

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

Ferenc Huszár

July 15, 2012

Part I

Introduction

Part II

Information geometry of probability distributions

Chapter 1

An introduction to scoring rules

In this section I describe scoring rules that can be used to assess the performance of probabilistic forecasting models. The scoring rule framework allows us to define useful generalisations of well-known information quantities, such as entropy, mutual information and divergence. Scoring rules allow for defining geometry of probabilistic models.

Let's say we want to have build a probabilistic forecaster that predicts the value of a random quantity X . We can describe any such probabilistic forecaster as a probability distribution $P(x)$ over the space of possible outcomes \mathcal{X} . After observing the outcome $X = x$ we want to assess how good our predictions were: scoring rule is a general term for a function that quantifies this: if the outcome was $X = x$, and our prediction was P we incur a score $S(x, P)$. A score by convention is interpreted as a loss, so lower values are better. A widespread example of scoring rules is the logarithmic score, or simply the log score: $S_{\log}(x, P) = -\log P(x)$. When comparing multiple probabilistic models, the score is often referred to as the negative log likelihood. I will give further examples of scoring rules in section ???. Mathematically scoring rule is any measurable function that maps an outcome-probability distribution pair onto real numbers: $s : \mathcal{X} \times \mathcal{M}_{\mathcal{X}}^1 \mapsto \mathcal{R} \cup \{\infty\}$.

1.1 Information quantities

A scoring rule allows us to define the following, useful quantities:

The first quantity is the generalised entropy of a distribution

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

that measures how hard is it to predict the outcome on average, when true distribution P itself is used as the forecasting model.

A further quantity of interest is the divergence between two distributions P and Q , defined as

$$d[P] = \mathbb{E}_{x \sim P} S(x, Q) - \mathbb{E}_{x \sim P} S(x, P), \quad (1.2)$$

which measures how much worse we are at forecasting a quantity X sampled from a distribution P when instead of using the true distribution P , we use an alternative model, Q . Ideally, you would expect that using the true model should always be better or at least as good as using any alternative model Q , but this is not automatically true for all scoring rules. A scoring rule that has this property is called a proper scoring rule.

Definition with respect to a class of models \mathcal{Q} , if has this property, in other words if $d(P\|Q) \geq 0$ for all $P, Q \in \mathcal{Q}$.

Furthermore, if the inequality is always strict, the scoring rule is called strictly proper

Definition with respect to a class of models \mathcal{Q} , if has this property, in other words if $d(P\|Q) > 0$ for all $P, Q \in \mathcal{Q}$.

Example The log score is strictly proper for the class of continuous probability densities.

1.2 Examples of proper scoring rules

1.2.1 The pseudolikelihood

Example: Markov random fields

Example: Population genetics, infinite sites model

1.2.2 Local proper scoring rules

1.2.3 Brier score

1.2.4 kernel scoring rule

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

$$\ell(\theta, a) = \sup_{f \in \mathcal{F}} (f(\theta) - a(f))^2 \quad (1.3)$$

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

$$a_p^* = \int f(\theta) p(\theta) d\theta \quad (1.4)$$

Thus, we can define the following scoring rule S :

$$S(\theta, p) = \sup_{f \in \mathcal{F}} \left(f(\theta) - \int f(\theta) p(\theta) d\theta \right)^2 \quad (1.5)$$

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

1.2.5 Scoring rules based on general decision problems

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

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

$$a_P^* = \operatorname{argmax}_{a \in \mathcal{A}} \mathbb{E}_{x \sim P} \ell(x, a) \quad (1.6)$$

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

$$S(x, P) = \ell(x, a_P^*) \quad (1.7)$$

Chapter 2

Information geometry

[?] showed that

Chapter 3

Visualising information geometries

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

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

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

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

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

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

3.0.6 Visualising the Shannon-information geometry

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

butions.

