Eqn. (6.1), however the algorithm is not based on a rearrangement to data space (Eqn. (6.5)).Posterior entropy calculations are made approximately on the n dimensional subspace corresponding to the n observed data points using the Gaussian process covariance matrix. To chose among a pool of k possible points, k posterior updates are required in each step. [**?**] proposes a quick method based on Assumed Density Filtering to perform these updates quickly. The IVM essentially optimises the same objective function, but makes different trade-offs and approximations. As we will see, BALD often performs better in practical situations.

Certain non-probabilistic methods have close analogues to information theoretic active learning. Perhaps the most ubiquitous is active learning for support vector machines [SVM, Seung et al., 1992; Tong and Koller, 2001] where the volume of version space is used as an objective function. This objective function is in fact closely related to Bayesian

active learning. If a uniform (improper) prior is used with a deterministic classification likelihood it can be shown that the logarithmic volume of version space and Bayesian posterior entropy are equivalent.

However, just as Bayesian posteriors become intractable after observing many data points, version space too can become very complicated and its volume intractable to compute. [Tong and Koller, 2001] proposes approximating version space with a simple shape, such as a hyper-sphere. This closely resembles approximating a Bayesian posterior using a Gaussian distribution via the Laplace or EP approximations. [Seung et al., 1992] sidesteps the problem by working in the space of observations, much like BALD does. The algorithm, Query by Committee (QBC), samples parameters from version space (committee members), which vote on the outcome of each possible measurement x in question. The x with the most balanced vote is selected; this is termed the 'principle of maximal disagreement'. If BALD is used in conjunction with a Monte Carlo approximation to the posterior, the resulting algorithm is very similar to query-by-committee, but with a theoretically sound, probabilistic measure of disagreement. QBC's deterministic vote criterion discards confidence in the predictions and so can exhibit similar pathologies as MES.

6.4 BALD for Gaussian Process Classification

BALD exploits the fact that in many active learning applications the output space $\mathcal{Y}$ is often simpler than the parameter space Θ. Here I consider the problem of active learning for binary classification, when the output takes one of two possible values $y \in \{-1, 1\}$. Given the simplicity of the outputs, binary classification is a highly relevant use-case for BALD.

I will use a non-parametric Bayesian classification model, Gaussian process classification [GPC, Rasmussen and Williams, 2006a] to demonstrate the usefulness of BALD: GPC appears to be an especially challenging problem for information-theoretic active learning because its parameter space is infinite. Therefore, computing entropy of the posterior involves nontrivial quantities. However, by using the BALD approach and Eqn. (6.5) we are able to fully calculate the relevant information quantities without having to work out entropies of infinite dimensional objects.

The probabilistic model underlying GPC is as follows:

$$f \sim \mathrm{GP}(\mu(\cdot), k(\cdot, \cdot)) \qquad y|x, f \sim \mathrm{Bernoulli}(\Phi(f(x))) \tag{6.8}$$

The latent parameter, now called f (previously denoted as $\boldsymbol{\theta}$), is a real-valued function $\mathcal{X} \to \mathbb{R}$, and is assigned a Gaussian process prior with mean $\mu(\cdot)$ and covariance function $k(\cdot, \cdot)$.

TODO: I need an appendix on kernels and Gaussian processes

We consider the probit case where, given the value of f, the binary label y takes a Bernoulli distribution with probability $\Phi(f(x))$, and Φ is the cumulative distribution function of the normal distribution. For further details on GPC see [Rasmussen and Williams, 2005].

Posterior inference in the GPC model is intractable; given some observations $\mathcal{D}$, the posterior over f becomes non-Gaussian and complicated. This is addressed by using approximate inference methods. The most commonly used approximate inference methods for Gaussian process classification are expectation propagation [EP, Minka and Lafferty, 2002], Laplace's approximation [Williams and Barber, 1998], assumed density filtering [ADF, Csató et al., 2000] and sparse methods [QuiÒonero-Candela and Rasmussen, 2005]. These all approximate the non-Gaussian posterior by a Gaussian [Nickisch and Rasmussen, 2008], but differ in the optimisation criterion and other restrictions. Throughout this chapter I will assume that we are provided with some Gaussian approximation to the GPC posterior resulting from one of these methods, though the active learning method is agnostic as to which method produced this estimate. Given the sequential nature of active learning, fast on-line methods [Csató and Opper, 2002] are particularly well suited for the task. In our derivation we will use $\overset{1}{\approx}$ to indicate where approximate inference is exploited.

6.4.1 Computing the value of information

Now, we will compute the informativeness of a query x using Eqn. (6.5). Recall that now the parameter $\boldsymbol{\theta}$ is the latent function f:

$$\operatorname*{argmax}_x H[y|x, \mathcal{D}] - \mathbb{E}_{f\sim p(\boldsymbol{\theta}|\mathcal{D})}\left[H[y|x, f]\right] \tag{6.9}$$

To evaluate the objective one has to compute expectations over the posterior $p_{\mathcal{D}}$. As discussed earlier, we use a Gaussian approximation to the posterior, so that for each x, $f_x = f(x)$ follows a Gaussian distribution with mean $\mu_{x,\mathcal{D}}$ and variance $\sigma^2_{x,\mathcal{D}}$. The exact value of $\mu_{x,\mathcal{D}}$ and $\sigma^2_{x,\mathcal{D}}$ depend on the approximation scheme used as well as the covariance kernel.

$$p(f_x|\mathcal{D}) \overset{1}{\approx} \mathcal{N}(f_x|\mu_{x,\mathcal{D}}, \sigma^2_{x,\mathcal{D}}) \tag{6.10}$$

To compute the two terms in Eqn. (6.5) we have to compute two entropy quantities. $H[y|x, f]$, the entropy of the binary output variable y given a fixed f can be expressed in terms of the binary entropy function h:

$$H[y|x, f] = h\left(\Phi(f(x)\right), \qquad h(p) = -p \log p - (1-p)\log(1-p) \tag{6.11}$$

Similarly, the first term in Eqn. (6.5), $H[y|x, \mathcal{D}]$ can be computed analytically:

$$H[y|x, \mathcal{D}] = h\left(\int \Phi(f_x) p(f_x|\mathcal{D}) df_x\right) \tag{6.12}$$

$$\overset{1}{\approx} h\left(\int \Phi(f_x) \mathcal{N}(f_x|\mu_{x,\mathcal{D}}, \sigma^2_{x,\mathcal{D}}) df_x\right) \tag{6.13}$$

$$= h\left(\Phi\left(\frac{\mu_{x,\mathcal{D}}}{\sqrt{\sigma^2_{x,\mathcal{D}} + 1}}\right)\right), \tag{6.14}$$

where we used the fact that the sum of Gaussian variables is Gaussian distributed. This trick is only possible because in probit classification the link function $\Phi(\cdot)$ can be interpreted as the cumulative distribution function of a Normal distribution. In logistic regression, instead of the Gaussian CDF, one would use a logistic sigmoid where the same trick cannot be applied and further approximations are required.

The missing piece in evaluating the BALD objective function is the expectation $\mathbb{E}_{f\sim p(\boldsymbol{\theta}|\mathcal{D})}\left[H[y|x, f]\right]$. Unfortunately, for GPC this quantity cannot be computed in closed form, even assuming a Gaussian approximation to the posterior. However, with a further approximation (denoted by $\overset{2}{\approx}$), the expectation can be accurately approximated as follows.

$$\mathbb{E}_{f\sim p(f|\mathcal{D})}\left[H[y|f]\right] \overset{1}{\approx} \int h(\Phi(f_x))\mathcal{N}(f_x|\mu_{x,\mathcal{D}},\sigma^2_{x,\mathcal{D}})df_x \tag{6.15}$$

$$\overset{2}{\approx} \int \exp\left(-\frac{f_x^2}{\pi\ln 2}\right)\mathcal{N}(f_x|\mu_{x,\mathcal{D}},\sigma^2_{x,\mathcal{D}})df_x \tag{6.16}$$

$$= \frac{C}{\sqrt{\sigma^2_{x,\mathcal{D}}+C^2}}\exp\left(-\frac{\mu^2_{x,\mathcal{D}}}{2\left(\sigma^2_{x,\mathcal{D}}+C^2\right)}\right), \tag{6.17}$$

where $C=\sqrt{\frac{\pi\ln 2}{2}}$. The first approximation, $\overset{1}{\approx}$, reflects the Gaussian approximation to the , as before.

The integral in the left hand side of Eqn. (6.15) is hard to compute; the non-linear function $h(\Phi(\cdot))$ must be integrated against a Gaussian distribution. However, $h(\Phi(t))$ can be approximated accurately by an unnormalised Gaussian curve, $\exp(-t^2/\pi\ln 2)$, making the integral trivial to compute via convolution formulas for Gaussian distributions.

Consider the Taylor expansion of $g := \log(h(\Phi(\cdot)))$.

$$g(t) = g(0) + \frac{g'(0)t}{1!} + \frac{g''(0)t^2}{2!} + \dots \tag{6.18}$$

The derivatives in the formula can be computed in closed form using the chain rule

$$g(t) = \log(h(\Phi(t))) \tag{6.19}$$

$$\begin{aligned} g'(t) &= -\frac{1}{\log 2}\frac{\Phi'(t)}{h(\Phi(t))}\left[\log\Phi(t) - \log(1-\Phi(t))\right) \\ f''(t) &= \frac{1}{\log 2}\frac{\Phi'(t)^2}{h(\Phi(t))^2}\left(\log\Phi(t) - \log(1-\Phi(t))\right) \\ &\quad - \frac{1}{\log 2}\frac{\Phi''(t)}{h(\Phi(t))}\left(\log\Phi(t) - \log(1-\Phi(t))\right) \\ &\quad - \frac{1}{\log 2}\frac{\Phi'(t)^2}{h(\Phi(t))}\left(\frac{1}{\Phi(t)} + \frac{1}{(1-\Phi(t)}\right)\Big) \end{aligned} \tag{6.20}$$

Therefore the following Taylor series approtimation holds:

Figure 6.1: *Left:* Analytic approximation ($\overset{1}{\approx}$) to the binary entropy of the error function (**??**) by a squared exponential (**??**). The absolute error (**??**) remains under $3 \cdot 10^{-3}$. *Right:* Percentage approximation error (± 1 s.d.) for different methods of approximate inference (*columns*) and approximation methods for evaluating Eqn.(6.15) (*rows*). The results indicate that $\overset{2}{\approx}$ is a very accurate approximation; EP causes some loss and Laplace significantly more, which is in line with the comparison presented in [**?**].

$$\log h(\Phi(t)) \overset{2}{\approx} -\frac{1}{\pi \log 2} t^2 \tag{6.21}$$

Consequently, exponentiating both sides, we can make the following Gaussian approximation to $h(\Phi(\cdot))$:

$$h(\Phi(t)) \overset{2}{\approx} \exp\left(-\frac{t^2}{\pi \log 2}\right) \tag{6.22}$$

Fig. 6.1 depicts the striking accuracy of this approximation. The maximum possible error that will be incurred when using this approximation in the BALD formula is if the approximate posterior $\mathcal{N}(f_x|\mu_{x,\mathcal{D}}, \sigma^2_{x,\mathcal{D}})$ is centred at $\mu_{x,\mathcal{D}} = \pm 2.05$ with $\sigma^2_{x,\mathcal{D}}$ tending to zero (see Fig. 6.1, absolute error **??**); even this yields only a 0.27% error in the integral in Eqn.(6.15).

In subsequent sections we investigate experimentally the information lost from approximations $\overset{1}{\approx}$ and $\overset{2}{\approx}$ as compared to the golden standard of extensive Monte Carlo simulation.

