# Information Geometric Measurements of Generalisation

Huaiyu Zhu and Richard Rohwer

NCRG/4350

Technical Report NCRG/4350

August 31, 1995

*Neural Computing Research Group*
Dept. of Computer Science and Applied Mathematics
Aston University, Aston Triangle
Birmingham B4 7ET, UK

**Tel: +44 (0)121 359-3611**
**Fax: +44 (0)121 333-6215**

# Information Geometric Measurements of Generalisation

Huaiyu Zhu and Richard Rohwer
Dept of Computer Science and Applied Mathematics
Aston University, Aston Triangle, Birmingham B4 7ET
Email: zhuh@aston.ac.uk, rohwerrj@aston.ac.uk

August 31, 1995

## Abstract

Neural networks can be regarded as statistical models, and can be analysed in a Bayesian framework. Generalisation is measured by the performance on independent test data drawn from the same distribution as the training data. Such performance can be quantified by the posterior average of the information divergence between the true and the model distributions. Averaging over the Bayesian posterior guarantees internal coherence; Using information divergence guarantees invariance with respect to representation.

The theory generalises the least mean squares theory for linear Gaussian models to general problems of statistical estimation. The main results are: (1) the ideal optimal estimate is always given by average over the posterior; (2) the optimal estimate within a computational model is given by the projection of the ideal estimate to the model. This incidentally shows some currently popular methods dealing with hyperpriors are in general unnecessary and misleading.

The extension of information divergence to positive normalisable measures reveals a remarkable relation between the $\delta$ dual affine geometry of statistical manifolds and the geometry of the dual pair of Banach spaces $L_{1/\delta}$ and $L_{1/(1-\delta)}$. It therefore offers conceptual simplification to information geometry.

The general conclusion on the issue of evaluating neural network learning rules and other statistical inference methods is that such evaluations are only meaningful under three assumptions: The prior $P(p)$, describing the environment of all the problems; the divergence $D_\delta$, specifying the requirement of the task; and the model $\mathcal{Q}$, specifying available computing resources.

# 1 Introduction

A neural network (either deterministic or stochastic) can be regarded as a parameterised model [Whi89]

$$P(y|x, w), \tag{1.1}$$

where $x \in X$ is the input, $y \in Y$ is the output and $w \in W$ is the weight matrix (Figure 1). In an environment with an input distribution $P(x)$, it is also equivalent to $P(x, y|w)$ or simply $P(z|w)$, where $z := [x, y] \in Z := [X, Y]$ denotes the combined input and output as data. This framework also includes unsupervised learning.

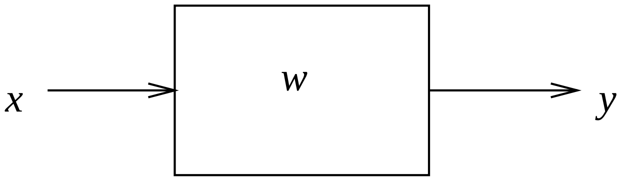

Figure 1: A neural network as a black box

Learning, or training, is the task of inferring $w$ from $z$. Therefore it is a typical statistical inference problem in which a neural network model acts as a likelihood function [Fis25, Zac71, CH74, KS79].

It is obvious that there can always be infinitely many $w$ which could have generated $z$ and it is logically impossible to select $w$ unless other information not contained in $z$ is available. In a Bayesian framework such auxiliary information is represented in the form of a prior $P(w)$, the distribution of $w$ before the data $z$ is seen. By the Bayes Theorem

$$P(w|z) = P(z|w)P(w)/P(z), \tag{1.2}$$

a posterior distribution of $w$ can be obtained, which contains the combined information of both the prior and the likelihood function. Nothing is lost by working in a Bayesian framework, since it is well known that, under mild regularity conditions, whatever the optimality condition is, for any given learning rule, there is always a Bayes learning rule which is not worse than the original learning rule under all circumstances [Fer67]. Sample-oriented statistics can often be interpreted as Bayesian statistics with improper priors [Aka80], although extra caution is needed dealing with improper priors [DSZ73].

The reason one can obtain a posterior about $w$ is that it is uniquely associated with a distribution by $w \leftrightarrow p := P(\cdot|w)$. Since $w$ does not have separate meaning apart from being a coordinate of $p$, it is natural to identify $w$ with $p$. With this view, both the prior and posterior are distributions of distributions, as noted in [Fis36, p. 247], and the Bayes Theorem can be more generally written as

$$P(p|z) = P(z|p)P(p)/P(z). \tag{1.3}$$

Let $\mathcal{P}$ be the space of probability distributions over $Z$. Those distributions which can be represented by the neural network with a particular weight form a subspace $\mathcal{Q} \subseteq \mathcal{P}$. The relation between $Z$, $\mathcal{P}$, $\mathcal{Q}$ and $W$ are illustrated in Figure 2.

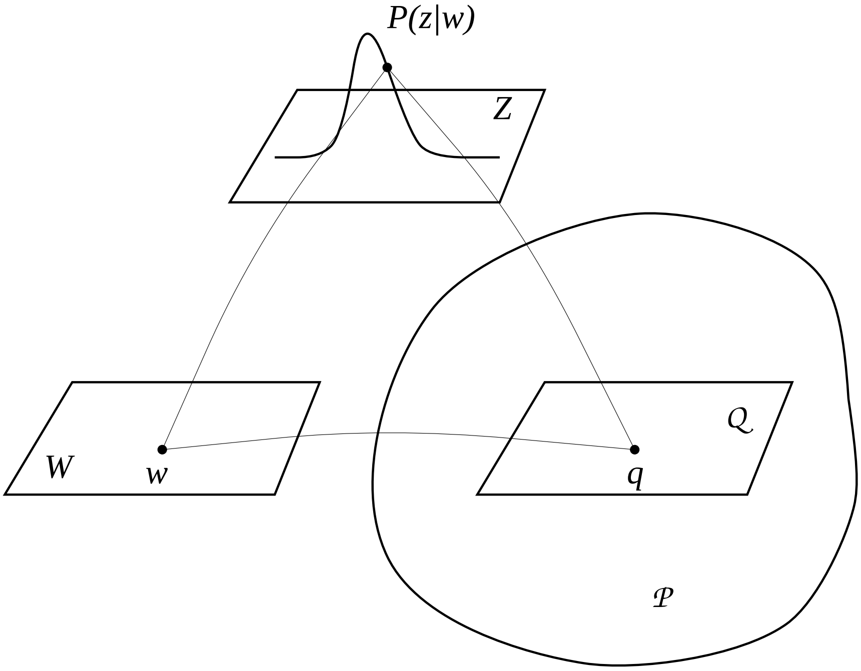

Figure 2: The relation between $Z$, $\mathcal{P}$, $\mathcal{Q}$ and $W$.

A neural network training algorithm $\tau : Z \to W$ must fix one distribution, say $q$, which in a sense approximates all the distributions $p$ which are possible *a posteriori*. [1] The relation between the learning rule and the Bayes theorem is shown in Figure 3.

How does one choose a particular $q$ from the posterior $P(p|z)$? Naturally, one wants the chosen $q$ to be closest to the true $p$ considered over the posterior. However, the metric in the space $W$ does not have an intrinsic meaning, as shown in Figure 4. What we need is a measure of "divergence"

$$D(p, q) \tag{1.4}$$

between two distributions, $p$ and $q$, which should be defined on the space of distributions $\mathcal{P}$, and should be invariant under reparameterisation. This is what exactly has been offered by the theory of information geometry (See [Ama82, Ama85, BNCR86, ABNK$^+$87, Kas89] for introductions, backgrounds and references), in which a parameterised family of distributions, $\mathcal{P}$, is regarded as a differentiable manifold, where the parameters $w$ act as

[1] Even those Bayesian methods which do not make a point estimate use an implicit point estimate when the network is applied. We shall come across this later.

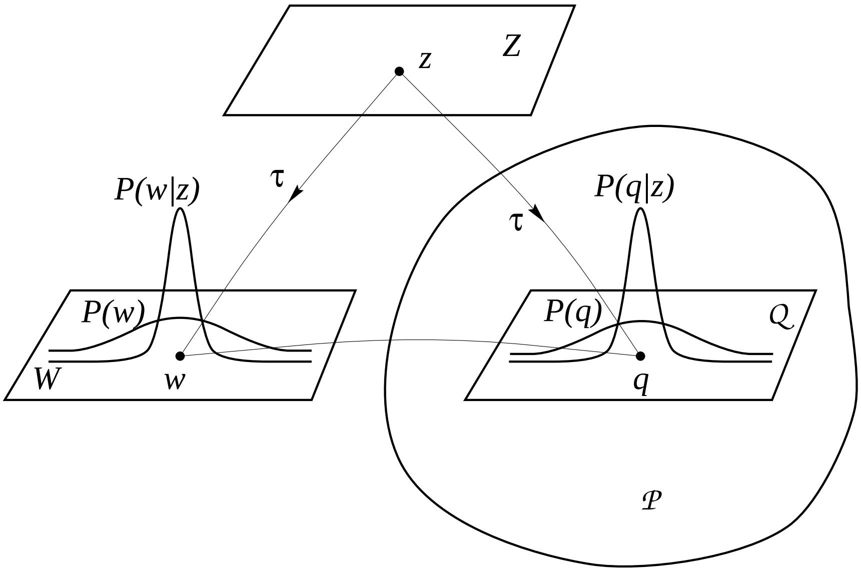

Figure 3: The role of Bayes Theorem on learning rules

coordinates. Each point in $\mathcal{P}$ represents a distribution. The word "geometry" captures two aspects of the theory: it studies quantities which can be measured with real numbers, in contrast to topology; and the measurements are independent of the coordinates, in contrast to algebra.