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

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

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

$$d_{KL}[\mathcal{N}_{\mu_1, \sigma_1} \parallel \mathcal{N}_{\mu_2, \sigma_2}] = \frac{(\mu_1 - \mu_2)^2}{2\sigma_2^2} + \frac{1}{2} \left(\frac{\sigma_1^2}{\sigma_2^2} - 1 - \log \frac{\sigma_1^2}{\sigma_2^2} \right) \quad (3.1)$$

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

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

$$f(x) = \beta^\alpha \frac{1}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\beta x) \quad (3.2)$$

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

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

$$d_{KL}[\Gamma_{\alpha_1, \beta_1} \parallel \Gamma_{\alpha_2, \beta_2}] = (\alpha_1 - \alpha_2) \psi(\alpha_1) - \log \frac{\Gamma(\alpha_1)}{\Gamma(\alpha_2)} + \alpha_1 \log \frac{\beta_1}{\beta_2} + \alpha_1 \frac{\beta_2 - \beta_1}{\beta_1} \quad (3.3)$$

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

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

$$\mu = \frac{\alpha}{\beta} \quad (3.4)$$

$$\sigma^2 = \frac{\alpha}{\beta^2} \quad (3.5)$$

Solving for α and β in these equations we get

$$\alpha = \frac{\mu^2}{\sigma^2} \quad (3.6)$$

$$\beta = \frac{\mu}{\sigma^2} \quad (3.7)$$

Plugging these into Eqn. (3.3) we can now map Gamma distributions with particular mean and variance. Figure 1 compares Normal and Gamma distributions with mean $\mu \in [0.5, 1.5]$ and standard deviation $\sigma \in [0.1, 1]$. We can observe that as the variance increases, the manifold of Gamma distributions shows a fan-like structure very similarly that of Normal distributions. However, for larger variance, the distributions look less Gaussian, and the manifold becomes more asymmetric. The effect of the central limit theorem would perhaps be even more prominent for smaller values of σ , but for those cases that case Eqn. (3.3) becomes numerically imprecise, as it relies on look-up-table implementation of the Gamma (Γ) and bigamma (ψ) functions.

3.0.7 Visualising geometries induced by divergences other than KL

Chapter 4

Score matching

Not all manifolds can be faithfully embedded in a two-dimensional space this way. For example the surface of a two-dimensional sphere is a two dimensional Riemannian manifold, which is can perfectly isometrically represented in a three-dimensional space, but is distorted when embedded into two dimensions.

4.1 Infinite dimensional sampling spaces

Chapter 5

The role of information geometry in approximate Bayesian analysis

5.1 Introduction

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

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

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

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

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

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

5.2 Loss-calibrated approximate inference

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

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

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

$$\mathcal{R}_{p_{\mathcal{D}}}(a) \doteq \int_{\Theta} \ell(\theta, a) p(\theta|\mathcal{D}) d\theta \quad (5.1)$$

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

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

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

Overview of KL minimisation in one way Variational methods

Overview of EP and minimizing KL in other way

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

Toy example

Framework

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

Gaussian process regression.

5.3 Loss-calibrated quasi-Monte-Carlo

Monte Carlo, powerful but

Chapter 6

Information-greedy Bayesian active learning

6.0.1 Model fitting

Maximum likelihood, and related methods

Herding as scoring rule minimisation

Introduce herding as a form of maximising average proper scoring rules.

6.0.2 Approximate Bayesian inference

Overview of existing approaches based on KL divergence and alpha divergence

variational, EP, power-EP, a generalisation of power-EP to general Bregman divergences

Loss-calibrated approximate inference

loss-calibrated Bayesian inference, examples of increasing sophistication from nuclear power plant toy example to something vaguely useful, such as loss-calibrated filtering

6.0.3 Bayesian experiment design

General case

A criterion for active learning based on proper scoring rules, decision-theoretic active learning and entropy minimisation as special cases.

Shannon's entropy case

derivation of the BALD criterion and exploration of its properties.

Part III

Bayesian analysis in experimental sciences

Chapter 7

Cognitive tomography: Bayesian analysis of choice probabilities

Optimal experiment design v.s. MCMC with People

An important comparison we always wanted to make.

Chapter 8

Quantum Tomography

8.1 Introduction

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

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

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

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

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

provide Bayesian analysis of the problem and then propose a

8.2 Overview of quantum statistics

quantum states

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

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

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

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

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

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

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

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

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

$$e = mc^2 \tag{8.1}$$

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

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

Born rule

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

8.2.1 Inference in quantum tomography

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

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

8.2.2 optimal experiment design and active tomography

8.3 Adaptive Bayesian Quantum Tomography

8.4 Results