To summarise, the BALD algorithm for Gaussian process classification consists of two steps. First it applies an approximate inference algorithm to obtain the posterior predictive mean $\mu_{x,\mathcal{D}}$ and $\sigma_{x,\mathcal{D}}$ for each point of interest x. Then, it selects a query x that maximises the following objective function:

$$\mathrm{h}\left(\Phi\left(\frac{\mu_{x,\mathcal{D}}}{\sqrt{\sigma^2_{x,\mathcal{D}} + 1}}\right)\right) - \frac{C \exp\left(-\frac{\mu^2_{x,\mathcal{D}}}{2\left(\sigma^2_{x,\mathcal{D}} + C^2\right)}\right)}{\sqrt{\sigma^2_{x,\mathcal{D}} + C^2}} \tag{6.23}$$

For most practically relevant kernels, the objective (6.23) is smooth, and differentiable function of x, so gradient-based optimisation procedures can be used to find the maximally informative query.

6.5 Experiments and Results

6.5.1 Quantifying approximation losses

Recall that to obtain Eqn. (6.23) we made two approximations: we perform approximate inference ($\overset{1}{\approx}$), and we approximated the binary entropy of the Gaussian CDF by a squared exponential ($\overset{2}{\approx}$). Both of these can be substituted with extensive Monte Carlo approximation, enabling us to compute a near-exact, unbiased estimate of the expected information gain. Using extensive Monte Carlo as the 'gold standard', we can evaluate how much we loose by applying these approximations. We quantify approximation error as:

$$\frac{\max_{x \in X} I(x) - I(\operatorname{argmax}_{x \in X} \hat{I}(x))}{\max_{x \in X} I(x)} \cdot 100\%, \tag{6.24}$$

where I is the objective computed using Monte Carlo, $\hat{I}$ is the approximate objective. We have run experiments on the Lung Cancer (*cancer*) classification dataset [Hong and Yang, 1991] downloaded from the UCI Machine Learning Repository [Bache and Lichman, 2013]. Results are shown and discussed in Figure 6.1. In general we see that the approximation error from approximate inference ($\overset{1}{\approx}$) typically outweighs the negligible error introduced by the Taylor-series approximation ($\overset{2}{\approx}$).

6.5.2 Pool based active learning

We test BALD for GPC and preference learning in the pool-based setting i. e. selecting x values from a fixed set of potential query locations. We compare to eight other algorithms discussed in this paper: random sampling, maximum entropy sampling (MES), query by committee (QBC), support vector machine (SVM) with version space approximation Tong and Koller [2001], decision theoretic approaches from [Kapoor et al., 2007; Zhu et al., 2003a] and direct minimisation of expected empirical error. The latter method is not widely used, but is included for analysis of Kapoor et al. [2007].

First we consider three artificial, but challenging, datasets in two dimensional input space $\mathcal{X} = \mathbb{R}^2$. These datasets are designed to exaggarate problematic edge cases that often cause active learning methods to fail. The first, *block in the middle*, has a block of noisy, randomly labeled points on the decision boundary between two otherwise clearly separable classes. The second, *block in the corner*, has a block of uninformative points far from the decision boundary. A capable active learning algorithm should avoid these

Figure 6.2: *Top:* Artificial data sets used in our evaluation of active learning methods. Exemplars of the two classes are shown with black squares (**??**) and red circles (**??**). *Bottom:* Results of active learning with nine methods: random query (**??**), BALD (**??**), MES (**??**), QBC with the vote criterion with 2 (QBC_2, **??**) and 100 (QBC_{100}, **??**) committee members, active SVM (**??**), IVM (**??**), Kapoor et al. [2007] (**??**), Zhu et al. [2003a] (**??**) and empirical error (**??**).

Figure 6.3: Evaluation of the accuracy of active learning algorithms on various real-world classification datasets. Methods used are random query (**??**), BALD (**??**), MES (**??**), QBC with 2 (QBC_2, **??**) and 100 (QBC_{100}, **??**) committee members, active SVM (**??**), IVM (**??**), decision theoretic [Kapoor et al., 2007] (**??**), semi-supervised [Zhu et al., 2003a] (**??**) and empicial error (**??**). The decision theoretic methods took a long time to run, so were not completed for all data sets. Plots (a-i) correspond to different classification datasets from the UCI Machine Learning Repository. Plot (i) includes BALD with hyper-parameter learning (**??**). See text for analysis.

uninformative regions. The third data set, similar to the *checkerboard* dataset used in Zhu et al. [2003a], is designed to test the algorithm's capabilities to find multiple disjoint islands of points from one class.

The three datasets and results using each algorithm are depicted in Fig. 6.2. We can see that BALD (**??**) handles all three artificial datasets well. Transductive criteria, Kapoor et al. [2007] (**??**) and Zhu et al. [2003a] (**??**), struggle on the *block in the middle* dataset, because the random noise added to the central block of points dominates the risk of classification. In the third example, they are very effective in exploring the centre of each island of data points early on.

In addition to this, we present results on 8 real-world classification datasets from the UCI Machine Learning Repository [Bache and Lichman, 2013]: *australia, crabs, wine, vehicle, isolet, cancer, letter* and *wdbc*. *Letter* is a multiclass dataset from which we select hard-to-distinguish letters E vs. F and D vs. P as binary sub-problems. Results on these data sets are plotted in Fig. 6.3.

We can see from Figs 6.2 and 6.3 that by using BALD we make significant gains over na ive random sampling in both the classification adn preference learning domains. Relative to other active learning algorithms BALD performs consistently well across all datasets, particularly when avoiding the block of points in Fig. 6.2 (a). Occasionally e. g. as Fig. 6.3 (i), it performs poorly on the first couple of queries.

In most reported experiments we have fixed the kernel k of the Gaussian process prior to the maximum likelihood estimate on the whole pool. This is of course cheating, as it uses information from the whole dataset before starting to select queries, but it provides us with a fair way of comparing various methods, including those that cannot handle hyperparameter learning. As described in [Houlsby et al., 2011, 2012], BALD can accommodate active learning of hyperparameters. In Fig. 6.3 (i) we also show the performance of this method, and on the *cancer* dataset it helps to overcome the initial poor performance of BALD.

MES often performs as well as BALD e. g. on Fig. 6.2(c), where there is no label noise. It never outperforms BALD though and on noisy datasets (e. g. Fig. 6.2(a)) performs particularly poorly as expected. QBC provides a close approximation to BALD and usually provides a small decrement in performance. However, there is a large decrease in performance on the noisy artificial dataset caused by the vote criterion not maintaining a notion of inherent uncertainty, like MES. The IVM occasionally performs well, but often exhibits highly pathological behaviour; by observing y values in advance it actively chooses noisy or mislabelled points, thinking them informative. The SVM-based approach exhibits variable performance (it does extremely well on Fig. 6.3 (f), but very poorly on 6.2 (c)).

On the real datasets though BALD usually performs as well, if not better, than transductive methods of Kapoor et al. [2007] and Zhu et al. [2003a], despite not having access to the locations of the test points and having a significantly lower computational cost. The [Kapoor et al., 2007] objective sometimes fails badly, this is likely to be because one term in their objective function is the empirical error. The weighting of this term is determined by the relative sizes of the training and test set. Directly minimizing empirical error usually performs very pathologically, picking only ''safe' points; when the [Kapoor et al., 2007] objective assigns too much weight to this term it also fails.

6.6 Extension to preference elicitation

It is possible to extend the above framework to handle a practically highly relevant problem problem of learning peoples' preferences. People have rich knowledge knowledge and information about which products, services, people, items they prefer, find attractive or like. Methods that uncover this knowledge is invaluable in a number of commercial and science applications. Examples include

market research and e-commerce: learning about users' preferences of products,

prices, or brands. Preferences can be exploited to maximise user satisfaction and to drive profit

social media: identifying people whom their peers find influential, reliable, trustworthy or knowledgeable on certain topics, and using this information to find domain experts, or drive influence marketing

recommendation: on review websites collecting and quantifying user feedback on restaurants, activities, movies, music albums, etc. to power recommendations

research: learning about difficult, subjective concepts such as attractiveness, for example investigating which features determine perceived attractiveness

equipment calibration: calibration of parameters to improve perceived subjective quality. Examples include calibrating sound quality of hearing aid or stereo equipment, high-dimensional parameter optimisation in digital image rendering

Many existing approaches – such as traditional market research surveys, restaurant review websites, DVD rental websites, etc – require human respondents[1] to give ratings of items on an absolute scale. Market research surveys often use a scale of 1 to 7, while review websites use star ratings, typically on a scale between 1 to 5 stars. There are multiple problems with this approach.

1. People's baseline level on the absolute scale may differ. One person's 4 star rating may describe the same level of satisfaction as someone else's 5 stars. This makes aggregating opinions from many different people a non-trivial task

2. The variance of responses may also differ across people: some more conservative reviewers would never use the extreme 1 star or 5 star ratings, whilst other respondents' opinions may be more polarised

3. To give informed ratings, the user has to know the distribution of the quality of items ahead of time. They may prefer not to give a maximal 5-star rating to an item, because they don't know if better items exist. Others may give 5-star to a mediocre item, because they have never seen a better alternative.

To overcome the limitations of an absolute scale, in an increasing number of applications preference elicitation is done via pairwise item comparisons. In this case, the

[1]Throughout this chapter I will use the words person, respondent and user interchangeably to mean the person whose preferences we are interested in predicting.

respondent is presented a pair of items, and they have to judge which of the two alternatives is more preferable. In cognitive science, this type of preference elicitation is known as two-alternative forced choice, or 2AFC for short [Fechner, 1860; Platt and Glimcher, 1999]. The machine learning community often refers to this kind problem as *binary preference elicitation*, preference learning or learning to rank [Chu and Ghahramani, 2005; Fürnkranz and Hüllermeier, 2010].

Crucially, binary preference elicitation sidesteps most of the problems that eliciting preference on an absolute scale exhibits. Because no absolute scale is used, it does not matter if the scale of ratings used by different respondents are not aligned, as long as there is a mostly monotonic relationship between them. Also, the respondent only needs to be knowledgeable about the two items being compared in order to give an informed response. This way respondents with much more limited scope of knowledge can provide highly informative data.

Despite these convenient properties, pairwise preference learning has a drawback. When learning preferences over n items, there are $\mathcal{O}(n^2)$ potential pairs of items that we can ask the respondents to compare. Not only would querying all item-pairs take prohibitively long time, it is also unnecessary. Most binary choices would be predictable from previous choices the respondent or other respondents have already made.

Therefore, in binary preference learning, active learning and optimal experiment design have increased importance, and increased potential to improve over passive learning.

In this section I will extend the BALD framework developed for Gaussian process classification to work on preference learning. I will do this by showing how a popular non-parametric model for preference learning can be interpreted as binary Gaussian process classification with an special positive definite kernel between item-pairs, that I call *the preference kernel*. This idea can be used in other kernel-based classifiers such as SVMs. I present experimental results where I compare BALD to other active learning approaches to this problem.

6.6.1 Reduction to classification

In preference learning each data point describes two items, i and j, which have been presented to a human judge. It is assumed that the items are described by their numeric feature vectors $x_i \in \mathcal{X}$ and $x_j \in \mathcal{X}$ respectively. Each item is assumed to have a fixed number, $\dim(\mathcal{X}) = d$, of features associated with them. Each training data point also has a binary label $y \in \{-1, 1\}$ such that $y = 1$ if the user prefers item i to item j, and $y = -1$ otherwise. The primary goal of preference learning is to accurately predict the

direction of human preference for a new pair of feature vectors not seen before.

Pairwise preference learning is a special case of binary classification, inasmuch as the main goal is to predict a class label $y \in \{-1, 1\}$ given an input feature vector $(x_i, x_j) \in \mathcal{X}^2$, composed by concatenating features of i and j. However, using a generic classifier, such as an SVM or Gaussian process classifier would be highly inefficient, as these classifiers do not know about the symmetries inherent in the ranking problem. Firstly, if one observes (x_i, x_j) pair with a positive label, that implies the pair (x_j, x_i) would have a negative label. Furthermore, if one observes x_i is preferred to x_j and x_j is preferred to x_k, one would predict x_i is preferred to x_k. A generic classifier trained on pairs cannot make such deductions.