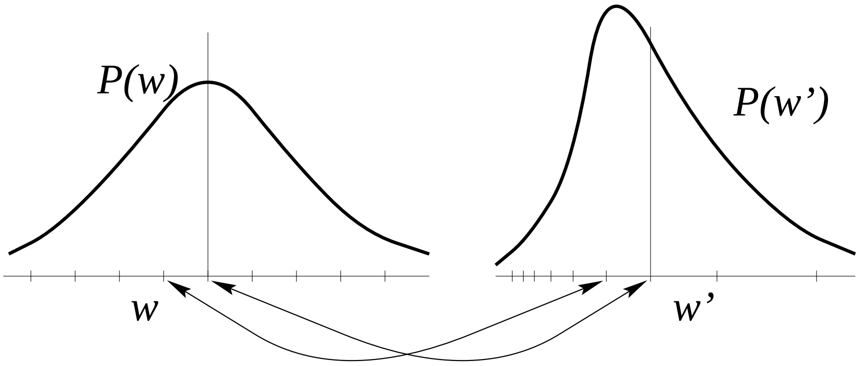

Figure 4: Weight space is not invariant
Mean and mode will change with a weight space transform.

To an uninitiated mind, it is quite surprising to know that with information geometry one can regard distributions as points, and talk about the length, angle, curvature, etc., of geometrical objects composed of distributions. One of the major contributions from information geometry which is particularly relevant to our current aim is that there is a one-parameter family of "divergences" defined on the space of distributions, which possess properties enabling them to be regarded as generalisations of the squared distances which have been so important in linear theory.

It will be shown that in order to develop a theory as comprehensive as that of linear regression, it is necessary to extend information geometry from the manifold $\mathcal{P}$ of all probability measures to the manifold $\widetilde{\mathcal{P}}$ of all the positive definite measures. This generalisation follows suggestions in [Ama85].

The main thesis of this paper is that generalisation should be measured as the posterior expected divergence between the true distribution and the model distribution. Using the Bayesian posterior guarantees that the optimal learning rule is a Bayes rule, which gives optimal estimates for almost all the data. Using the information divergence guarantees that the criterion of optimality is invariant of representation. That is, it remains the same under any invertible transforms on the input, output and weight spaces.

The theory presented here generalises least mean square estimation for linear Gaussian models to any statistical estimation problems, in that there is always a unique optimal estimate which is obtained by appropriate average of the posterior, and that the optimal estimate within a model can be obtained by an appropriate projection of the ideal optimal estimate onto the model. The added complexity is comparable to, and related to, that of the generalisation from Hilbert spaces to Banach spaces.

A condensed version of this report was presented at the MANNA conference, 1995 [ZR95c].

## 2 Information Geometry — Basic Concepts

This section is basically a review of the part of information geometry immediately relevant to our current study. Most of our notation on information geometry follows [Ama85, Ama87], although for technical convenience, we use $\delta = (1-\alpha)/2$ instead of $\alpha$, following [Hou82, Kas84]. The latter references also give interpretations of $\delta$ in the sense of classical statistics. The proof of results stated in this section can all be found in [Ama85], the first part of which is perhaps still the best introduction to information geometry to-date.

Let $Z$ be a sample space with a base measure $\mu$. We shall only consider measures which are absolutely continuous with respect to $\mu$. With this in mind we can identify positive measures with density functions, in the sense of the Radon-Nikodym Theorem [HS49]. This point will always be implicitly assumed throughout this paper, while the base measure is often implicit. Denote by $\mathcal{P}$ the space of all the probability measures on $Z$. Each $p \in \mathcal{P}$ is a distribution and can be denoted as $P(\cdot|p)$. Following [Ama85], we denote by $\widetilde{\mathcal{P}}$ the space of all positive normalisable measures on $Z$, ie., all the positive measures with a finite mass. Probability distributions are characterised by the fact that their mass is unity. Intuitively, $\mathcal{P}$ can be thought of as a section of a sphere, while $\widetilde{\mathcal{P}}$ can be thought of as a cone.

For each $p \in \mathcal{P}$, denote the log-likelihood function $l := \log p$. The space $\mathcal{P}$ can be regarded as an (infinite dimensional) manifold in the sense to be introduced below. [2] Here we shall

[2] Some regularity conditions are needed for these considerations to be meaningful; they are often trivially

only consider the finite dimensional case. This is technically easier to manage, intuitively more tractable, and provides motivation for our later definition for the infinite dimensional case. For any finite dimensional submanifold $\mathcal{Q}$ parameterised by $\theta^i$, the tangent space can be identified with the linear space $T\mathcal{Q}$ spanned by Fisher's "score functions" $\partial_i l$. Each tangent vector is a random variable, a function of the sample $z$, with zero mean.

$$\langle u \rangle = 0, \qquad \forall u \in T\mathcal{Q}. \tag{2.1}$$

Alternatively, $T\mathcal{Q}$ can be considered as a linear subspace of measures on $Z$.

A Riemannian metric can be introduced on $\mathcal{P}$ through the Fisher information matrix

$$g_{ij} := \langle \partial_i l \partial_j l \rangle . \tag{2.2}$$

The inner product for $u, v \in T\mathcal{P}$ is denoted as $\langle u, v \rangle$. This symbol is also used to denote covariance. Obviously there is no conflict between these two usages here.

The skewness tensor is a symmetric tensor of order three, defined as

$$T_{ijk} := \langle \partial_i l \partial_j l \partial_k l \rangle . \tag{2.3}$$

Let $\delta \in [0, 1]$, the $\delta$-connection on $\mathcal{P}$ is defined through corresponding Christoffel symbols,

$$\overset{\delta}{?}_{ijk} := \langle \partial_i \partial_j l \partial_k l \rangle + \delta T_{ijk}. \tag{2.4}$$

A whole bunch of related concepts, such as $\delta$-geodesic, $\delta$-parallel translate, $\delta$-curvature, $\delta$-flatness, $\delta$-convex, can be defined in the conventional way [Ama85]. It turns out that $\mathcal{P}$ is $\delta$-flat for $\delta \in \{0, 1\}$ but not for $\delta \in (0, 1)$. This will lead us, in sections to follow, to an extension of the $\delta$-geometry to $\tilde{\mathcal{P}}$, the space of positive measures.

For a distribution $p \in \tilde{\mathcal{P}}$, its $\delta$-representation (coordinate) is a measure given by

$$\overset{\delta}{l}(p) := \begin{cases} p^\delta/\delta, & \delta > 0, \\ \log p, & \delta = 0. \end{cases} \tag{2.5}$$

Denote by $\overset{1/\delta}{l}$ the inverse mapping of $\overset{\delta}{l}$. Let $\nu := 1/\delta$. The image of the mapping $\overset{\delta}{l}(p)$ is a member of the Banach space $L_\nu(Z)$, the space of all the $\nu$th power integrable functions on $Z$, since

$$\|\overset{\delta}{l}(p)\|_\nu^\nu = \int (\overset{\delta}{l}(p))^\nu = \int p/\delta^\nu < \infty. \tag{2.6}$$

If this is regarded as an (infinite dimensional) diffeomorphism, $\tilde{\mathcal{P}}$ can be identified with the positive cone of Banach space $L_\nu(Z)$, while $\mathcal{P}$ corresponds to the positive sector of a sphere

satisfied by commonly used families of probability distributions. We shall not consider regularity conditions in this report, except that we shall always make sure that we are not studying properties of members of the empty set.

(in the norm $\|\cdot\|_\nu$). The 0-representation is usually called the exponential representation ($e$-representation), while the 1-representation is usually called the mixture representation ($m$-representation). They are also called the likelihood function and the log-likelihood functions, respectively.

**Example 2.1** Consider a sample space of only three points, $Z = \{1, 2, 3\}$. Any function on $Z$ can be identified with a vector in $\mathbb{R}^3$. Any positive measure on $Z$ can be identified with a "positive vector" in $\mathbb{R}^3_+$. Consider a particular distribution $p \in \mathcal{P}$ on $Z$, defined by $p(1) = 1/2$, $p(2) = 1/3$, $p(3) = 1/6$. It can be denoted as $p = [1/2, 1/3, 1/6]$. This notation, however, is exactly the 1-coordinate, if, as usually the case, the base measure is chosen as $\mu = [1, 1, 1]$. That is, $\overset{1}{l}(p) = [1/2, 1/3, 1/6]$. Other representations are also legitimate. The 0-coordinate, for example, is $\overset{0}{l}(p) = [-\log 2, -\log 3, -\log 6]$.

Through the $\delta$-representation, the Riemannian metric and the $\delta$-divergence can be expressed as

$$g_{ij} = \int \partial_i \overset{\delta}{l}\, \partial_j \overset{1-\delta}{l}, \tag{2.7}$$

$$\overset{\delta}{?}_{ijk} = \int \partial_i \partial_j \overset{\delta}{l}\, \partial_k \overset{1-\delta}{l}. \tag{2.8}$$

The $\delta$ and $(1-\delta)$ divergences are dual to each other with respect to the Riemannian metric

$$u\nabla \langle v, w\rangle = \left\langle u\overset{\delta}{\nabla} v, w \right\rangle + \left\langle v, u \overset{1-\delta}{\nabla} w \right\rangle. \tag{2.9}$$

The symbol $\nabla$ at the left hand side is the uniquely defined covariant derivative operator. It is independent of the affine connections since $\langle v, w\rangle$ is a scalar field. The $\delta = 1/2$ connection is self-dual, and is therefore the Levi-Civita connection corresponding to the given Riemannian metric. The following holds as a consequence of the duality.

**Theorem 2.1** *There are dual-affine coordinates* $[\theta^i, \eta_i]$ *and dual potential fields* $\phi$ *and* $\psi$, *such that*

$$\eta_i = \partial_i \psi, \qquad g_{ij} = \partial_i \eta_j, \tag{2.10}$$

$$\theta^i = \partial^i \phi, \qquad g^{ij} = \partial^i \theta^j, \tag{2.11}$$

*where* $\partial_i := \partial/\partial\theta^i$ *and* $\partial^i := \partial/\partial\eta_i$. *The potentials* $\phi$ *and* $\psi$ *are related by the Legendre relation*

$$\psi + \phi = \theta^i \eta_i. \tag{2.12}$$

Many of the interesting dualities between $\delta$ and $(1-\delta)$ geometries on $\mathcal{P}$ [Ama85] can be attributed to the duality between corresponding Banach spaces, as we shall see below.

The $\delta$-divergence between two distributions $p, q \in \mathcal{P}$ is defined as

$$D_\delta(p,q) := \psi(p) + \phi(q) - \theta^i(p)\eta_i(q). \tag{2.13}$$

It turns out that

$$D_\delta(p,q) = \frac{1}{\delta(1-\delta)}\left(1 - \int p^\delta q^{1-\delta}\right).$$

This is obviously invariant of the $\delta$-representation. Let $r$ be any measure absolutely continuous with respect to $p+q$, then

$$\int p^\delta q^{1-\delta} = \int r(p/r)^\delta (q/r)^{1-\delta}, \tag{2.14}$$

where $p/r$ and $q/r$ are the Radon-Nikodym derivatives. The corresponding divergences for $\delta \in \{0,1\}$ are defined through limits. It turns out that $D_1$ is the cross entropy, $D_0$ is the reverse cross entropy, and $D_{1/2}(p,q)$ is the Hellinger distance.

Not only can the $\delta$-divergence be defined in terms of the $\delta$-connections, the converse is also true by way of the Eguchi relations [Egu83]. The $\delta$-divergence is unique in a certain sense, as discussed in [Ama85, p.96–100]: By some invariance considerations, the most general definition of information divergence is the $f$-divergence. However, since $\delta$-connections are the only invariant connections, the $\delta$-divergence is unique if the Eguchi relations are to hold.

These conditions are weaker than Shannon's conditions for the definition of entropy[Sha48]. If an additive constraint is added to the definition of $\delta$-divergence, then $\delta = 1$, and the definition of cross entropy is obtained [Kul59]. A further constraint relating to the uniform distribution gives Shannon's definition of entropy (with a linear transform).

The most important consequence of these definitions is the "Generalised Pythagorean Theorem"[Ama85]. It can be presented in many forms, an important one of them is as follows.

**Theorem 2.2 (Generalised Pythagorean Theorem)** *Let $\mathcal{P}$ be a $\delta$-flat manifold. Let $\mathcal{Q}$ be $(1-\delta)$-convex submanifold of $\mathcal{P}$. Then for an arbitrary point $p \in \mathcal{P}$, there is a unique point $\hat{q} \in \mathcal{Q}$ which minimises $D_\delta(p, \hat{q})$. This is called the $\delta$-projection of $p$ onto $\mathcal{Q}$. For any $q \in Q$,*

$$D_\delta(p,q) = D_\delta(p,\hat{q}) + D_\delta(\hat{q},q). \tag{2.15}$$

This is analogous to the linear projection of quadratic distance. In fact, we shall see later that it is a proper generalisation to the later.

# 3 Coherent Invariant Measurements of Generalisation

Given sample space $Z$, consider the space $\mathcal{P} := \mathcal{P}(Z)$ of all the probability measures on $Z$. [3] Each $p \in \mathcal{P}$, corresponding to $P(\cdot|p)$, is a distribution from which the data could have been drawn.

A statistical model is a subspace $\mathcal{Q}$ of $\mathcal{P}$. In the case of neural networks, $\mathcal{Q}$ is always a proper subspace of $\mathcal{Q}$, restricted by the fact that input distribution $P(x)$ cannot be freely changed, and that not all conditional distributions $P(y|x)$ can be represented by the network.

A sample, or a data set, of size $n$, is a point $z^n = [z_1, \ldots, z_n] \in Z^n$. We shall only be interested in independent data in this paper. That is, we shall always assume $P(z^n|p) = \prod_i P(z_i|p)$. [4] Since the sample size $n$ is always available, there is a trivial one-one correspondence between $P(z^n|p)$ and $P(z_i|p)$. Therefore, to avoid excessive use of complicated notation, we shall suppress explicit distinction between these two. The formulas in the rest of this paper, except in the following paragraph acting as an example, look as if both the training data and the test data consist of only a single point, although they are applicable to the general case, and it is always quite obvious how to write them in full detail.

A learning rule is a mapping

$$\tau : \; Z \to \mathcal{Q}, \tag{3.1}$$

which maps the observed data $z \in Z$ generated by an unknown distribution $p \in \mathcal{P}$ to the estimated distribution $q = \tau(z) \in \mathcal{Q}$. In view of the convention on notation just introduced, this definition actually means the following. Let

$$\mathcal{Z} := \bigcup_{n=1}^{\infty} Z^n, \tag{3.2}$$

A learning rule is a mapping $\tau : \mathcal{Z} \to \mathcal{Q}$. For each data size $n$ and each observed data set $z^n \in Z^n$, generated by $p^n \in \mathcal{P}^n$, where $p^n$ is defined by $P(z^n|p^n) = \prod_i P(z_i|p)$, the learning rule $\tau$ maps $z^n$ to an estimate $q = \tau(z^n) \in \mathcal{Q}$. Such explanation will not be repeated in the rest of the paper. See [Ama87] for notation in which the sample size is explicit. The above discussion is schematically illustrated in Figure 5.

How does one measure the performance of a learning rule? It is obvious that the performance should be measured on test data which is independent of the training data, for otherwise a learning rule which uses the "empirical distribution" of the training data as the estimate would be regarded as optimal [GBD92]. However, the test data is a random

[3] As mentioned before, we only consider measures which are absolutely continuous with respect to a base measure $\mu$ on $Z$

[4] The case of non-independent data is usually called "temporal learning", and requires techniques such as the adaptive critic [BSA83, BSW90]. The mathematical formulation is different, although statistical estimators can act as modules in an adaptive critic architecture [Zhu93].

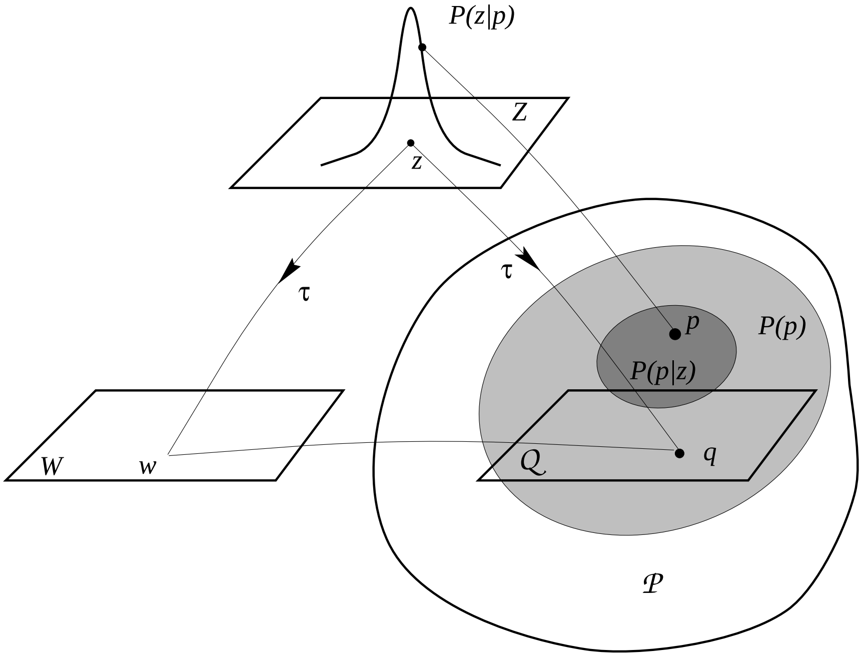

Figure 5: Relation between weight space, learning rule, and Bayes Theorem

variable. Any particular sample test data cannot be directly used to measure the performance, for that will label a learning rule which always produces the same test data as the optimal one, although admittedly such a learning rule is very unlikely to be hit upon before the test data is known. Therefore it is clear that what we really want is to calculate the "divergence", between the true distribution which will generate all the future test data and the model distribution given by the learning rule. Of course the true distribution is not known. Therefore we take the posterior average of the divergence as a performance measure for the learning rule. In other words, a learning rule should not only be able to make a good prediction by training on a particular data set; It should be able to do so for any training set which naturally arises in a given environment. These considerations lead to the following definition.

**Definition** 3.1 (Generalisation error) The measurement of generalisation of learning rule $\tau$ is defined as the expected divergence between the true distribution $p$ and the model distribution $q = \tau(z)$ given by the rule, averaged over any possible true distribution $p$ and data $z$,

$$E_\delta(\tau) := \langle D_\delta(p, \tau(z)) \rangle = \int_p P(p) \int_z P(z|p) D_\delta(p, \tau(z)). \tag{3.3}$$

A learning rule $\tau$ which minimises $E_\delta(\tau)$ is called a $\delta$-optimal learning rule (estimator). It is denoted $\tau_\delta$ if it is unique.

Given some observed data $z$, we also want to choose an estimate that minimises the posterior expected divergence between the true distribution $p$ and the model distribution $q$, averaged over all possible $p$,

$$E_\delta(q|z) := \langle D_\delta(p,q) \rangle_z = \int_p P(p|z) D_\delta(p,q), \tag{3.4}$$

An estimate $q \in \mathcal{Q}$ which minimises $E_\delta(q|z)$ is called a $\delta$-optimal estimate based on $z$.

It is important to note that the definitions of $E_\delta(\tau)$ and $E_\delta(q|z)$ depend on the prior $P(p)$. A learning rule which is optimal in one environment may not be good in another environment.

It is easy to verify that these two criteria are compatible with each other in the sense of the following theorem.

**Theorem 3.1 (Coherence of optimality of estimators and estimates)** *A learning rule $\tau$ is $\delta$-optimal if and only if for any data $z$, excluding a subset of zero probability, $\tau(z)$ is a $\delta$-optimal estimate based on $z$.*

***Proof:*** By applying Bayes' rule, we have

$$E_\delta(\tau) = \int_z P(z) \int_p P(p|z) D_\delta(p, \tau(z)) = \int_z P(z) E_\delta(\tau(z)|z) \tag{3.5}$$

The conclusion then directly follows the fact that the divergence is non-negative. □

The inferences by Bayes rule, and only such inferences, are coherent, in the sense that the information carried in the posterior is exactly the combined information carried in the prior and the data (equivalently in the likelihood function). An estimator is regarded as optimal if and only if it gives optimal estimates for almost all the data. The $\delta$-divergences are invariant with respect to invertible transforms in $W$ and $Z$. Therefore the generalisation measurement defined above is both coherent and invariant. This resolves the controversy in [Mac92, Wol93, Mac93]. We shall come back to this point later (§8).