This problem is often addressed by introducing a latent preference function $f : \mathcal{X} \mapsto \mathbb{R}$ such that $f(x_i) > f(x_j)$ whenever the user prefers item i to item j and $f(x_i) < f(x_j)$ otherwise [Chu and Ghahramani, 2005]. This latent representation implies a pre-defined ordering of items. However, the observed data are not always consistent with a single fixed ordering, and sometimes are contradictory. Furthermore people's choices may be contaminated by noise, or be inaccurate because of lack of attention. To account for this randomness, the model presented by [Chu and Ghahramani, 2005] assumes that when respondents decide between options, their preference function is contaminated by evaluation noise.

When the evaluations of f are contaminated with Gaussian noise with zero mean and (without loss of generality) variance $1/2$, we obtain the following likelihood function for the underlying preference function f given the data point x_i, x_j and corresponding label y:

$$\mathbb{P}(y|x_i, x_j, f) = \Phi[(f[x_i] - f[x_j])y]\,, \tag{6.25}$$

where Φ is the standard Normal cumulative distribution function. The preference learning problem can be solved via Bayesian inference, combining a GP prior on f with the likelihood function in (6.25) [Chu and Ghahramani, 2005]. The posterior for f can then be used to make predictions on the user preferences for new pairs of items.

Note that the likelihood (6.25) depends only on the difference between $f(x_i)$ and $f(x_j)$. Let $g : \mathcal{X}^2 \mapsto \mathbb{R}$ be the latent function $g(x_i, x_j) = f(x_i) - f(x_j)$. We can recast the inference problem in terms of g and ignore f. When the evaluation of f is contaminated with standard Gaussian noise as before, the likelihood for g given x_i, x_j and y is

$$\mathbb{P}(y|x_i, x_j, g) = \Phi[g(x_i, x_j)y]\,. \tag{6.26}$$

Note this likelihood functon for g is the same as the probit classification likelihood defined in Eqn. (6.8). Since g is obtained from f through a linear operation, the GP prior on f induces a GP prior on g. The prior covariance function k_{pref} of the GP prior on g can be computed from the covariance function k of the GP on f as follows (assuming zero mean prior for f)

$$\begin{aligned}
k_{\text{pref}}((x_i, x_j), (x_k, x_l)) &= \mathbb{C}\text{ov}_{f\sim GP_k}\left[g(x_i, x_j), g(x_i, x_j)\right] && (6.27)\\
&= \mathbb{C}\text{ov}_{f\sim GP_k}\left[f(x_i) - f(x_j), f(x_k)(x_l)\right] && (6.28)\\
&= \mathbb{E}_{f\sim GP_k}\left[\left(f(x_i) - f(x_j)\right)\left(f(x_k) - f(x_l)\right)\right] && (6.29)\\
&= \mathbb{E}_{f\sim GP_k} f_{x_i} f_{x_k} + \mathbb{E}_{f\sim GP_k} f_{x_j} f_{x_l} - \mathbb{E}_{f\sim GP_k} f_{x_i} f_{x_l} - \mathbb{E}_{f\sim GP_k} f_{x_j} f_{x_k} && (6.30)\\
&= k(x_i, x_k) + k(x_j, x_l) - k(x_i, x_l) - k(x_j, x_k) && (6.31)
\end{aligned}$$

We call k_{pref} the *preference kernel*. This kernel function for preference learning is not new: the same kernel has been derived from a large margin classification viewpoint by Fürnkranz and Hüllermeier [2010]. However, to our knowledge, the preference kernel has not been used previously in Bayesian, Gaussian process-based models, only in our previous work [Houlsby et al., 2011, 2012]

The combination of (6.26) with a GP prior based on the preference kernel allows us to transform the pairwise preference learning problem into binary classification with GPs. This means that state-of-the-art methods for GP binary classification, such as expectation propagation [Minka, 2001b], can be applied readily to preference learning. Furthermore, the simplified likelihood (6.26) allows us to implement complex inference problems such as the multi-user hierarchical model described in [Houlsby et al., 2012].

Experiments

To test the performance of BALD on preference learning we use the *cpu, cart* and *kinematics* regression datasets from the LIACC Machine Learning Data Set Repository [**?**]. We processed each of these data sets to yield a binary preference task following the procedure described in in Chu and Ghahramani [2005]: Pairs of items were randomly selected and regression target values were used to decide the direction of preference to

Figure 6.4: Evaluation of the accuracy of active learning algorithms on various binary preference learning datasets. Methods used are random query (**??**), BALD (**??**), MES (**??**), QBC with 2 (QBC_2, **??**) and 100 (QBC_{100}, **??**) committee members, active SVM (**??**), IVM (**??**), decision theoretic [Kapoor et al., 2007] (**??**). For prediction, each method uses Gaussian process classification or SVM with the preference kernel from Eqn. (6.31). See text for analysis.

form a pool of training points for pool-based active learning. The performance was tested on a held-out test data set generated in a similar fashion.

Results on the three datasets are shown in Fig. 6.4. Surprisingly, despite our expectations, we found that random sampling from the pool provides reasonable performance, and active selection of measurements does not radically improve the speed of learning. Across the three data sets all studied methods perform well. BALD seems to struggle in the initial phase of learning on the *cpu* dataset, for which we did not find a convincing explanation. On the *kinematics* dataset the differnece between methods is more apparent, and here BALD provides the best performance among all methods studied.

In follow-up work we extended BALD to collaborative preference learning, where preference judgements are elicited from multiple experts [Houlsby et al., 2012]. The interested reader is referred to this paper for further experimental validation involving real-world preference datasets.

6.7 Summary and Conclusions

In this chapter I presented BALD, or Bayesian active learning by Disagreement, a practical algorithm for Bayesian active learning based on Shannon's mutual information. BALD exploits the symmetry of Shannon's mutual information and expresses the well known criterion in terms of the Shannon entropy of predictive distributions in the observation space. We have shown how BALD can be implemented to perform Bayesian active classification and preference learning in state-of-the art Gaussian process-based models in a computationally efficient way. Our experiments show that BALD compares favourably to other active learning methods, including the popular active SVMs [**?**].

Unfortunately, as BALD is built on the unique symmetry property of Shannon's mutual information, it does not generally apply to other instances of the scoring rule-based active learning framework described in Chapter 5. Indeed, among the examples of scoring rules given in Chapter 2, no other value of information functional is symmetric.

Chapter 7

Adaptive Bayesian Quantum Tomography

Summary of contributions: The work presented in this section is joint work with Neil M. T. Houlsby. Both authors contributed equally to all aspects of research. The work forms the basis of the journal article 'Adaptive Bayesian Quantum Tomography' [?].

7.1 Introduction

Quantum computing and quantum communication are rapidly exploding areas of modern computer science. Exploiting quantum physical phenomena such as entanglement and quantum teleportation, these computers and communication protocols radically extend the scope of efficient computing. Over the past couple of decades it has been shown that large classes of practically important algorithms, such as integer factorisation or discrete logarithm, can be implemented in polynomial time using quantum computers [Shor, 1994, 1997]. For these problems, no efficient classical (non-quantum) algorithms are known.

However, there is an important experimental limitation which is a barrier to progress towards studying large quantum computers: state reconstruction. At the hearth of this problem lies the fact that the end result of a quantum computation is not simply a series of bits, but instead 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 perform *measurements* on it.

Measurements in quantum physics have two surprising properties :

Firstly, the outcome of a quantum measurement is generally non-deterministic. That

is, even if the state of the system being measured is fully known, the outcome of the measurement is random. Therefore, in most cases, a single measurement doesn't provide full information about the state of the system, so repeated measurements and statistical inference are needed.

Secondly, a measurement destroys – or at least modifies – the quantum state of the system itself. The phenomenon is known as wave function collapse. This means that repeated measurements on the same system are pointless, as the first measurements destroy the state. These two properties of quantum measurements imply that one can (nearly) never obtain full information about a quantum state in any single experiment, because it is destroyed before sufficient information is extracted.

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 to 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 classically hard algorithms can be implemented using polynomial number of quantum operations, reading out the result can still take exponentially long.[1]

In current experimental quantum physics, when researchers study properties a new physical implementation of a quantum gate, they have to demonstrate that in multiple situations their equipment produces a state predicted by theory. Often due to noise and other environmental factors these implementations are imperfect and the produced state isn't exactly as desired. The degree of discrepancy between the theoretical predictions and the actually produced state is called infidelity. To be able to measure infidelity, 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 chapter I explore the possibility of speeding up quantum tomography by using state-of-the art active learning. I will first introduce the mathematical framework

[1]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

for studying quantum tomography, and provide a Bayesian analysis of the problem. Then I present an active learning method to perform quantum tomography, that we call Adaptive Bayesian Quantum Tomographu (ABQT). I present simulated experimental results that show that active learning can speed up the process of tomography by orders of magnitude.

7.2 A primer to quantum statistics

The central concept in quantum physics is the quantum state of a system. The state describes the system's behaviour when measurements are carried out. An example of a simple, two-dimensional quantum state is the polarisation state of a single photon. Photons are indeed one of the most widely used quantum physical model systems. Throughout this introduction I will use photons as an example to illustrate physical analogues of mathematical formalism. Other examples of quantum systems include spins of electrons, nuclear spins, quantum dot pairs, atomic spins. For recent reviews on the state of experimental quantum computing see [Ladd et al., 2010] and references therein.

Even a simple system like a photon can be in a variety of states. For example, the photon can be linearly polarised. The angle of linear polarisation can be either horizontal (denoted as $|H\rangle$) or vertical ($|V\rangle$) or any angle in between. Light can also be circularly polarised. The direction of circular polarisation can be left (denoted $|L\rangle$) right ($|R\rangle$), or anything in between. In addition, the system can be in any superposition of linear and circular polarisation which is generally called elliptic polarisation.

Mathematically, the polarisation state of a photon is fully described by a two-dimensional unit-length complex vector. In this chapter I use the so called Dirac or bra-ket notation [Dirac, 1939] for complex valued vectors $|x\rangle$, to clearly distinguish them from real valued vectors and variables that appeared in other sections of the thesis. In this notation $\langle x|$ denotes the conjugate transpose of $|x\rangle$, $\langle x|y\rangle$ the scalar product between $|x\rangle$ and $|y\rangle$. Between complex vectors the scalar product is defined as $\langle x|y\rangle = \sum_{d=1}^{D} x_d \overline{y_d}$, where $\overline{y_d}$ is complex conjugation.

A two-dimensional unit-length complex vector $|\phi\rangle$, which one can parametrise using three continuous parameters θ, a_x and a_y as follows:

$$|\phi\rangle = \begin{pmatrix} \cos(\theta)\exp(ia_x) \\ \sin(\theta)\exp(ia_y) \end{pmatrix} \tag{7.1}$$

This complex vector may represent the quantum state of the photon. Parameter θ can be interpreted as the angle of linear polarisation relative to horizontal, with $\theta = 0^\circ$

meaning horizontal, $\theta = 90°$ vertical polarisation:

$$|H\rangle = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \tag{7.2}$$

and

$$|V\rangle = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \tag{7.3}$$

Any real valued linear combination of these two basis vectors (as long as unit-norm constraint is preserved) describes a general linearly polarised state at a certain angle. To represent circular polarisation, we need to use complex vectors:

$$|R\rangle = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ i \end{pmatrix} \tag{7.4}$$

and

$$|L\rangle = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ -i \end{pmatrix} \tag{7.5}$$

7.2.1 Measurements

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 (like the ones used in displays). Mathematically, the polarising filter is described by a two-dimensional unit-length complex vector $|\phi\rangle$, similarly to the quantum state. Depending on the state of the photon $|\phi\rangle$ and the measurement describing the filter $|\psi\rangle$, 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.

$$\mathbb{P}\left(\text{pass} \mid |\phi\rangle, |\psi\rangle\right) = |\langle\psi|\phi\rangle|^2 \qquad = \text{tr}\left(|\phi\rangle\langle\phi|\right)\left(|\psi\rangle\langle\psi|\right) \tag{7.6}$$

$$\mathbb{P}\left(\text{bounce} \mid |\phi\rangle, |\psi\rangle\right) = 1 - |\langle\psi|\phi\rangle|^2 \qquad = \left(\text{tr}\,|\phi\rangle\langle\phi|\right)\left(I - |\psi\rangle\langle\psi|\right) \tag{7.7}$$

The outcome probabilities are expressed as straightforward quadratic forms, and the

unit norm constraint ensures that the outcome probabilities are non-negative and sum to one.