**Theorem 3.2 (Projected Average Theorem)** *Let $\mathcal{Q}$ be a $\delta$-flat manifold. Let $P(p)$ be a prior on $\mathcal{Q}$. Then $\forall q \in \mathcal{Q}$, $\forall z \in Z$,*

$$E_\delta(q|z) = E_\delta(\widehat{p}|z) + D_\delta(\widehat{p}, q), \tag{3.6}$$

*where $\widehat{p}$ is the $\delta$-optimal estimate in $\mathcal{Q}$.*

***Proof:*** Let $(\theta, \eta)$ be the $(\delta, 1-\delta)$-dual coordinates. Then the theorem is equivalent to the vanishing of the following entity,

$$\langle D_\delta(p,q) - D_\delta(p,\widehat{p}) - D_\delta(\widehat{p},q) \rangle_z = \left\langle (\theta^i(p) - \theta^i(\widehat{p}))(\eta_i(\widehat{p}) - \eta_i(q)) \right\rangle_z \tag{3.7}$$

Define $q_t$ by its $\eta$ coordinates, for $t$ in a small neighbourhood of 0,

$$\eta(q_t) := \eta(\widehat{p}) + t(\eta(q) - \eta(\widehat{p})). \tag{3.8}$$

Then $q_t \in \mathcal{Q}$. For $|t| \ll 1$, since $D_\delta(\widehat{p}, q_t)$ is quadratic in $t$, and non-negative[Ama85],

$$\begin{aligned} \langle D_\delta(p, q_t) - D_\delta(p, \widehat{p}) \rangle_x &= D_\delta(\widehat{p}, q_t) + \left\langle \theta^i(p) - \theta^i(\widehat{p}) \right\rangle_x (\eta_i(\widehat{p}) - \eta_i(q_t)) \\ &\approx At^2 + Bt \geq 0, \end{aligned}$$

it follows that $B = 0$, or equivalently

$$\left\langle \theta^i(p) - \theta^i(\widehat{p}) \right\rangle_x \cdot (\eta_i(\widehat{p}) - \eta_i(q)) = 0. \tag{3.9}$$

This shows that the quantity (3.7) is identically zero. This completes the proof. □

This theorem is a direct generalisation of the well-known corresponding result for least squares estimates. Let $p$ and $q$ be Gaussians with means $a$ and $b$, respectively, and the same fixed variance. Let $\widehat{a}$ be defined by minimising $\langle \|a - \widehat{a}\|^2 \rangle_z$. Then the above theorem reduces to the following

$$\left\langle \|a - b\|^2 \right\rangle_z = \left\langle \|a - \widehat{a}\|^2 \right\rangle_z + \|\widehat{a} - b\|^2. \tag{3.10}$$

The above proof itself is also a generalisation of the usual proof for the least squares estimate.

## 4 Optimal Estimators on the Whole Probability Space

As was mentioned previously, neural network models are proper subspaces of $\mathcal{P}$. However, in order to gain some intuition about the above abstract definitions, it is beneficial to find out the $\delta$-optimal estimates for the unrestricted case. That is, in this section, we assume $\mathcal{Q} = \mathcal{P}$.

**Theorem 4.1** *Given a prior $P(p)$ on $\mathcal{P}$, and data $z$, the $\delta$-optimal estimate $q$ based on $z$ can be obtained by taking the average of the $\delta$-representations of the true distribution over the posterior as the $\delta$-representation of $q$ and then normalising the results. That is,*

$$\operatorname*{Min}_{q \in \mathcal{P}} E_\delta(q|z) \iff q \sim \overset{1/\delta}{l}\left( \left\langle \overset{\delta}{l}(p) \right\rangle_z \right), \tag{4.1}$$

*for any data $z$ excluding a set of zero probability, where $\langle \cdot \rangle_z$ denotes the posterior expectation.*

***Proof:*** This is proved in [ZR95a], by a variational argument. □

The reason that the $\delta$-optimal estimate is only proportional to the positive measure obtained by averaging the $\delta$-coordinates over the posterior is that probabilities are normalised. Later we shall extend the $\delta$-geometry to the space of normalisable positive measures, $\widetilde{\mathcal{P}}$. It will be seen that the mathematics become much simpler, and the intuition much clearer. The unfamiliar $\delta$-geometry will be connected with the familiar geometry of the Banach space $L_{1/\delta}$. It will be shown that projection from $\widetilde{\mathcal{P}}$ to $\mathcal{P}$ is always achieved by renormalisation, so that the above theorem becomes a simple corollary.

It is interesting to observe that for $\delta = 1$, the $\delta$-estimate reduces to $q = \langle p \rangle_z$. In other words, the 1-optimal estimate is simply the posterior marginal distribution. The Monte Carlo method proposed by R. Neal[Nea93] consists of sampling from the posterior of $w$ for each instance of $z$, which gives a predictive distribution of $z$ which is exactly the posterior marginal distribution.

Two examples from classical statistics will help to clarify these discussions. The detailed derivations are given in [ZR95b, ZR95a].

**Example 4.1** Consider the multinomial family of distributions

$$M(m|p), \qquad m \in \mathbb{N}^n, \ p \in \mathcal{P} = \Delta^{n-1}, \tag{4.2}$$

with a Dirichlet prior

$$D(p|a), \qquad a \in \mathbb{R}^n_+. \tag{4.3}$$

The posterior is also a Dirichlet distribution $D(p|a+m)$. The $\delta$-optimal estimate $\widehat{q} \in \mathcal{P}$ is given by

$$(\widehat{q}_i)^\delta \sim (a_i + m_i)_\delta, \tag{4.4}$$

where $(a)_b := \Gamma(a+b)/\Gamma(a)$. In particular,

$$\widehat{q}_i = (m_i + a_i) / \sum_j (m_j + a_j), \qquad \delta = 1, \tag{4.5}$$

$$\widehat{q}_i \sim \exp(\Psi(a_i + m_i)), \qquad \delta = 0, \tag{4.6}$$

where $\Psi$ is the the digamma function, the logarithmic derivative of $\Gamma$ function.

**Example 4.2** Consider the Gaussian family of distributions

$$f(z|\mu) = N(z - \mu|h), \qquad z, \mu \in \mathbb{R}, \ h \in \mathbb{R}_+, \tag{4.7}$$

with fixed variance $\sigma^2 = 1/h$. Let the prior be another Gaussian

$$f(\mu) = N(\mu - a_0|n_0 h), \qquad a_0 \in \mathbb{R}, \ n_0 \in \mathbb{R}_+. \tag{4.8}$$

Then the posterior after seeing the sample $z^k = [z_1, \dots, z_k]$ is also a Gaussian

$$f(\mu|z) = N(\mu - a_k|n_k h), \qquad a_k = \left(\sum z + n_0 a_0\right)/(n_0 + k), \tag{4.9}$$

where $a_1$ is also the posterior least squares estimate. The $\delta$-optimal estimate $\widehat{q}$ is given by the density

$$f(z'|\widehat{q}) = N\left(z' - a_k \middle| \frac{h}{1 + \delta/n_k}\right). \tag{4.10}$$

This show that if we are only interested in estimating the mean of a Gaussian, then the least squares estimate should be used, whatever the choice of $\delta$. Note that the $\delta$-optimal estimate is not a member of the original Gaussian family ($h' \neq h$) unless $\delta = 0$, since this family is only 0-flat.

The entities $|a|$ for the Dirichlet prior and $n$ for the Gaussian prior are effective previous sample sizes. This fact was of course well known since Fisher's time. In a restricted model, the sample size might not be well reflected, and some ancillary statistics, as introduced by Fisher, may be used for information recovery.

## 5 Information Geometry — Extension to Positive Measures

As alluded in the previous section, the procedure for obtaining the $\delta$-optimal estimate is not entirely aesthetically appealing. Intuitively, from the Averaged Projection Theorem, the $\delta$-estimate should be obtained simply by averaging the $\delta$-coordinates. However, since we have been seeking $\delta$-optimal estimates in $\mathcal{P}$, the space of positive measures with total mass of unity, a renormalisation is necessary. Intuitively, the space $\mathcal{P}$ looks like a (very high dimensional) sphere. Averaging on the sphere will in general result in a point in the interior of the ball. (Figure 6) Renormalisation is the projection back onto the sphere. If we are ready to accept all positive measures as possible candidates of a statistical estimation problem, the mathematical machinery becomes much simpler. There is nothing lost in this change of concept, since in practice only the proportionality of measures are necessary for applications of probability theory. [5] A renormalisation can always be performed at the final stage if such an answer is sought, but it will be seen that positive measures provide a more versatile tool to hold intermediate results. [6]

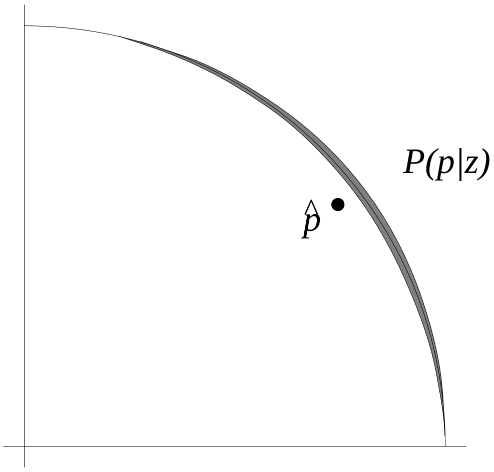

Figure 6: The $\delta$-average of the posterior is not on $\mathcal{P}$

The extension of $\delta$-geometry to $\tilde{\mathcal{P}}$ follows suggestions in [Ama85]. The resulting theory is surprisingly elegant. It also provides an explicit link between the duality between the

[5] This is evident from all the Monte Carlo algorithms, such as Gibbs samplers.

[6] One of the most important techniques in stochastic computation, the Gibbs sampler [MRR$^+$53, KGV83, GG84, Yor92], only requires a detailed balance [Ami89], which is expressible as ratios between probability distributions.

geometry of statistical manifolds, and the thoroughly studied duality between Banach spaces (complete normed linear spaces).

We shall present the theory for a finite sample space $Z$. The general situation is quite similar, although the notation will be more abstract. [7] We shall state the corresponding results for an infinite sample space, however, without rigorous proof.

Following [Ama85], the extension to $\widetilde{\mathcal{P}}$ is most conveniently carried out through the $\delta$-representatives. We shall first consider $\delta \in (0,1)$, and then extend the results to $\delta \in \{0,1\}$. Note that the tangent vectors of $\widetilde{\mathcal{P}}$ are no longer zero mean random variables. If a vector $u \in T\widetilde{\mathcal{P}}$ has a positive mean, for example, it points in the direction of increasing total mass, and vice versa.

For $p \in \widetilde{\mathcal{P}}$, let $[\theta, \eta]$ be the $[\delta, 1-\delta]$-dual coordinates,

$$\theta^i := \overset{\delta}{l}_i = \frac{p_i^\delta}{\delta}, \qquad \eta_i := \overset{1-\delta}{l}_i = \frac{p_i^{1-\delta}}{1-\delta}. \tag{5.1}$$

It follows easily that

$$\partial_i \theta^j = \mathbf{1}_i^j, \qquad \partial_i \eta_j = p_i^{1-2\delta} \mathbf{1}_{ij}. \tag{5.2}$$

The Riemannian metric is defined as

$$g_{ij} := \int \partial_i \overset{\delta}{l}\, \partial_j \overset{1-\delta}{l} = \partial_j \eta_i = p_i^{1-2\delta} \mathbf{1}_{ij}. \tag{5.3}$$

Note that the alternative formula $g_{ij} = -\langle \partial_i \partial_j l \rangle$, which depends on $\langle \partial_i l \rangle = 0$, is no longer true.

For any parameterisation, the $\delta$-connection $\overset{\delta}{\nabla}$ is defined through

$$\begin{aligned}
\left\langle u \overset{\delta}{\nabla} v w \right\rangle :&= \int u \partial (v \partial \overset{\delta}{l})\, w \partial \overset{1-\delta}{l} \\
&= \int u^i \partial_i (v^j \partial_j \overset{\delta}{l})\, w^k \partial_k \overset{1-\delta}{l} \\
&= \int \left( u^i \partial_i v^j \partial_j \overset{\delta}{l} + u^i v^i \partial_i \partial_j \overset{\delta}{l} \right) w^k \partial_k \overset{1-\delta}{l} \\
&= \left\langle \left( u^i \partial_i v^j \partial_j l + u^i v^i \partial_i \partial_j l + \delta u^i \partial_i l v^i \partial_j l \right) w^k \partial_k l \right\rangle \\
&= u^i \partial_i v^j w^k g_{jk} + u^i v^j w^k \overset{\delta}{?}_{ijk} \\
&= u^i \overset{\delta}{\nabla}_i v^k w_k,
\end{aligned}$$

[7] Finite dimensional linear algebra can be conveniently presented in either coordinate-dependent or coordinate-free fashion. The infinite dimensional counterpart can only be reasonably well presented in a coordinate-free fashion. However, such description might obscure the geometric meaning to those not familiar with such presentations.

where

$$(5.4) \qquad \overset{\delta}{?}_{ijk} := \left\langle\left(\partial_i\partial_j l + \delta\partial_i l\partial_j l\right)\partial_k l\right\rangle, \qquad \overset{\delta}{\nabla}_i v^k := \partial_i v^j + v^j \overset{\delta}{?}_{ij}{}^k.$$

It has simple formula under the $\delta$-representation,

$$\begin{aligned} \left\langle u\overset{\delta}{\nabla} v w\right\rangle &= \int u^i\partial_i(v^j\partial_j\overset{\delta}{l})\, w^k\partial_k \overset{1-\delta}{l} \\ &= \int\left(u^i\partial_i v^j\partial_j\overset{\delta}{l} + u^i v^i\partial_i\partial_j\overset{\delta}{l}\right)\; w^k\partial_k\overset{1-\delta}{l} \\ &= \int u^i\partial_i v^j\partial_j\overset{\delta}{l}\, w^k\partial_k\overset{1-\delta}{l} \\ &= u^i\partial_i v^j w_j. \end{aligned}$$

This implies $\overset{\delta}{?}_{ij}{}^k = 0$. Therefore the coordinate curves of $\delta$-representation are $\delta$-geodesics,

$$(5.5) \qquad \partial_i\partial_j\theta^k = 0 \implies e_i\overset{\delta}{\nabla} e_j = 0.$$

Since the space $\widetilde{\mathcal{P}}$ is $\delta$-affine, the $\delta$-potential $\psi$ can be obtained by integration

$$(5.6) \qquad \psi(p) = \int \eta_k(p)d\theta^k(p) = \int\sum_k \frac{p_k^{1-\delta}}{1-\delta}\, d\frac{p_k^{\delta}}{\delta} = \int\sum_k \frac{1}{1-\delta}\, dp_k = \frac{\sum_k p_k}{1-\delta}.$$

It can be verified that this indeed satisfies (by $\partial_i p_i = p_i^{1-\delta}$)

$$(5.7) \qquad \partial_i\psi = \eta_i.$$

Since $[\theta, \eta]$ are $[\delta, 1-\delta]$ dual affine coordinates with potentials

$$(5.8) \qquad \psi(p) = \frac{\sum_k p_k}{1-\delta}, \qquad \phi(p) = \frac{\sum_k p_k}{\delta},$$

the $\delta$-divergence between two distributions $p, q \in \widetilde{\mathcal{P}}$ is well defined, [8]

$$D_\delta(p, q) := \psi(p) + \phi(q) - \theta^i(p)\eta_i(q) = \frac{\sum p}{1-\delta} + \frac{\sum q}{\delta} - \frac{\sum p^\delta q^{1-\delta}}{\delta(1-\delta)}.$$

In the case of continuous distributions, the sum should be replaced by integration, and we *define* the $\delta$-divergence between $p, q \in \widetilde{\mathcal{P}}(Z)$ for any measurable space $Z$ to be

$$D_\delta(p, q) := \frac{\int p}{1-\delta} + \frac{\int q}{\delta} - \frac{\int p^\delta q^{1-\delta}}{\delta(1-\delta)}.$$

[8] We were informed during the MANNA conference that Prof. S. Amari had independently obtained results of the same nature, which, after all, is a simple corollary from his book.

The corresponding divergences for $\delta \in \{0,1\}$ are defined through limits. It is easy to verify that

$$\lim_{\delta\to 0} \frac{1}{\delta(1-\delta)} \left(\delta p + (1-\delta)q - p^{\delta} q^{1-\delta}\right) = p - q + q \log \frac{q}{p}. \tag{5.9}$$

Therefore the correct generalisation of the Kullback-Leibler-divergence to $\widetilde{\mathcal{P}}$ is

$$K(p,q) := \int \left( q - p + p \log \frac{p}{q} \right). \tag{5.10}$$

It can be easily verified that $D_\delta$ is invariant with respect to the base measure. It is also obvious that this definition is compatible to the previous definition for $p, q \in \mathcal{P}$, since in this situation,

$$\frac{\int p}{1-\delta} + \frac{\int q}{\delta} = \frac{1}{\delta} + \frac{1}{1-\delta} = \frac{1}{\delta(1-\delta)}.$$

We shall prove in the next section that $D_\delta(p,q)$ is approach the same quadratic form when $p$ and $q$ are close to each other, independent of $\delta$. This defines a topology on $\widetilde{\mathcal{P}}$. Since $D_{1/2}$ is a squared Hilbert-norm, $\widetilde{\mathcal{P}}$ becomes an infinite dimensional manifold modelled on a Hilbert space [AMR83]. Similar ideas for considering $\widetilde{\mathcal{P}}$ as infinite dimensional manifold appeared more or less explicitly in [Ama85, Ama87].

By defining the $\delta$-divergence this way, we have largely avoided the issue of regularity conditions. In the following sections we shall prove several theorems which would be quite difficult under a more general treatment of infinite dimensional manifolds, but are quite elementary under the current definition. On the other hand, the $\delta$-divergence obviously introduces a topology on $\widetilde{\mathcal{P}}$, by which $\overset{\delta}{l}$ are diffeomorphisms. The Banach spaces $L_{1/\delta}$ can be viewed as tangent spaces. It is not clear to us at the moment in what sense transformations among these Banach spaces can be viewed as diffeomorphisms, but we note that most of the commonly used Banach spaces are isomorphic to each other [Tri83].

The drawback is that we cannot discuss the uniqueness or generality of such a definition. It is adopted merely because it is useful. With such a definition of $\delta$-divergence, we can work backwards and define the Riemannian metric and $\delta$-connections by using the Eguchi relations[Egu83]

$$\langle u, v \rangle := u\underset{p}{\nabla}\, v\underset{q}{\nabla}\, D_\delta(p,q)\Big|_{q=p}, \tag{5.11}$$

$$\left\langle u\overset{\delta}{\nabla} v, w \right\rangle := u\underset{p}{\nabla}\, v\underset{p}{\nabla}\, w\underset{q}{\nabla}\, D_\delta(p,q)\Big|_{q=p}, \tag{5.12}$$

where $u\nabla$, etc., are Frechet derivatives in Banach spaces. We have not studied the details of the regularity conditions required for these formal manipulations to be rigorous. Since our main objective here is to find a definition of information divergence, we shall not be concerned with the exact meaning of other aspects of information geometry.