Let us look at the outcome probabilities in concrete examples.

Example 7.2.1. A linearly polarised photon with polarisation angle $\theta = 30°$ is measured using a horizontally polarised filter. The photon's state is $|\phi\rangle = \cos(\theta)|H\rangle + \sin(\theta)|V\rangle$. The filter is described as $|\psi\rangle = |H\rangle$. Since $\langle H|V\rangle = 0$, we can see that the probability that the photon passes the filter is

$$\mathbb{P}\left(\text{pass} \mid |\phi\rangle, |\psi\rangle\right) = \cos(\theta)\langle H|H\rangle + \sin(\theta)\langle H|V\rangle = \cos(\theta) = \frac{\sqrt{3}}{2} \tag{7.8}$$

Example 7.2.2. A linearly polarised photon with polarisation angle $\theta = 30°$ from the horizontal is measured using a right circularly polarised filter. The photon's state is $|\phi\rangle = \cos(\theta)|H\rangle + \sin(\theta)|V\rangle$. The filter is described as $|\psi\rangle = |R\rangle$. Using Eqns. (7.2),(7.3) and (7.4) one can see that $\langle V|R\rangle = 1/\sqrt{2}$, $\langle H|R\rangle = i/\sqrt{2}$. Thus, the probability of the photon passing through is

$$\mathbb{P}\left(\text{pass} \mid |\phi\rangle, |\psi\rangle\right) = \frac{1}{2}\left|\cos(\theta) + i\sin(\theta)\right|^2 = \frac{1}{2}. \tag{7.9}$$

So the photon passes or bounces back completely randomly with a probability 0.5 irrespective of the angle of the linear polarisation of the photon. Thus, performing a measurement using a circular polarisation filter provides no information about the linear polarisation of the photon.

Crucially, measuring a quantum system alters the state. This phenomenon is called wave function collapse. In quantum tomography we assume that after only a single measurement on a system, its state gets completely destroyed and we cannot use it anymore for the purposes of inference. After each measurement, once the outcome is recorded, the measured system is discarded, and a new, independent copy of the system is generated and measured.

There are alternative approaches that use sequences of measurements which only partially destroy the state; these approaches are referred to as continuous weak measurements and quantum control [Smith et al., 2006]. Weak measurements also have a high importance in quantum cryptanalysis [Pryde et al., 2004].

7.2.2 Density matrices

In the previous paragraphs we have seen that quantum measurements are inherently non-deterministic. But in some cases there is another source of uncertainty effecting the

outcome of measurements. These can be due to environmental noise, lack of information, and various other reasons. We will call these other sources of uncertainty classical uncertainty, as opposed to quantum uncertainty which is the inherent uncertainty in measurements. When both kinds of uncertainty are present, we will say that the quantum system is in a *mixed state*. The states we have discussed so far, that is when there is no classical uncertainty present are called *pure states*.

Example 7.2.3 (Observationally equivalent sources)**.** Consider two noisy sources of photons. Source A produces a horizontally polarised photon with probability 0.5 or a vertically polarised one with probability 0.5. Source B produces the state $\frac{1}{\sqrt{2}}|H\rangle + \frac{1}{\sqrt{2}}|V\rangle$ with probability 0.5 or $\frac{1}{\sqrt{2}}|H\rangle - \frac{1}{\sqrt{2}}|V\rangle$ with probability 0.5. These correspond to 45° and −45° polarisation angles respectively.

Let us see what happens if we perform a measurement $\langle\psi|$ on the two noisy systems. First, for system A the probability is given by

$$\mathbb{P}_A\left(\text{pass}|\ |\psi\rangle\right) = \frac{1}{2}\langle\psi|H\rangle^2 + \frac{1}{2}\langle\psi|V\rangle^2 \tag{7.10}$$

For system B we get

$$\mathbb{P}_B\left(\text{pass}|\ |\psi\rangle\right) = \frac{1}{2}\left(\frac{1}{\sqrt{2}}\langle\psi|H\rangle + \frac{1}{\sqrt{2}}\langle\psi|V\rangle\right)^2 + \frac{1}{2}\left(\frac{1}{\sqrt{2}}\langle\psi|H\rangle - \frac{1}{\sqrt{2}}\langle\psi|V\rangle\right)^2 \tag{7.11}$$

$$= \frac{1}{2}\langle\psi|H\rangle^2 + \frac{1}{2}\langle\psi|V\rangle^2 + \frac{1}{2}\langle\psi|H\rangle\langle\psi|V\rangle - \frac{1}{2}\langle\psi|H\rangle\langle\psi|V\rangle \tag{7.12}$$

$$= \frac{1}{2}\langle\psi|H\rangle^2 + \frac{1}{2}\langle\psi|V\rangle^2 \tag{7.13}$$

In both cases the probabilities of observing the photon passing the filter or bouncing back are exactly the same. The probability is only a function of the measurement, but it does not depend on which source the photon came from. So no matter what measurement we make, the two sources are indistinguishable experimentally. I will call these two sources *observationally equivalent*.

Incidentally, because $1/2\langle\psi|H\rangle^2 + 1/2\langle\psi|V\rangle^2 = 1/2$, the measurement outcome does not even depend on the measurement.

In more general terms, when classical and quantum uncertainty are both present in a system, they cannot be disambiguated by measuring the system. There are examples of quantum systems that cannot be distinguished experimentally. This concept is intimately related to the issue of non-identifiable likelihood models [Teicher, 1961] in statistics. However, we can define equivalence classes of systems, and in this case the

goal of quantum tomography becomes to infer the equivalence class, or *the mixed state* of the system. These mixed states can be parametrised via the so called density matrix ρ [Fano, 1957].

Definition 12 (Density matrix)**.** Consider a random quantum system that is in pure state $|\phi\rangle$ with probability $P_{|\psi\rangle}$. The density matrix describing this system is defined as the following linear operator:

$$\rho = \mathbb{E}_{|\psi\rangle\sim P}|\phi\rangle\langle\phi| \tag{7.14}$$

The density matrix is Hermitian and has unit trace.

As we have seen, the two noisy systems in Example 7.2.3 were equivalent, and indeed they both are described by the same density matrix $\rho = \frac{1}{2}I$. In the context of photon sources, a light source, whose density matrix is $\rho = \frac{1}{2}I$ is called *unpolarised*. There are several 'different', but observationally equivalent unpolarised light sources, and in fact, randomly mixing unpolarised light with unpolarised light remains unpolarised.

Pure states are special cases of mixed states. A system with pure state $|\phi\rangle$ can be described by its density matrix $\rho = |\phi\rangle\langle\phi|$. This follows from the definition, using a probability distribution that concentrates on the single pure state $|\phi\rangle$. Density matrices corresponding to pure states are therefore always rank-one.

Similarly to mixed states, measurements also have a more general form. The most general class of measurements considered in this thesis are Positive Operator Valued Measures (POVMs). A POVM is defined by a set, $\mathbb{M}$, of Hermitian operators M_γ, indexed by possible outcomes $\gamma \in \{1, \ldots, \Gamma\}$, satisfying $\sum_{\gamma=1}^{\Gamma} M_\gamma = I$. These Hermitian operators determine the probability of observing outcome γ in configuration α when the measured system is in state ρ via Born's rule:

$$\mathbb{P}(\gamma|\rho, \mathbb{M}) = \mathrm{tr}\{M_\gamma \rho\}$$

The condition $\sum_{\gamma=1}^{\Gamma} M_\gamma = I$ ensures that the probabilities sum to 1, and they are always non-negative because all operators involved are Hermitian.

So far we have only considered probably the simplest two-dimensional model system, the polarised photon. In quantum computing, two-dimensional quantum systems are called qubits. As a generalisation of horizontal and vertical polarisation $|H\rangle$ and $|V\rangle$ the notation $|0\rangle$ and $|1\rangle$ is often used to denote the two most important states.

For general D-dimensional systems, ρ is a $D\times D$ Hermitian matrix, satisfying $\mathrm{tr}\,\rho = 1$. For m-qubit systems used in quantum computers $D = 2^m$, and the notation $|000\rangle$, $|001\rangle$, $\ldots|111\rangle$ is used to denote important states. Importantly, in multipartite systems

(systems that are composed of multiple parts) the components of the system

7.3 Quantum Tomography as Bayesian Inference

Quantum state tomography involves determining from experimental data the quantum state, ρ, of a system by performing measurements on several identical copies. As ρ is Hermitian and have unit trace, $D^2 - 1$ real degrees of freedom must be estimated. Note that the number of degrees of freedom is exponential in the number of qubits in a multi-qubit system.

The apparatus for a tomographic experiment may be configured to perform several different measurements; we use $\alpha \in \mathcal{A}$ to index all accessible configurations. Each measurement configuration α is characterised by a positive operator-valued measure (POVM). For each configuration, a measurement results in observing one of a finite number, Γ, of distinguishable outcomes. A POVM is defined by the set, $\mathbb{M}_\alpha$, of Hermitian operators $M_{\alpha\gamma}$, indexed by possible outcomes $\gamma \in \{1, \ldots, \Gamma\}$, satisfying $\sum_{\gamma=1}^{\Gamma} M_{\alpha\gamma} = I$. These POVMs jointly constitute our tomographic model $\mathcal{M} = \{\mathbb{M}_\alpha : \alpha \in \mathcal{A}\}$ and determine the probability of observing outcome γ in configuration α when the measured system is in state ρ via Born's rule:

$$\mathbb{P}\left(\gamma|\rho, \alpha; \mathcal{M}\right) = \mathrm{tr}\left\{M_{\alpha\gamma}\rho\right\}$$

State reconstruction has been approached with several methods, the most popular being maximum likelihood estimation (MLE). MLE finds a physically feasible state ρ that is most likely to have produced the observed data, $\mathcal{D}$, by maximising the likelihood:

$$\mathcal{L}(\rho; \mathcal{D}) = \prod_{n=1}^{N} \mathbb{P}\left(\gamma_n|\rho, \alpha_n\right) = \prod_{\alpha\in\mathcal{A}} n_\alpha! \prod_{\gamma=1}^{\Gamma} \frac{\mathrm{tr}\{M_{\alpha\gamma}\rho\}^{n_{\alpha\gamma}}}{n_{\alpha\gamma}!} \tag{7.15}$$

where n_α is the number of times configuration α was used, $n_{\alpha\gamma}$ is the number of times outcome γ was observed in configuration α. All probabilities are conditional on $\mathcal{M}$, for brevity this is omitted. A well-known drawback of MLE is that it often yields rank-deficient estimates, and thus assigns zero predictive probability to certain observations [Blume-Kohout, 2010]. This seems an unreasonable conclusion on the basis of a finite sample. This phenomenon is not unusual in maximum likelihood methods, and is related to *over-fitting* in statistics. Additionally, MLE provides no measure of uncertainty in its point estimate.