## 6 Properties of Information Divergence on $\widetilde{\mathcal{P}}$

**Theorem 6.1** *The following properties hold for $\delta \in (0,1)$ and $p, q \in \tilde{\mathcal{P}}$.*

1. *The $\delta$-divergence is non-negative:*

$$D_\delta(p,q) \geq 0. \tag{6.1}$$

2. *The $\delta$-divergence of two measures is zero if and only if they are identical to each other,*

$$D_\delta(p,q) = 0 \iff p = q. \tag{6.2}$$

3. *The $\delta$-divergence is approximately quadratic when two distributions are close to each other, independent of $\delta$.*

$$D_\delta(p, p+\Delta p) \approx \frac{1}{2}\int \frac{\Delta p^2}{p}, \tag{6.3}$$

*where right hand side is usually called the $\chi^2$ distance [Kas89].*

4. *The $\delta$ and $(1-\delta)$-divergences are dual to each other:*

$$D_\delta(p,q) = D_{1-\delta}(q,p). \tag{6.4}$$

5. *The $\delta$-divergence is invariant of reparameterisation. That is, $D_\delta(p,q)$ does not depend on the base measure $\mu$ implicit in its definition.*

6. *The $\delta$-divergence is homogeneous in both arguments.*

$$D_\delta(ap, aq) = aD_\delta(p,q), \qquad \forall a \in \mathbb{R}_+. \tag{6.5}$$

***Proof:***

1. The integrand in the definition of the $\delta$-divergence is a positive measure, since it is the difference between an arithmetic mean and a corresponding geometric mean, hence alway positive. The situation for $\delta = 1$ is also easy to prove, using the inequality given in Figure 7.

2. The mass of a positive measure is zero only when the measure is itself zero. The difference between the arithmetic mean and geometric mean of $p$ and $q$ can only be zero when $p = q$.

3. Expansion in terms of $\Delta p$ gives

$$\delta p + (1-\delta)(p+\Delta p) - p^\delta (p+\Delta p)^{1-\delta} \approx \frac{\delta(1-\delta)}{2}\int \frac{\Delta p^2}{p} + O\left(\left(\frac{\Delta p}{p}\right)^3\right). \tag{6.6}$$

4. Obvious from definition.

5. Use (2.14).

6. Obvious from definition.

This completes the proof of the theorem. □

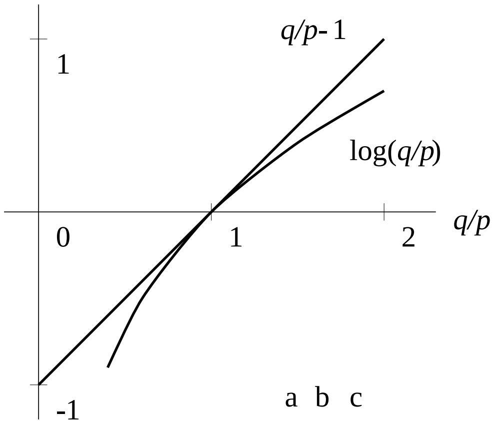

Figure 7: Positiveness of $q/p - 1 - \log(q/p)$

The positiveness of the integrand in the definition of the $\delta$-divergence has the additional advantage of not only telling us the total divergence between any two positive measures, but also telling us the divergence on each event (measurable subsets of $Z$). For example, in Figure 8, it is possible to speak of the divergence between $p$ and $q$ over $E$.

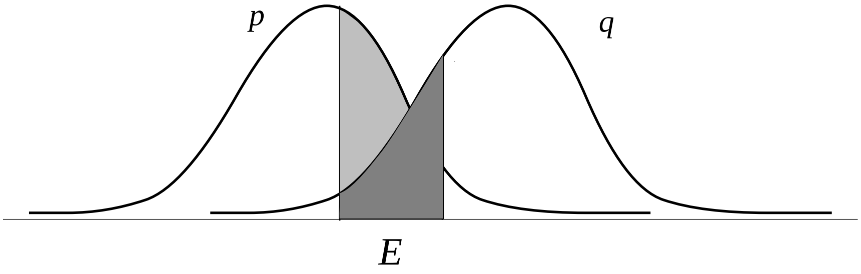

Figure 8: Divergence between two measures on a event

Since the space $\widetilde{\mathcal{P}}$ is $\delta$-flat for any $\delta$, the Projected Average Theorem automatically holds. However, it has a more conventional proof which is analogous to the corresponding theorem for linear models. In fact, it is a direct generalisation of the corresponding proof for $\delta = 1/2$

**Theorem 6.2 (Projected Average Theorem)** *Let $P(p)$ be a distribution on $\widetilde{\mathcal{P}}$. Let $q \in \widetilde{\mathcal{P}}$. Then*

$$\langle D_\delta(p, q) \rangle = \langle D_\delta(p, \widehat{p}) \rangle + D_\delta(\widehat{p}, q), \tag{6.7}$$

*where $\widehat{p}$ is the $\delta$-average of $p$ defined as*

$$\widehat{p}^{\delta} := \left\langle p^{\delta} \right\rangle . \tag{6.8}$$

***Proof:*** This is proved by straightforward evaluation. In doing the following calculations, it is important to remember that $p$ is a random quantity while $\widehat{p}$ and $q$ are not. From the definition of $\widehat{p}$, it is easy to show that

$$\begin{aligned}\delta(1-\delta)\left\langle D_\delta(p,\widehat{p})\right\rangle &= \delta \int \langle p \rangle + (1-\delta)\int \widehat{p} - \int \left\langle p^{\delta}\right\rangle \widehat{p}^{1-\delta} \\ &= \delta \int \langle p \rangle + (1-\delta)\int \widehat{p} - \int \widehat{p}^{\delta}\widehat{p}^{1-\delta} \\ &= \delta \int \langle p \rangle - \delta \int \widehat{p}.\end{aligned}$$

Therefore,

$$\begin{aligned}\delta(1-\delta)\left\langle D_\delta(p,q)\right\rangle &= \delta \int \langle p \rangle + (1-\delta)\int q - \int \left\langle p^{\delta}\right\rangle q^{1-\delta} \\ &= \delta \int \langle p \rangle - \delta \int \widehat{p} + \delta \int \widehat{p} + (1-\delta)\int q - \int \widehat{p}^{\delta} q^{1-\delta} \\ &= \delta(1-\delta)\left\langle D_\delta(p,\widehat{p})\right\rangle + \delta(1-\delta) D_\delta(\widehat{p},q).\end{aligned}$$

This proves the theorem. □

**Corollary 6.3** *Given sample $z \in Z$ and a prior $P(p)$ over $\widetilde{\mathcal{P}}$, the $\delta$-estimate $q = \tau(z)$ in $\widetilde{\mathcal{P}}$ is given by the $\delta$-average over the posterior,*

$$q^{\delta} = \left\langle p^{\delta} \right\rangle_z . \tag{6.9}$$

***Proof:*** Substitute the posterior $P(p|z)$ into the previous theorem. □

This generalises the classical result that the least mean square estimate is given by $\widehat{a} = \langle a \rangle_z$. Therefore we immediately obtains an extension of the Kalman filter to any $\delta$-flat statistical models. The drawback is that, as a Kalman filter requires a matrix inversion at each step, the generalised algorithm requires the use of Bayes' rule at each step. How difficult this is depends on the exact model.

**Example 6.1** Return to the multinomial family of distributions as considered before. The $\delta$-optimal estimate $q \in \mathcal{P}$ is given by

$$(q_i)^{\delta} = (a_i + m_i)_\delta / (|a+m|)_\delta, \tag{6.10}$$

where $|a| := \sum_i a_i$. In particular,

$$q_i = (a_i + m_i)/|a+m|, \qquad \delta = 1, \tag{6.11}$$

$$\log q_i = \Psi(a_i + m_i) - \Psi(|a+m|), \qquad \delta = 0. \tag{6.12}$$

# 7 Optimal Estimators for Restricted Models

As said earlier, neural networks are statistical models such that $\mathcal{Q}$ is a proper subspace of $\mathcal{P}$. In this section we shall discuss such situations.

**Theorem 7.1 (Decomposition of errors)** *Let $\mathcal{Q} \subseteq \widetilde{\mathcal{P}}$ be a $(1-\delta)$-convex submanifold. Let $P(p)$ be a prior on $\mathcal{P} \subseteq \widetilde{\mathcal{P}}$. Denote by $\widehat{p}$ and $\widehat{q}$ the $\delta$-optimal estimates in $\widetilde{\mathcal{P}}$ and $\mathcal{Q}$, respectively, based on data $z \in Z$. Then*

$$E_\delta(q|z) = E_\delta(\widehat{p}|z) + D_\delta(\widehat{p}, q) \tag{7.1}$$

$$= \langle D_\delta(p, \widehat{p}) \rangle_z + D_\delta(\widehat{p}, \widehat{q}) + D_\delta(\widehat{q}, q). \tag{7.2}$$

This is an immediate generalisation of least mean square estimates for linear models. Consider the family of normal distributions with a fixed covariance matrix $\Sigma$. Let $\|\cdot\|$ be the norm associated with the inner product defined by $\Sigma^{-1}$. Then the above reduces to

$$\left\langle \|a-b\|^2 \right\rangle_z = \left\langle \|a-\widehat{a}\|^2 \right\rangle_z + \|\widehat{a}-b\|^2 \tag{7.3}$$

$$= \left\langle \|a-\widehat{a}\|^2 \right\rangle_z + \|\widehat{a}-\widehat{b}\|^2 + \|\widehat{b}-b\|^2, \tag{7.4}$$

where $a, b, \widehat{a}, \widehat{b}$ are posterior mean of $p, q, \widehat{p}, \widehat{q}$, respectively.

The practical interpretations of these quantities can be summarised as follows:

- The manifold $\mathcal{P}$ is the mathematical model. It is usually infinite dimensional.
- The manifold $\mathcal{Q}$ is the computational model. It is always finite dimensional.
- $p \in \mathcal{P}$ is the true distribution which generates the data. It is random even when conditional on the data. In other words, it is unknown even after the data is observed.
- $\widehat{p} \in \widetilde{\mathcal{P}}$ is the ideal $\delta$-estimate, which is not random conditional on the data. It is the $\delta$-average of the posterior $P(p|z)$.
- $\widehat{q} \in Q$ is the $\delta$-estimate within the model. It is the $\delta$-projection of $\widehat{p}$ onto $\mathcal{Q}$.
- $q \in \mathcal{Q}$ is the estimate which is actually computed.