More sophisticated methods for quantum tomography use Bayesian inference and

suffer from neither of these problems [Blume-Kohout, 2010, and refs.]. In Bayesian inference a prior probability density, $p(\rho)$, over feasible states is specified. This prior is then augmented with the likelihood from Eqn. (7.15) using Bayes' rule to yield a posterior distribution:

$$p(\rho|\mathcal{D}) \propto \mathcal{L}(\rho;\mathcal{D})p(\rho) \tag{7.16}$$

Should we want a point estimate, we may report, say, the Bayesian mean estimate (BME) which is known to maximise expected operational divergences [Blume-Kohout and Hayden, 2006; Blume-Kohout, 2010]. But importantly, Bayesian inference also provides *error bars*, and more: the posterior captures richly our remaining uncertainty in the true state having seen the data $\mathcal{D}$.

For Bayesian inference one has to provide the prior $p(\rho)$, which is typically chosen to be non-informative or uniform. Here we adopt the representation and prior introduced in [Blume-Kohout, 2010], that treats the original system of interest as part of a larger, $D \times K$ dimensional bipartite system. Our prior over the mixed state ρ is then defined as the measure induced by the uniform (Haar) measure over pure states in $D \times K$ dimensions. It is easy to see that, tracing out[1] the K dimensional ancillary part leaves us with a prior distribution that concentrates on rank-K density matrices. By tuning parameter K one can trade off between computational efficiency and estimation accuracy, in a similar manner to compressed sensing [Gross *et al*, 2010].

Unfortunately, normalisation of the posterior distribution (Eqn. (7.16)) becomes analytically intractable, and therefore we have to approximate it. Several MCMC approaches have been suggested in this context [Blume-Kohout, 2010, and refs. therein]. These methods require evaluation of the full likelihood (7.15), which has $\mathcal{O}(n)$ cost with the number of different configurations used so far.

If we want to implement active learning, inference has to be performed after each measurement or after mini-batches of measurements. Therefore, MCMC methods whose cost increases with the number of measurements are inappropriate. To address this problem we developed a fast sequential importance sampling (SIS) algorithm, with $\mathcal{O}(1)$ likelihood evaluation cost. See [Doucet et al., 2001] for a thorough overview of these techniques.

In SIS, one keeps track of a number of samples, often called particles, ρ_s, $(s = 1 \dots S)$ and corresponding weights w_s, $(\sum_s w_s = 1)$ which are updated sequentially, every time

[1] 'tracing out' is an operation on the density matrix of a multi-partitite quantum system, which is analogous to computing the marginal of multivariate probability distributions. For further details please refer to [Blume-Kohout, 2010].

a new measurement is made. Assume that after n measurements, having observed data $\mathcal{D}_n$, the particles and weights $w_s^{(n)}$ constitute an approximation to the posterior:

$$p(\rho|\mathcal{D}_n) \approx \sum_{s=1}^{S} w_s^{(n)} \delta(\rho - \rho_s) \tag{7.17}$$

Using this approximation, and Bayes' rule, one can derive an approximation to the next posterior, after observing a new outcome γ_{n+1} in configuration α_{n+1}, as:

$$\begin{aligned} p(\rho|\alpha_{n+1}, \gamma_{n+1}, \mathcal{D}_n) &= \frac{\mathbb{P}(\gamma_{n+1}|\rho, \alpha_{n+1}) p(\rho|\mathcal{D}_n)}{\int \mathbb{P}(\gamma_{n+1}|\rho, \alpha_{n+1}) p(\rho|\mathcal{D}_n) d\rho} \\ &\approx \sum_{s=1}^{S} \underbrace{\frac{\mathbb{P}(\gamma_{n+1}|\rho_s, \alpha_{n+1}) w_s^{(n)}}{\sum_{r=1}^{S} \mathbb{P}(\gamma_{n+1}|\rho_r, \alpha_{n+1}) w_r^{(n)}}}_{w_s^{(n+1)}} \delta(\rho - \rho_s) \end{aligned} \tag{7.18}$$

The new weights $w_s^{(n+1)}$ are the re-normalised product of our current weights $w_s^{(n)}$ and observation probabilities $\mathbb{P}(\gamma_{n+1}|\rho_s, \alpha_{n+1})$. This update is fast, and only requires computing one term of the full likelihood, thus its complexity is independent of how many configurations have been tried before. This computational efficiency comes at a price; as time progresses, several weights decay to almost zero, and thus the quality of our approximation drops. This issue can be detected and handled by monitoring the effective sample size and resampling appropriately [Doucet et al., 2001].

7.4 Active learning in Quantum Tomography

Most existing approaches to optimal experiment design for quantum tomography determine, prior to collecting data, an optimal set of measurements to be used throughout the experiment. In this sense, whenever they exist, mutually unbiased bases (MUBs) are known to be optimal [Adamson and Steinberg, 2010; Wootters and Fields, 1989]. Research since has focused mainly on proving or disproving existence of, and implement MUBs in various dimensions [Adamson and Steinberg, 2010; Raynal et al., 2011; Yan et al., 2010]. Other work, [Kosut et al., 2004; Nunn *et al*, 2010] considered OED based on the Cramér-Rao bound. Here we argue that these approaches, including MUBs, provide only a partial solution to the problem of optimal experiment design inasmuch as they do not take partial data into account. If we are allowed to revise our choice of measurements during the experiment based on data collected so far, we may be in a better

position to reduce redundancy. This strategy is generally known as active learning or adaptive sampling; such adaptive experimental design have been applied in a number of other fields, such as clinical trails [Berry, 2006], cognitive science [Cavagnaro et al., 2010] and computer vision [Vondrick et al., 2011]. In physics, this approach has been referred to as self-learning measurements [Fischer et al., 2000; Hannemann et al., 2002]. However, due to the expensive computations that are involved, these methods have been restricted to two dimensional pure quantum states, or very few measurements. Recently advances in Bayesian methods allow us to build a fast, online algorithm that allows self-learning in arbitrary dimensions with many measurements.

Here we propose a new algorithmic framework that we call *Adaptive Bayesian Quantum Tomography* (ABQT), that builds on full Bayesian inference and Shannon information. To achieve adaptivity in practice, we need a fast algorithm for performing Bayesian state reconstruction from partial data after each measurement. Current sampling methods such as in [Blume-Kohout, 2010] are inappropriate as their costs increase with the number of measurement configurations tried so far. As a solution, we present a sequential importance sampling scheme [Doucet et al., 2001], that does not suffer from this. We then use the developed algorithm in conjunction with an information theoretic objective to adaptively optimise measurements. We assess the relative performance of our adaptive method in Monte Carlo simulations of qubit systems, and demonstrate a ten-fold reduction in the number of measurements needed for full tomography of two-qubit pure states. We also investigate the trade off between entangling and separable measurements in multipartite systems. Our central finding is that via adaptive tomography one can achieve, and even surpass, the statistical efficiency of MUB tomography using only separable measurements, that require experimental apparatus that is substantially easier to build using current technology.

Having discussed our method for estimating the state based on partial data, we now turn to the problem of optimal experiment design. Different state determination schemes have different OED strategies associated with them. Maximum likelihood methods usually use some form of the Cramér-Rao bound [Kosut et al., 2004; Nunn *et al*, 2010]. Bayesian experiment design on the other hand is based on Shannon information [Patra, 2007; Wootters and Fields, 1989]. The posterior characterises our remaining uncertainty in the parameter, and this uncertainty can be quantified using Shannon's entropy. A sensible aim is to pick an experimental configuration α, such that after observing the

outcome γ, the entropy $\mathbb{H}$ of the new posterior is reduced the most:

$$\underset{\alpha\in\mathcal{A}}{\operatorname{argmax}} \left\{ \mathbb{H}\left[p(\rho|\mathcal{D})\right] - \mathbb{E}_{p(\gamma|\alpha,\mathcal{D})}\left[\mathbb{H}\left[p(\rho|\gamma,\alpha,\mathcal{D})\right]\right] \right\} \tag{7.19}$$

The expectation with respect to γ is needed as the measurement outcome is unknown *a priori*. This objective naturally allows us to address the question 'Having seen the outcome of the first few measurements, which measurement should we carry out next?' Rather, it was used to determine a single best set of measurements which are then uniformly sampled throughout the experiment [Patra, 2007; Wootters and Fields, 1989]. Under these circumstances mutually unbiased bases (MUBs) are optimal, whenever they exist. We exploit the dependence of Eqn. (7.19) on past observations, and allow for measurements to be re-optimised adaptively as the experiment progresses.

However, Eqn. (7.19) is impractical to work with directly, as it involves computing entropies of high-dimensional intractable posterior densities. Recall that we approximate our posterior by samples, with which it is notoriously hard to estimate differential entropies. Furthermore, in Eqn. (7.19) the posterior has to be re-computed for every possible outcome γ. Therefore, instead of working with Eqn. (7.19) directly, we propose to use an equivalent reformulation thereof in terms of predictive distributions [Patra, 2007]:

$$\underset{\alpha\in\mathcal{A}}{\operatorname{argmax}} \left\{ \mathbb{H}\left[\mathbb{P}(\gamma|\alpha,\mathcal{D})\right] - \mathbb{E}_{p(\rho|\mathcal{D})}\left[\mathbb{H}\left[\mathbb{P}(\gamma|\alpha,\rho)\right]\right] \right\} \tag{7.20}$$

In previous studies [Fischer et al., 2000] the system is limited to pure single qubit states, calculating the intractable Bayesian normalising constant can be realised with simple numerical integration; this could not be extended easily to higher dimensions. They consider two active learning algorithms: firstly, uncertainty sampling, which uses an approximate version of Eqn. (7.20), where the second term was ignored. This arguably leads to suboptimal selection behaviour; the experimenter's uncertainty may be confounded with inherent uncertainty of quantum measurements.

Within the Bayesian framework, one could use other measures of uncertainty about the Quantum state; we note that minimising the Shannon Entropy (Eqn. (7.19)) is equivalent to minimisation of the Bayes risk when one uses the log loss to evaluate probabilistic estimate of the state [Dawid, 2007]. One could construct a number of other loss functions - the second algorithm in [Fischer et al., 2000] seeks to minimise the Bayes risk, but using fidelity as the loss function. Although this loss is theoretically attractive, only the log loss allows the particular analytic reformulation to Eqn. (7.20)

that permits efficient online computations. Other loss functions require one full posterior update for every possible outcome for each measurement under consideration, ABQT requires only one update per complete cycle. Therefore, if one does not use log loss online computation is infeasible; in [Hannemann et al., 2002] experimental designs for all 2^N possible experimental outcome successions are pre-computed, they are therefore limited to very short experiments (< 20 measurements). Combining Eqn. (7.20) with our SIS Bayesian update scheme allows for fast online experimental design.

The equivalence between Eqns. (7.19) and (7.20) becomes clear realising that they both express the conditional mutual information between ρ and γ. Eqn. (7.20) offers computational advantages over Eqn. (7.19): it only involves computing discrete entropies $\mathbb{H}\left[\mathbb{P}(\gamma|\alpha,\rho)\right]$ and expectations of these under the posterior. This objective function is generally non-convex in α, but its value - and derivatives with respect to α - can now be efficiently computed using our weighted posterior samples from Eqn. (7.17), allowing us to find the most informative α by direct optimisation. It is interesting to note that Eqn. (7.20) is equivalent to the Jensen-Shannon divergence of the class of predictive distributions, with the posterior being the mixing distribution.

In summary, we propose the following algorithm, called Adaptive Bayesian Quantum Tomography. After each single measurement, ABQT updates its approximate posterior using Eqn. (7.18), then chooses the next measurement by direct numerical maximisation of the information theoretic objective in Eqn. (7.20).

7.5 Results

7.5.1 single qubit tomography

In our first simulated experiments we study tomography of single qubits ($D = 2$). Mixed state qubits have three real degrees of freedom, ρ is represented as a point in a unit ball, called the Bloch sphere. For illustration purposes we first omit the third component, and only infer two remaining parameters, which will lie in a unit (Bloch) disk. This corresponds to e. g. determining linear polarisation of a photon, assuming that the circular polarisation is zero. We allow for arbitrary projective measurements with binary ($\Gamma = 2$) outcomes. These are represented by pairs of antipodal points on the perimeter of the Bloch disk. Now $\alpha \in [0, \pi)$ codes for the orientation. Fig. 7.1 shows the progression of measurement bases chosen by ABQT. The first two measurements are mutually unbiased, however, the third measurement is equally biased with respect to both previous bases, demonstrating that using a fixed MUB set is suboptimal in the adaptive

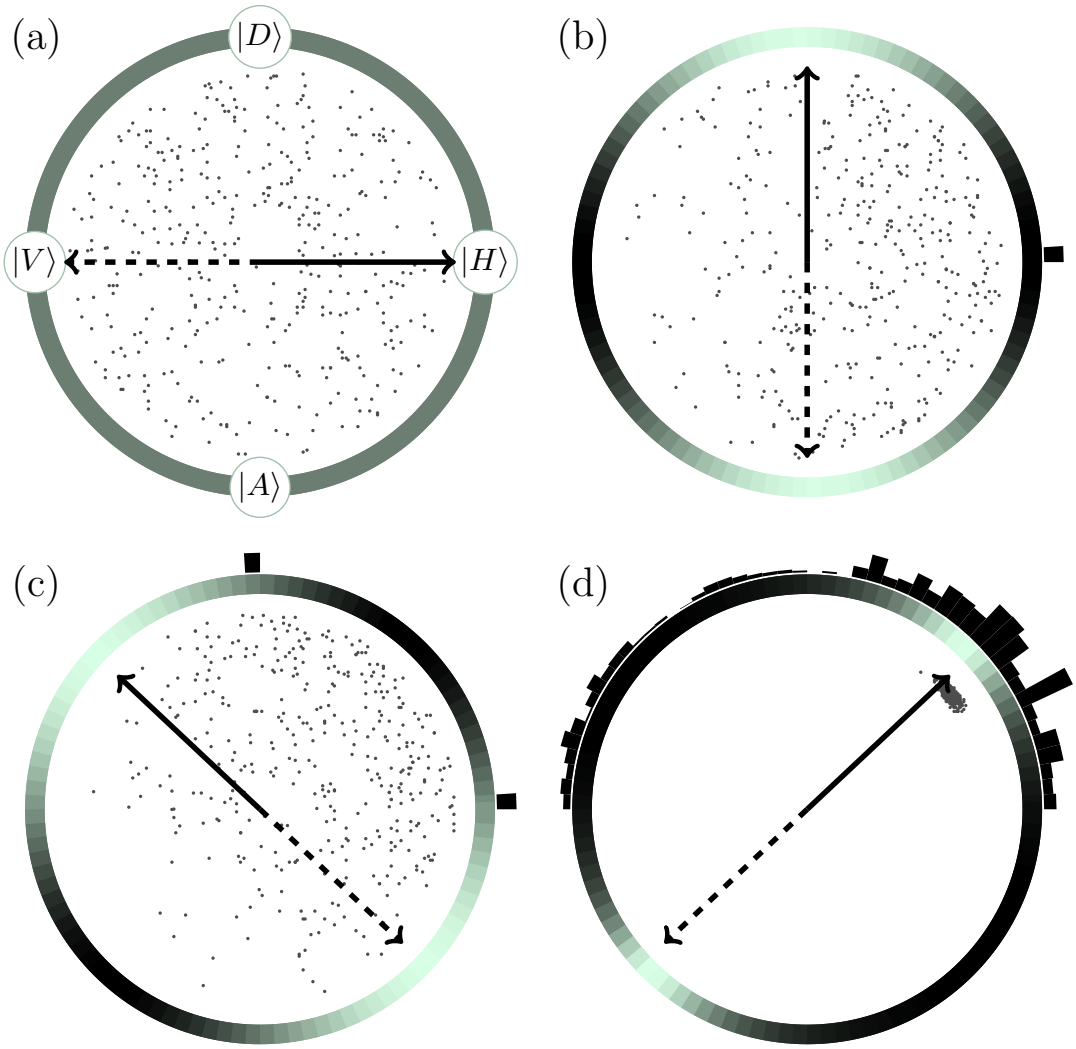

Figure 7.1: Adaptive selection of measurements based of partial data. Scatter plots show 400 samples from current posterior. Shaded circles around the 'Bloch disk' show relative value of the objective in Eqn. (7.20) for different measurement directions (lighter is higher). Pairs of arrows show the most informative next measurement. Circular histograms show the number of times measurement directions have been used. **(a)** Initially, no observations are made, samples shown are from the uniform prior. All measurements are equally informative, we chose to start with $\{|H\rangle, |V\rangle\}$. **(b)** After one measurement, the posterior is updated, the next best measurement is mutually unbiased w.r.t. the first one. It is now $\{|D\rangle, |A\rangle\}$. **(c)** After two observations, the next best measurement is equally biased to the first two bases. **(d)** Posterior after 1000 observations concentrates around true state. The method tries a range of measurements, with a tendency to point towards the solution.

Figure 7.2: One qubit tomography using projective measurements. **(a)** Improvement of mean posterior fidelity as the experiment progresses. Results are shown for uniformly sampled measurements (**??**), uniformly sampled Pauli measurements (**??**), ABQT selecting adaptively amongst the 3 Pauli measurements (**??**) and ABQT picking general measurements (**??**). Adaptive optimisation of measurements allows for an almost n^{-1} rate of convergence, while other methods are more consistent with a $n^{-\frac{1}{2}}$ rate. **(b)** Final value of the mean posterior infidelity after 6000 measurements using the four methods as before, as a function of purity of the state to be estimated. The advantage of ABQT is greatest for purer states.

Figure 7.3: Two qubit QST with uniformly chosen amongst MUB (**??**) or SSQT bases (**??**) and AB QT picking from the same set of MUBs (**??**), SSQT bases (**??**) or a more flexible set of 81 separable bases (**??**). Cases (a)-(c) are the same as those in [Adamson and Steinberg, 2010], (d) shows average results over 20 randomly generated entangled pure states. **(a)** As expected, for the maximally mixed state the choice of measurement strategy has little effect. **(b)** On the entangled state $(|HH\rangle + |VV\rangle)/\sqrt{2}$ MUB outperforms SSQT when uniformly sampled, but by allowing for adaptivity we can close the performance gap. **(c)** SSQT outperforms MUBs on the separable state $|HV\rangle$, but again, picking measurements adaptively the two sets perform similarly. **(d)** For random pure states a large improvement in performance is made when performing ABQT with the flexible set of separable measurements. Using this set, ABQT only needs 10^4 measurements to achieve $\approx 98.7\%$ mean fidelity for which MUB needs 10^5 measurements.

framework. Throughout the rest of the experiment the algorithm explores a wide range of measurements.

Fig. 7.2 shows that the posterior mean fidelity - this time inferring all three coordinates in the full Bloch sphere - improves at a faster rate when measurements are adaptively optimised. We quantify performance as mean posterior fidelity, rather than the fidelity of the Bayesian mean estimate, as the latter gives no indication of the confidence in our estimate. The rate is more consistent with a n^{-1} law rather than $n^{-\frac{1}{2}}$ as predicted for non-adaptive methods [Adamson and Steinberg, 2010, and refs.]. Fig. 7.2.b shows a larger advantage for states of high purity (defined as sum of squared eigenvalues).

7.5.2 Separable vs. MUB tomography of two qubits

In multipartite systems, such as m-qubit registers, there are two fundamentally different classes of measurements one can apply: separable or entangling. Separable tomographic

experiments are straightforward and cheap to implement, while entangling measurements are statistically more powerful. Notably, entanglement is required for implementing MUBs. These differences are discussed extensively in [Adamson and Steinberg, 2010]. To investigate this trade-off in the light of adaptive tomography, we reproduce and extend the experiments in [Adamson and Steinberg, 2010]. Results are shown in Fig. 7.3. Notably, all substantial differences between MUB and standard separable tomography (SSQT) vanish as we allow for adaptivity (Fig. 7.3.a–c). Furthermore, for random pure states, we are able to realise a ten-fold improvement over MUBs when using flexible separable measurements (Fig. 7.3.d). The results indicate that allowing for adaptivity with an imperfect, but flexible set of measurements offers greater advantages than using a fixed set of MUBs.

7.5.3 Conclusions and outlook

In summary, we have presented a new adaptive optimal experimental design framework and method based on Bayesian inference and Shannon's information. We showed that mutually unbiased bases, widely accepted as *the* optimal measurements, represent only a partial solution and are suboptimal in the adaptive framework. Moreover, the adaptive framework applies regardless of dimensionality, and can be applied to spaces where MUBs do not even exist [Patra, 2007; Raynal et al., 2011]. This motivates a shift in experimental focus from implementing complex entangling measurements to implementing quickly reconfigurable simpler measurements. In quantum optics, this could be feasibly achieved via mechanically or electronically controlled liquid crystal wave plates.

Although our algorithm demonstrated a substantial leap forward in terms of empirical performance, it is important to keep in mind that it still does not resolve the curse of dimensionality: the size of the parameter space still scales exponentially with the number of qubits in question. Other successful approaches address the question of dimensionality by restricting the search space. Compressed sensing [Gross *et al*, 2010] constrains estimation onto a lower-dimensional manifold of rank-deficient states. It is even possible to carry out quantum homodyne tomography in infinite dimensional spaces, assuming the Wigner function is infinitely differentiable and falls into a particular smoothness class [Butucea et al., 2007]. These simplifying assumptions and smoothness constraints can be incorporated into a Bayesian framework via priors.

Chapter 8

Conclusions

Machine learning is a mess. This thesis finally brings order and clarity to what otherwise appears as a mix of muddled thoughts. Just kidding.

Machine learning is fragmented. Researchers in the field publish and endless stream of ideas, methods, optimisation problems and probabilistic models. Substantial progress is made when, from this system of thoughts, unifying frameworks and generalisations emerge. Unifying views help researchers establish connections between concepts and techniques they previously thought were unrelated, and develop new ones in a more systematic fashion.

A good example of this unification process is the emergence of probabilistic generative models. By the early 90's, several unsupervised learning algorithms existed. Principal components analysis, Gaussian mixture models, slow feature analysis, Kalman-filters, autoencoders, independent component analysis, Boltzmann machines, factor analysis, hidden Markov models and others co-existed for years or decades, but each of these methods were by large studied in isolation. The relationship between these methods became clear when a unifying framework, based on probabilistic generative models emerged [Lauritzen, 1996; Roweis and Ghahramani, 1999; Tipping and Bishop, 1999; Turner and Sahani, 2007].

Over the past decades, graphical models [Lauritzen, 1996] have become the standard language for describing complicated, hierarchical probabilistic models. Each new probabilistic model is explained in terms of conditional independence statements represented graphically. The movement introduced the useful distinction between models and algorithms: inference methods are usually developed for general classes of graphical models, rather than in the context of a particular model. The relationship between models is now clear, and it became relatively straightforward to adopt probabilistic models to new

problems.

I hope this thesis contributes to the unification of machine learning by presenting a unifying framework for Bayesian machine learning problems based on scoring rules and information geometry. The unifying view allows one to uncover relationships between seemingly unrelated methods. There are four main connections presented in this thesis.

Score matching, approximate inference and active learning Information geometry provides a common basis for score matching, approximate Bayesian inference and Bayesian active learning. Scoring-rule-based divergences play a central role in all three of these problems. In score matching and approximate inference one tries to minimise divergences, in Bayesian active learning the goal is to maximise them.

Maximum mean discrepancy and kernel scoring rule A minor contribution in chapter 2 is the connection between maximum mean discrepancy [Gretton et al., 2012] and the kernel scoring rule [Jose et al., 2008]. These related concepts have been developed studied by two distinct communities, derived from different first principles. Establishing this connection allowed me to define a new concepts, such as the *kernel spherical divergence* in Eqn. (2.112) and the *kernel value of information* in Eqn. (2.83).