By the theorem for the decomposition of errors, the total error of an estimate consists of three "orthogonal" components: the statistical error $\langle D_\delta(p, \widehat{p}) \rangle_z$ caused by the lack of knowledge about the true $p$ from finite sample $z$; the approximation error $D_\delta(\widehat{p}, \widehat{q})$ caused by the fact that the ideal estimate is not representable by the given model; and the avoidable error $D_\delta(\widehat{q}, q)$. It follows that the most one can do is to reduce $D_\delta(\widehat{q}, q)$. This is shown in Figure 9.

In summary, we introduce the following terms:

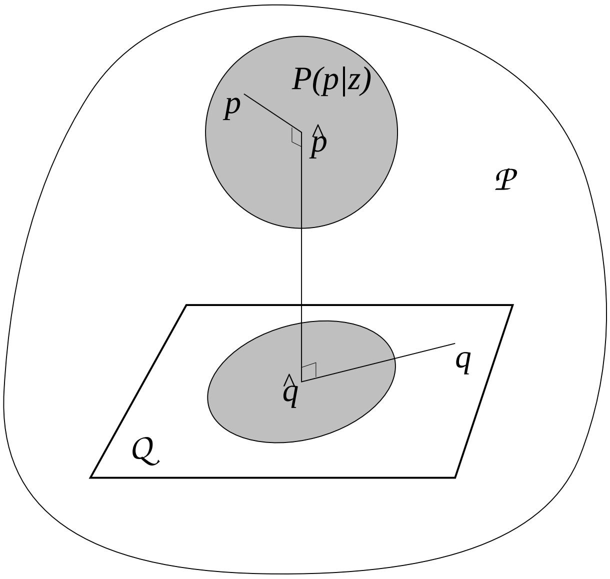

Figure 9: Decomposition of errors

- $E_\delta(q|z) = \langle D_\delta(p,q)\rangle_z$ — total error.
- $E_\delta(\widehat{p}|z) = \langle D_\delta(p,\widehat{p})\rangle_z$ — statistical error, intrinsic uncertainty (minimum total error).
- $D_\delta(\widehat{p},q)$ — additional error.
- $D_\delta(\widehat{p},\widehat{q})$ — approximation error (minimum additional error for the model).
- $D_\delta(\widehat{q},q)$ — Avoidable error.

It appears that many commonly used statistical techniques can be expressed in terms of these concepts:

- "Test of significance", "alternative-free" test — Compare additional error with intrinsic error.
- "Classification", "test among two alternatives", "goodness of fit test" — Compare total error of several models.
- "Model adequacy" — Compare approximation error.
- "Cross validation" — Sample estimate of total error.

Since multilayer neural networks are usually not $\delta$-convex for any $\delta$, there may exist local optima of $E_\delta(\cdot|z)$ on $\mathcal{Q}$. Let $p \in \mathcal{P}$ be the environmental distribution. Let $q \in \mathcal{Q}$ be the network distribution, parameterised by $w$. A practical learning rule is usually a gradient descent rule which moves $w$ in the direction which reduces $E_\delta(q|z)$.

**Example 7.1** To optimise the 1-divergence is equivalent to

$$\min_q K(p,q) \iff \Delta w \sim \left\langle \partial_w \overset{0}{l}(q) \right\rangle_p - \left\langle \partial_w \overset{0}{l}(q) \right\rangle_q \tag{7.5}$$

This can be approximately implemented as a supervised learning rule, the Boltzmann machine learning rule [AHS85].

**Example 7.2** To minimise the 0-divergence is equivalent to

$$\min_q K(q,p) \iff \Delta w \sim \left\langle \partial_w \overset{0}{l}(q),\ \overset{0}{l}(p) - \overset{0}{l}(q) \right\rangle_q \tag{7.6}$$

This can be approximately implemented as a reinforcement learning rule, the simulated annealing reinforcement learning rule for stochastic networks[Zhu93]. [9]

The above should, in theory, be averaged over the posterior. However, since in neural network situations the prior is usually very week, the empirical distribution of $z$ can be used in place of $p$. The result will approach the correct result in the long run, which is to converge to a point where the generalisation error has a (local) minima.

Rather than simply regarding neural networks as non-flat models, some further structure may be used. Many neural networks, such as multilayer perceptrons, can be viewed as mixture models, for which a very powerful learning rule, the EM algorithm, exists [DL77]. The EM algorithm is recently re-analysed in light of information geometry [Ama94, Ama95].

If the model $\mathcal{Q}$ is further restricted such that neurons are independent of each other, then $\mathcal{Q}$ can be parameterised by the mean state of each neuron, resulting in a mean field model. This immediately gives the corresponding learning rules for feed-forward perceptrons and (continuous) Hopfield nets [Zhu93].

## 8 A Hyperprior is Unnecessary

In the proper Bayes framework, there is no difference between the prior and hyperprior. If the distribution of a variable in to change with new data, this variable is called parameter.

Let $z$ be the data which we can observe directly. Let $p$ be a parameter which we cannot observe but would like to infer from the observation of $z$. Then by definition, the prior $P(p)$ is the unconditional distribution of $p$ while the posterior is the conditional distribution.

The change from prior to posterior is effected by the Bayes rule through the likelihood function which is dependent on the data. Therefore any attempt to infer about the prior from the data is fundamentally wrong and breaks the coherence only enjoyed by Bayes methods. Technically speaking, coherent use of probability theory requires the assignment of joint distributions, and Bayes rule is a method which enables one to achieve this.

[9] Let $e(z)$ be an evaluation function specified by the environment. Define the environmental distribution by $\overset{0}{l}(p)(z) = \beta e(z)$, where $\beta$ is usually called the reciprocal temperature.

It is well known that due to Cox's Lemmas[Cox46], the only coherent method of real valued inference is through probability, which means, in particular, through the application of Bayes Theorem. All the other procedures will yield incoherent results sooner or later in a long chain of inference. Therefore, even if an approximation is valid at some stage from the point of view of certain applications, its validity is unlikely to hold if other consequences are derived from it. The incoherence will manifest itself, for example, if there is no learning rule which is optimal for almost all given data.

In the sense of pure mathematics, there is only one unique theory of probability, that defined by Kolmogorov's Axioms [Kol56], which specifies what kinds of operations with probability distributions are logically coherent, without any restriction on the possible interpretations. Although various statistical methodologies assign different meanings to probability, they can be divided into two categories: either it is possible to assign a joint distribution of all the objects under consideration, or it is not. If it is possible, then for any two objects, the probabilities of $a|b$, $b|a$, $a$ and $b$ cannot be given arbitrarily; they must be related by the Bayes' rule. If it is not possible to assign a joint distribution, then it is always possible to find a counterexample, in which the probability of an object, whatever its meaning, can be different depending on the way it is calculated, even allowing infinitely accurate calculations. Therefore we shall restrict our discussion to the Bayesian framework.

The objects which are called hyperparameters in the literature are actually a mixture of different things. The reason to treat hyperparameters differently than the ordinary parameters is not always clear, and is sometimes wrong. Here we shall provide an unambiguous definition of the term "hyperparameter". Then we shall show that if a parameter is qualified as such then it should always be integrated out before analysing the data.

Now suppose the parameter $p$ can be decomposed into two parts $[w, r]$ such that $P(z|p) = P(z|w)$, ie. the $r$ part is not contained in the likelihood function. Then we call $r$ a "hyperparameter". This appears to be the intended usages in [LS72] when the term was first introduced. Decomposing the prior into $P(p) = P(w, r) = P(w|r)P(r)$, the joint distribution of $z, w, r$ is given by

$$P(z, w, r) = P(z|w)P(w|r)P(r), \tag{8.1}$$

from which all the marginal and conditional distributions of combinations of these three variables are well defined. In particular, this is equivalent to the conditional independence of $z$ and $r$ given $w$, defined by any of the following three relations,

$$P(z, w|r) = P(z|w)P(w|r), \tag{8.2}$$

$$P(r, w|z) = P(r|w)P(w|z), \tag{8.3}$$

$$P(z, r|w) = P(z|w)P(r|w). \tag{8.4}$$

For a highly stimulating discussion of the concepts of conditional independence in statistics see [Daw79], where the notation $z{-}r|w$ was also introduced to denote any of the above relations.

Given data $z$, The $\delta$-optimal estimate $P_\delta(z'|z)$ for the distribution of new data $z'$ is given by

$$P_\delta(z'|z)^\delta = \int_{w,r} P(w,r|z)P(y|w)^\delta \tag{8.5}$$

$$= \int_w P(w|z) \int_r P(r|w)P(z'|w)^\delta = \int_w P(w|z)P(z'|w)^\delta \tag{8.6}$$

$$= \int_r P(r|z) \int_w P(w|r,z)P(z'|w)^\delta = \int_r P(r|z)P_\delta(z'|r,z)^\delta. \tag{8.7}$$

From (8.6), it is seen that the optimal estimate is determined by the prior distribution of $w$, obtained by marginalising out $r$. The hyperparameter $r$ only serves as a tool in the definition of prior $P(w)$ and is otherwise irrelevant.

It is usually the case that $P(w)$ is analytically more troublesome to deal with than $P(w|r)$. This might prompt one to do Bayes inference by conditioning on a particular value of $r$. However, from (8.7), it is clearly seen that such results should then be integrated over the posterior of the hyperparameter, $P(r|z)$, which is monstrously more difficult to deal with analytically. Furthermore, there is no hope that $P(r|z)$ would be more concentrated then $P(r)$, at least for the purpose of estimating $w$, due to the conditional independence of $r$ and $z$. If to the required precision a certain representative value of $r$ can be chosen from $P(r|z)$, then a similar one can be chosen from $P(r)$, so that $P(w|r)$ can be used as the prior of $w$. Therefore there is no circumstance in which it would be beneficial to keep a true hyperparameter in an inference process. This is also confirmed by numerical simulations of controlled experiments [Zhu95].