Kernel herding, Bayesian quadrature and loss-calibrated quasi-Monte Carlo Herding was originally introduced by Welling [2009] as a heuristic procedure obtained by "taking the zero temperature limit of the corresponding maximum likelihood problem". Although the paper studied several properties of this method, it was unclear whether and how herding is related to existing methods in machine learning. In [Huszár and Duvenaud, 2012] we showed that herding is closely related to Bayesian quadrature. In chapter 4 I also show that it is a special case of the general procedure I call loss-calibrated quasi-Monte Carlo.

Bayesian optimisation, Bayesian quadrature and active learning In chaper 5 I introduce a unifying framework for Bayesian active learning. This unifying framework naturally accommodates Bayesian optimisation, Bayesian quadrature, information theoretic active learning and transductive learning. Scoring rules provide a unifying language for these models inasmuch as they can all be described via the scoring rule they use to quantify the usefulness of the Bayesian posterior.

In addition to theoretical contributions, the thesis also presents applications of state-of-the-art information theoretic active learning to binary classification, preference elici-

tation and quantum tomography.

References

R. B. A. Adamson and A. M. Steinberg. Improving quantum state estimation with mutually unbiased bases. *Phys. Rev. Lett.*, 105(3):030406, 2010. doi: 10.1103/PhysRevLett.105.030406. 124, 129, 130

S. Amari and A. Cichocki. Information geometry of divergence functions. *Bulletin of the Polish Academy of Sciences: Technical Sciences*, 58(1):183–195, 2010. 8

F. Bach, S. Lacoste-Julien, and G. Obozinski. On the equivalence between herding and conditional gradient algorithms. *Arxiv preprint arXiv:1203.4523*, 2012. 67, 70, 83

K. Bache and M. Lichman. UCI machine learning repository, 2013. URL `http://archive.ics.uci.edu/ml`. 106, 107

James O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer, New York, 1985. 57

D.A. Berry. Bayesian clinical trials. *Nat. Rev. Drug Disc.*, 5(1):27–36, 2006. 125

J. Besag. Efficiency of pseudolikelihood estimation for simple gaussian fields. *Biometrika*, pages 616–618, 1977. 14, 15

R Blume-Kohout and P Hayden. Accurate quantum state estimation via "keeping the experimentalist honest". 2006. 123

Robin Blume-Kohout. Optimal, reliable estimation of quantum states. *New J. Phys.*, 12(4):043034, 2010. doi: 10.1088/1367-2630/12/4/043034. 122, 123, 125

G.W. Brier. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3, 1950. 16

Cristina Butucea, Mădălin Guţă, and Luis Artiles. Minimax and adaptive estimation of the wigner function in quantum homodyne tomography with noisy data. *The Annals*

of Statistics, 35(2):pp. 465–494, 2007. ISSN 00905364. URL http://www.jstor.org/stable/25463565. 130

D.R. Cavagnaro, J.I. Myung, M.A. Pitt, and J.V. Kujala. Adaptive design optimization: A mutual information-based approach to model discrimination in cognitive science. *Neur. Comp.*, 22(4):887–905, 2010. 125

Moses S Charikar. Similarity estimation techniques from rounding algorithms. In *Proceedings of the thiry-fourth annual ACM symposium on Theory of computing*, pages 380–388. ACM, 2002. 28

Y. Chen, M. Welling, and A. Smola. Super-samples from kernel herding. UAI, 2010. 64, 67, 76, 83

Y. Chen, M. Welling, and A. Smola. Super-samples from kernel herding. *arXiv preprint arXiv:1203.3472*, 2012. 21, 54

W. Chu and Z. Ghahramani. Preference learning with gaussian processes. In *Proceedings of the 22nd international conference on Machine learning*, pages 137–144. ACM, 2005. 110, 111, 112

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. 14, 15

Mary K. Cowles and Bradley P. Carlin. Markov chain monte carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association*, 91(434):883–904, 1996. ISSN 01621459. doi: 10.2307/2291683. URL http://dx.doi.org/10.2307/2291683. 61, 64

L. Csató and M. Opper. Sparse on-line gaussian processes. *Neural Computation*, 14(3): 641–668, 2002. 56, 102

Lehel Csató, Ernest Fokoué, Manfred Opper, Bernhard Schottky, and Ole Winther. Efficient approaches to gaussian process classification. 2000. 102

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. 15

S. Dasgupta and D. Hsu. Hierarchical sampling for active learning. In *Proceedings of the 25th international conference on Machine learning*, pages 208–215. ACM, 2008. 100

A.P. Dawid. The geometry of proper scoring rules. *Annals of the Institute of Statistical Mathematics*, 59(1):77–93, 2007. 8, 17, 18, 22, 126

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. 13, 15, 17

Phillip Dawid. Proper measures of discrepancy uncertainty and dependence with applications to predictive experimental design. Technical Report 139, University College London, 1994. 87, 89

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. 19

P. A. M. Dirac. A new notation for quantum mechanics. *Mathematical Proceedings of the Cambridge Philosophical Society*, 35:416–418, 6 1939. ISSN 1469-8064. doi: 10.1017/S0305004100021162. URL http://journals.cambridge.org/article_S0305004100021162. 117

Arnaud Doucet, Nando de Freitas, and Neil Gordon, editors. *Sequential Monte Carlo in Paractice*. Springer-Verlag, 2001. ISBN 0-387-95146-6. 123, 124, 125

R.M. Dudley. Metric entropy of some classes of sets with differentiable boundaries. *Journal of Approximation Theory*, 10(3):227–236, 1974. 19

M. L. Eaton. *A method for evaluating improper prior distributions.*, pages 329–352. Academic Press, New York, 1982. 22

Morris L Eaton, Alessandra Giovagnoli, and Paola Sebastiani. A predictive approach to the bayesian design problem with application to normal regression models. *Biometrika*, 83(1):111–125, 1996. 18, 22

Emre Ertin, John Fisher, and Lee Potter. Maximum mutual information principle for dynamic sensor query problems. In Feng Zhao and Leonidas Guibas, editors, *Information Processing in Sensor Networks*, volume 2634 of *Lecture Notes in Computer Science*, pages 558–558. Springer Berlin / Heidelberg, 2003. URL http://dx.doi.org/10.1007/3-540-36978-3-27. 10.1007/3-540-36978-3-27. 92

Ugo Fano. Description of states in quantum mechanics by density matrix and operator techniques. *Reviews of Modern Physics*, 29(1):74–93, 1957. 121

Gustav Theodor Fechner. *Elemente der Psychophysik (Elements of Psychophysics)*. Breitkopf and H artel, 1860. 110

C.A.T. Ferro. Comparing probabilistic forecasting systems with the brier score. *Weather and Forecasting*, 22(5):1076–1088, 2007. 16

Dietmar G. Fischer, Stefan H. Kienle, and Matthias Freyberger. Quantum-state estimation by self-learning measurements. *Phys. Rev. A*, 61:032306, Feb 2000. doi: 10.1103/PhysRevA.61.032306. URL http://link.aps.org/doi/10.1103/PhysRevA.61.032306. 125, 126

Marcus Frean and Phillip Boyle. Using gaussian processes to optimize expensive functions. In *AI 2008: Advances in Artificial Intelligence*, pages 258–267. Springer, 2008. 93

Y. Freund, H.S. Seung, E. Shamir, and N. Tishby. Selective sampling using the query by committee algorithm. *Machine Learning*, 28(2):133–168, 1997. 99

D.R. Fuhrmann. *Active Testing Surveillance Systems, or, Playing Twenty Questions with a Radar*. Defense Technical Information Center, 2003. 92

K. Fukumizu, L. Song, and A. Gretton. Kernel bayes' rule. *arXiv preprint arXiv:1009.5736*, 2010. 19

J. Fürnkranz and E. Hüllermeier. *Preference learning*. Springer-Verlag New York Inc, 2010. 110, 112

Alex Gammerman, Volodya Vovk, and Vladimir Vapnik. Learning by transduction. In *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence*, pages 148–155. Morgan Kaufmann Publishers Inc., 1998. 91

T. Gneiting and A.E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. 18, 19, 22

Michel X. Goemans and David P. Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *J. ACM*, 42(6):1115–1145, November 1995. ISSN 0004-5411. doi: 10.1145/227683.227684. URL http://doi.acm.org/10.1145/227683.227684. 28

Irving John Good. discussion of proper scores for probability forecasters by hendrickson and buehler. discussion paper, 1971. 17

Thore Graepel, Ralf Herbrich, and Klaus Obermayer. Bayesian transduction. *Advances in Neural Information Processing Systems*, 12:456–462, 1999. 91

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. 12, 19, 21, 132

Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *Algorithmic learning theory*, pages 63–77. Springer, 2005a. 24

Arthur Gretton, Alex Smola, Olivier Bousquet, Ralf Herbrich, Andreas Belitski, Mark Augath, Yusuke Murayama, Jon Pauls, Bernhard Schölkopf, and Nikos Logothetis. Kernel constrained covariance for dependence measurement. 2005b. 24

David Gross *et al*. Quantum state tomography via compressed sensing. *Phys. Rev. Lett.*, 105(15):150401, Oct 2010. doi: 10.1103/PhysRevLett.105.150401. URL http://link.aps.org/doi/10.1103/PhysRevLett.105.150401. 123, 130

Th. Hannemann, D. Reiss, Ch. Balzer, W. Neuhauser, P. E. Toschek, and Ch. Wunderlich. Self-learning estimation of quantum states. *Phys. Rev. A*, 65:050303, May 2002. doi: 10.1103/PhysRevA.65.050303. URL http://link.aps.org/doi/10.1103/PhysRevA.65.050303. 125, 127

P. Hennig and M. Kiefel. Quasi-Newton methods – a new direction. In *Int. Conf. on Machine Learning (ICML)*, volume 29, 2012. 88

Philipp Hennig and Christian J Schuler. Entropy search for information-efficient global optimization. *The Journal of Machine Learning Research*, 98888:1809–1837, 2012. 88, 93

Zi-Quan Hong and Jing-Yu Yang. Optimal discriminant plane for a small number of samples and design method of classifier on the plane. *pattern recognition*, 24(4):317–324, 1991. 106

Neil Houlsby, Ferenc Huszár, Mate Lengyel, and Zoubin Ghahramani. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*, 2011. 88, 91, 97, 99, 108, 112

Neil Houlsby, José Miguel Hernández-Lobato, Ferenc Huszar, and Zoubin Ghahramani. Collaborative gaussian processes for preference learning. In Peter L. Bartlett, Fernando C. N. Pereira, Christopher J. C. Burges, Léon Bottou, and Kilian Q. Weinberger, editors, *NIPS*, pages 2105–2113, 2012. 97, 99, 108, 112, 113

S.J. Huang, R. Jin, and Z.H. Zhou. Active learning by querying informative and representative examples. *Advances in neural information processing systems*, 23:892–900, 2010. 100

Ferenc Huszár and David Duvenaud. Optimally-weighted herding is Bayesian quadrature. In *Uncertainty in Artificial Intelligence*, 2012. 53, 132

Ferenc Huszár and Neil MT Houlsby. Adaptive bayesian quantum tomography. *Physical Review A*, 85(5):052120, 2012. 91, 99

A. Hyvärinen. Consistency of pseudolikelihood estimation of fully visible boltzmann machines. *Neural Computation*, 18(10):2283–2292, 2006. 14, 15

Edwin T Jaynes. Information theory and statistical mechanics. *Physical review*, 106(4): 620, 1957. 87

Jianqiu Ji, Jianmin Li, Shuicheng Yan, Bo Zhang, and Qi Tian. Super-bit locality-sensitive hashing. In *Advances in Neural Information Processing Systems 25*, pages 108–116, 2012. 28

Shihao Ji, Ya Xue, and Lawrence Carin. Bayesian compressive sensing. *Signal Processing, IEEE Transactions on*, 56(6):2346–2356, 2008. 91

Donald R. Jones. A taxonomy of global optimization methods based on response surfaces. *J. of Global Optimization*, 21(4):345–383, December 2001. ISSN 0925-5001. doi: 10.1023/A:1012771025575. URL http://dx.doi.org/10.1023/A:1012771025575. 93

Donald R. Jones, Matthias Schonlau, and William J. Welch. Efficient global optimization of expensive black-box functions. *J. of Global Optimization*, 13(4):455–492, December 1998. ISSN 0925-5001. doi: 10.1023/A:1008306431147. URL http://dx.doi.org/10.1023/A:1008306431147. 93

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. 18, 132

A. Kapoor, E. Horvitz, and S. Basu. Selective supervision: Guiding supervised learning with decision-theoretic active learning. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2007. 88, 92, 106, 107, 108, 113

Robert Kosut, Ian A. Walmsley, and Herschel Rabitz. Optimal experiment design for quantum state and process tomography and 2004. 124, 125

A. Krause and C. Guestrin. Nonmyopic active learning of gaussian processes: an exploration-exploitation approach. In *Proceedings of the 24th international conference on Machine learning*, pages 449–456. ACM, 2007. 89

A. Krause, C. Guestrin, A. Gupta, and J. Kleinberg. Near-optimal sensor placements: Maximizing information while minimizing communication cost. In *Proceedings of the 5th international conference on Information processing in sensor networks*, pages 2–10. ACM, 2006a. 88, 91

A. Krause, C. Guestrin, A. Gupta, and J. Kleinberg. Near-optimal sensor placements: Maximizing information while minimizing communication cost. In *Proceedings of the Fifth International Conference on Information Processing in Sensor Networks (IPSN '06)*, pages 2–10, Nashville, Tennessee, USA, 2006b. 70

Andreas Krause and Volkan Cevher. Submodular dictionary selection for sparse representation. In Johannes Fürnkranz and Thorsten Joachims, editors, *ICML*, pages 567–574. Omnipress, 2010. ISBN 978-1-60558-907-7. 65, 74, 75, 81

B. Krishnapuram, D. Williams, Y. Xue, A. Hartemink, L. Carin, and M. Figueiredo. On semi-supervised classification. *Advances in neural information processing systems*, 17: 721–728, 2004. 98

S Lacoste-Julien, F Huszar, and Z Ghahramani. Approximate inference for the loss-calibrated bayesian. In *Proceedings of 14th International Conference on Artificial Intelligence and Statistics*, volume 15, pages 416–424, 2011. 53, 54

Thaddeus D Ladd, Fedor Jelezko, Raymond Laflamme, Yasunobu Nakamura, Christopher Monroe, and Jeremy L O'Brien. Quantum computers. *Nature*, 464(7285):45–53, 2010. 117

Steffen L Lauritzen. *Graphical models*, volume 17. Oxford University Press, 1996. 131

Neil D. Lawrence and John C. Platt. Learning to learn with the informative vector machine. In *ICML 2004*, pages 65+, 2004. 91, 98

D.V. Lindley. On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, 27(4):986–1005, 1956. 87, 98

Daniel James Lizotte. *Practical bayesian optimization*. PhD thesis, Edmonton, Alta., Canada, 2008. AAINR46365. 93

David J. C. MacKay. *Information Theory, Inference & Learning Algorithms*. Cambridge University Press, New York, NY, USA, 2002. ISBN 0521642981. 13

D.J.C. MacKay. Information-based objective functions for active data selection. *Neural computation*, 4(4):590–604, 1992. 88, 90, 91, 92, 98

T. Minka and J. Lafferty. Expectation-propagation for the generative aspect model. In *Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence*, pages 352–359, 2002. 102

T. P. Minka. Deriving quadrature rules from Gaussian processes. Technical report, Statistics Department, Carnegie Mellon University, 2000. 70

Tom Minka. *A family of algorithms for approximate Bayesian inference*. PhD thesis, MIT, 2001a. 56

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., 2001b. 56, 112

J Mockus. *Systems Modeling and Optimization*, volume 38, chapter The Bayesian approach to global optimizationMockus, J. The Bayesian approach to global optimization, pages 473 – 481. Springer, 1982. 93

Iain Murray. *Advances in Markov chain Monte Carlo methods*. PhD thesis, Gatsby computational neuroscience unit, University College London, 2007. 61

Hannes Nickisch and Carl Edward Rasmussen. Approximations for binary Gaussian process classification. *The Journal of Machine Learning Research*, 9:2035–2078, 2008. 57, 102

J. Nunn *et al*. Optimal experiment design for quantum state tomography: Fair, precise, and minimal tomography. *Phys. Rev. A*, 81(4):042109, Apr 2010. doi: 10.1103/PhysRevA.81.042109. URL http://link.aps.org/doi/10.1103/PhysRevA.81.042109. 124, 125

A. O'Hagan. Bayes-Hermite quadrature. *Journal of Statistical Planning and Inference*, 29:245–260, 1991. 64, 88

J.K. Ord, S.F. Arnold, A. O'Hagan, and J. Forster. *Kendall's advanced theory of statistics*. A. Arnold, 1999. 20

Senatore R. Montemurro M.A. Panzeri, S. and R.S. Petersen. Correcting for the sampling bias problem in spike train information measures. *Journal of neurophysiology*, 98(3): 1064, 2007. 98

Matthew Parry, A. Philip Dawid, and Steffen Lauritzen. Proper local scoring rules. *Annals of Statistics*, 40(1):561–592, 2012. 13

M K Patra. Quantum state determination: estimates for information gain and some exact calculations. *J. Phys. A*, 40(35):10887, 2007. URL http://stacks.iop.org/1751-8121/40/i=35/a=011. 125, 126, 130

Michael L Platt and Paul W Glimcher. Neural correlates of decision variables in parietal cortex. *Nature*, 400(6741):233–238, 1999. 110

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. 14

G. J. Pryde, J. L. O'Brien, A. G. White, S. D. Bartlett, and T. C. Ralph. Measuring a photonic qubit without destroying it. *Phys. Rev. Lett.*, 92:190402, May 2004. doi: 10.1103/PhysRevLett.92.190402. URL http://link.aps.org/doi/10.1103/PhysRevLett.92.190402. 119

J. QuiÒonero-Candela and C. E. Rasmussen. A unifying view of sparse approximate Gaussian process regression. *Journal of Machine Learning Research*, 6:1935–1959, 2005. 102

C. E. Rasmussen and Z. Ghahramani. Bayesian monte carlo. In S. Becker and K. Obermayer, editors, *Advances in Neural Information Processing Systems*, volume 15. MIT Press, Cambridge, MA, 2003. 64, 67, 70, 88

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, Cambridge, MA, USA, 2006a. 101

C.E. Rasmussen and C.K.I. Williams. Gaussian processes for machine learning. 2005. 102

C.E. Rasmussen and CKI Williams. Gaussian Processes for Machine Learning. *The MIT Press, Cambridge, MA, USA*, 2006b. 81

Philippe Raynal, Xin Lü, and Berthold-Georg Englert. Mutually unbiased bases in six dimensions: The four most distant bases. *Phys. Rev. A*, 83(6):062303, Jun 2011. doi: 10.1103/PhysRevA.83.062303. URL http://link.aps.org/doi/10.1103/PhysRevA.83.062303. 124, 130

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. 16

Christian P. Robert. *The Bayesian Choice*. Springer, New York, 2001. 57

Sam Roweis and Zoubin Ghahramani. A unifying review of linear gaussian models. *Neural computation*, 11(2):305–345, 1999. 131

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. 16

Mark J. Schervish. *Theory of Statistics*. Springer, New York, 1995. 58

P. Sebastiani and H.P. Wynn. Maximum entropy sampling and optimal Bayesian experimental design. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(1):145–157, 2000. 100

Burr Settles. Active learning literature survey. *University of Wisconsin, Madison*, 2010. 91

HS Seung, M. Opper, and H. Sompolinsky. Query by committee. In *COLT*, pages 287–294. ACM, 1992. 98, 100, 101

Claude E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 1948. 13

Peter W Shor. Algorithms for quantum computation: discrete logarithms and factoring. In *Foundations of Computer Science, 1994 Proceedings., 35th Annual Symposium on*, pages 124–134. IEEE, 1994. 115

Peter W Shor. Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer. *SIAM journal on computing*, 26(5):1484–1509, 1997. 115

Greg A. Smith, Andrew Silberfarb, Ivan H. Deutsch, and Poul S. Jessen. Efficient quantum-state estimation by continuous weak measurement and dynamical control. *Phys. Rev. Lett.*, 97:180403, Oct 2006. doi: 10.1103/PhysRevLett.97.180403. URL http://link.aps.org/doi/10.1103/PhysRevLett.97.180403. 119

A. Smola, A. Gretton, L. Song, and B. Schölkopf. A hilbert space embedding for distributions. In *Algorithmic Learning Theory*, pages 13–31. Springer, 2007. 19, 20

Le Song, Xinhua Zhang, Alex Smola, Arthur Gretton, and Bernhard Schölkopf. Tailoring density estimation via reproducing kernel moment matching. In *Proceedings of the 25th international conference on Machine learning*, ICML '08, pages 992–999, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-205-4. doi: 10.1145/1390156.1390281. URL http://doi.acm.org/10.1145/1390156.1390281. 19, 21, 23, 67, 70

D.J. Spiegelhalter. Probabilistic prediction in patient management and clinical trials. *Statistics in Medicine*, 5(5):421–433, 2006. 16

Niranjan Srinivas, Andreas Krause, Sham M Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. *arXiv preprint arXiv:0912.3995*, 2009. 93

B.K. Sriperumbudur, A. Gretton, K. Fukumizu, G. Lanckriet, and B. Schölkopf. Injective hilbert space embeddings of probability measures. 2008. 12, 19, 20, 65, 66, 81

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. 19

Henry Teicher. Identifiability of mixtures. *The annals of Mathematical statistics*, 32(1): 244–248, 1961. 120

Michael E Tipping and Christopher M Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61 (3):611–622, 1999. 131

S. Tong and D. Koller. Support vector machine active learning with applications to text classification. *Journal of Machine Learning Research*, 2:45–66, 2001. 100, 101, 106

Richard Turner and Maneesh Sahani. A maximum-likelihood interpretation for slow feature analysis. *Neural computation*, 19(4):1022–1038, 2007. 131

C. Vondrick, UC Irvine, and D. Ramanan. Video annotation and tracking with active learning. *NIPS*, 2011. 125

M. Welling. Herding dynamical weights to learn. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1121–1128. ACM, 2009. 21, 65, 132

Christopher K. I. Williams and David Barber. Bayesian classification with gaussian processes. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 20(12): 1342–1351, 1998. 102

John Winn and Christopher M Bishop. Variational message passing. *Journal of Machine Learning Research*, 6(1):661, 2006. 55

William Wootters and Brian Fields. Optimal state-determination by mutually unbiased measurements. *Ann. Phys.*, 191(2):363 – 381, 1989. ISSN 0003-4916. doi: DOI:10.1016/0003-4916(89)90322-9. URL http://www.sciencedirect.com/science/article/pii/0003491689903229. 124, 125, 126

Fei Yan, Ming Yang, and Zhuo-Liang Cao. Optimal reconstruction of the states in qutrit systems. *Phys. Rev. A*, 82:044102, Oct 2010. doi: 10.1103/PhysRevA.82.044102. URL http://link.aps.org/doi/10.1103/PhysRevA.82.044102. 124

X. Zhu, Z. Ghahramani, and J. Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In *Proceedings of the 20th international conference on Machine learning*, volume 20, page 912, 2003a. 106, 107, 108

Xiaojin Zhu, John Lafferty, and Zoubin Ghahramani. Combining active learning and semi-supervised learning using gaussian fields and harmonic functions. In *ICML 2003 workshop on the continuum from labeled to unlabeled data in machine learning and data mining*, pages 58–65, 2003b. 88, 92