For those parameters which are usually called hyperparameters but nonetheless do appear in the likelihood function, one has no other choice but to give them a full Bayesian treatment, as if they are part of $w$. Indeed, any attempt to separate such parameters from the weight $w$ is artificial.

The above analysis completely resolves the controversy in [Mac92, Wol93, Mac93]. The conclusion is that one can and should obtain an estimate which is both coherent and invariant, by employing both the Bayes rule and the information divergence criteria. There is no conflict between treating all parameters in Bayesian framework and using an invariant criteria to select an optimal estimate.

## 9 Optimal Solutions to Classification Problems

In this section subscripts are used to represent classes instead of sample points. Such usage is different from the rest of the paper.

Suppose there are several classes $i \in I$, each with a distribution of data $p_i \in \mathcal{P}$. Suppose there is a prior $P(p_i)$ for each class $i$ and a data set $z_i$ is observed for class $i$. From these the posterior $P(p_i|z_i)$ and $\delta$-optimal estimate $\widehat{p}_i$ of class $i$ are well defined.

Suppose there is a new data set $z'$, with prior $P(p')$. The posterior of the distribution of observed data is also well defined, together with its $(1-\delta)$ optimal estimate.

Suppose we want to find the class $i$ whose true distribution $p_i$ is closest, in terms of $\delta$-divergence, to the true distribution $p'$ which generates the new data $z'$, averaged over the posterior. That is, we want to find the $\delta$-optimal classification, $i$, which minimises

$$(9.1) \qquad \langle D_\delta(p_i, p')\rangle_{z_1,\ldots,z_n,z'} = \langle D_\delta(p_i, \widehat{p}_i)\rangle_{z_i} + D_\delta(\widehat{p}_i, \widehat{p}') + \langle D_\delta(\widehat{p}', p')\rangle_{z'}.$$

It is obvious that the third term, which represents intrinsic uncertainty about the true distribution of the class which generates the new data, is independent of $i$. Therefore the $\delta$-optimal classification $i$ is obtained by minimising

$$(9.2) \qquad \langle D_\delta(p_i, \widehat{p}_i)\rangle_{z_i} + D_\delta(\widehat{p}_i, \widehat{p}').$$

**Classification Algorithm:**

1. Given data set $z_1, \ldots, z_n$, Find the $\delta$-optimal estimate $\widehat{p}_1, \ldots, \widehat{p}_n$. Compute $\langle D_\delta(p_i, \widehat{p}_i)\rangle_{z_i}$ for each class.
2. For any new data set $z'$, find the $(1-\delta)$-optimal estimate $p'$. Compute $D_\delta(\widehat{p}_i, \widehat{p}')$ for each class.
3. Assign class $i$ to $z'$ in order to minimise $\langle D_\delta(p_i, \widehat{p}_i)\rangle_{z_i} + D_\delta(\widehat{p}_i, \widehat{p}')$.

It is important to note that in order to obtain the $\delta$-optimal classification $i$, it is necessary to find the $(1-\delta)$-estimate of $p'$, instead of the $\delta$-optimal estimate. In most applications, the model space is exponential, and the class information is kept in 1-estimates $\widehat{p}_i$. Therefore for each new data set to be classified, a 0-optimal estimate is required. If a symmetrical formula is required, the $1/2$ estimations for $p_i$ and $p'$ are required. In any case, this shows clearly that $\delta$-optimal estimates for $\delta$ other than 1 are useful.

In practice, the model space $\mathcal{Q}$ is a subset of $\mathcal{P}$. It is in general impossible to express the optimal solution in $\mathcal{Q}$. Suppose we compute the $\delta$-projection of $\widehat{p}_i$ to $\mathcal{Q}$ as $\widehat{q}_i$, and the $(1-\delta)$ projection of $\widehat{p}'$ to $\mathcal{Q}$ as $\widehat{q}'$, then the following holds,

(9.3)
$$\langle D_\delta(p_i, p')\rangle_{z_1,\ldots,z_n,z'} \leq \langle D_\delta(p_i, \widehat{p}_i)\rangle_{z_i} + D_\delta(\widehat{p}_i, \widehat{q}_i) + D_\delta(\widehat{q}_i, \widehat{q}') + D_\delta(\widehat{q}', \widehat{p}') + \langle D_\delta(\widehat{p}', p')\rangle_{z'}.$$

The reason that the above is an inequality is that $D_\delta(\widehat{p}_i, \widehat{q}_i)$ and $D_\delta(\widehat{q}', \widehat{p}')$ may actually cancel each other. However, if the model is large enough so that both terms are small enough, it is sensible to ignore them altogether. This leads to the following algorithm.

**Restricted Classification Algorithm:**

1. Given data set $z_1, \ldots, z_n$, Find the $\delta$-optimal estimate $\widehat{p}_1, \ldots, \widehat{p}_n$. Find the $\delta$-projection $\widehat{q}_i$ of $\widehat{p}_i$ onto $\mathcal{Q}$. Compute $\langle D_\delta(p_i, \widehat{p}_i)\rangle_{z_i}$ for each class. Compute $D_\delta(\widehat{q}_i, \widehat{q}')$ for each class.

2. For any new data set $z'$, find the $(1-\delta)$-optimal estimate $p'$. Find the $(1-\delta)$ projection $\widehat{q}'$ of $\widehat{p}'$ onto $\mathcal{Q}$.

3. Assign class $i$ to $z'$ in order to minimise $\langle D_\delta(p_i, \widehat{p}_i)\rangle_{z_i} + D_\delta(\widehat{q}_i, \widehat{q}')$.

Note that this will not always result in the $\delta$-optimal classification. However, the procedure is well behaved when the model reasonably adequately represents the classes. This means that the projection error is small compared with the the "intra-class uncertainty" $\langle D_\delta(p_i, \widehat{p}_i)\rangle_{z_i}$ and the "inter class divergence" $D_\delta(\widehat{p}_i, \widehat{p}')$.

In applications the data set $z'$ often contains a single point, but this only simplifies the matter.

It might appear that the prescription given here for solving classification problems is non-Bayesian, since one might expect that a Bayesian treatment would assign a prior distribution to the classes which appears to be missing from the above treatment. This is not the case since the estimate $\widehat{p}'$ depends on the prior $P(p')$, which can be specified in an arbitrary fashion. If the prior knowledge is such that $z'$ is drawn from a distribution belonging to one of the classes, as is usually the case, then the prior of $p'$ is a mixture of the priors of $p_i$, and is therefore dependent on the mixing distribution. The general treatment here also allows the situations in which the class frequency in the test data is known to be different from that of the training data. This happens in practice if one tries to spend more computing resources training on more difficult classes, while the test set comes from applications with a different frequency.

## 10 Summary and Conclusions

The problem of finding a measurement of generalisation is solved in the framework of Bayesian inference, with machinery developed in the theory of information geometry.

By working in the Bayesian framework, this ensures that the measurement is internally coherent, in the sense that a learning rule is optimal if and only if it produces optimal estimates for almost all the data. By adopting an information geometric measure of divergence between distributions, this ensures that the theory is independent of parameterisation.

To guarantee a unique and well defined solution to the learning problem, it is necessary to generalise the concept of information divergence to the space of all the normalisable positive measures. This development reveals certain elegant relations between information geometry and the theory of Banach spaces, showing that the dually-affine geometry of statistical manifolds is in fact intricately related to the dual linear geometry of Banach spaces.

The framework and main results are summarised as the following:

- Neural networks are parameterised statistical models. Weights are coordinates. Learning rules are estimators.

- Bayes Rule gives a posterior over the distributions.
- Information geometry defines a divergence between two arbitrary distributions.
- Generalisation is measured by the posterior expected divergence between the true distribution and the estimated distribution.
- An estimator is optimal if and only if it almost always produces an optimal estimate.
- The ideal $\delta$-estimate is the $\delta$-average of the posterior.
- The $\delta$-estimate within a model $\mathcal{Q}$ is given by the $\delta$-projection of the ideal $\delta$-estimate onto $\mathcal{Q}$.
- The error of an estimate is the sum of the intrinsic uncertainty in the ideal estimate and the divergence between the ideal estimate and the actual estimate.
- Hyperparameters should be integrated out before statistical inference is performed.

The ideal optimal solution thus defined summaries, together with a statistic indicating the data size, all the information contained in the prior and data. The extension of information geometry to $\widetilde{\mathcal{P}}$ also offers a new perspective to the understanding of ancillary statistics.

This theory generalises linear Gaussian regression theory to general statistical estimation and function approximation problems.

The implication of this theory on the various methodologies for comparing neural network learning rules and other statistical inference methods is that a meaningful evaluation can and only can be achieved under three assumptions:

- A prior $P(p)$, describing the environment of all the problems.
- A divergence $D_\delta$, specifying the requirement of the task.
- A model $\mathcal{Q}$, specifying available computing resources.

Any evaluation not having these three assumptions either assumes them implicitly, or the evaluation itself is influenced by uncontrollable random chances, which would be similar to evaluating the behaviour of gambling based on whether one actually wins.

## A Relation with the Geometry of $L_{1/\delta}$

There are great conceptual simplifications when the $\delta$-geometry of $\widetilde{\mathcal{P}}(Z)$ is associated with the geometry of $L_{1/\delta}(Z)$, the space of $(1/\delta)$th power integrable functions on $Z$. Most material in this section is not presented in a rigorous manner. Some of it is even speculative.

It does not form an integral part of this report, but provides intuitive explanations which also serve as a motivation for future research.

The condition for $D_\delta(p,q)=0$ is easily seen from the following inequalities [HLP52],

$$\int p^\delta q^{1-\delta} \le \left(\int p\right)^\delta \left(\int q\right)^{1-\delta} \le \delta \int p + (1-\delta)\int q. \tag{A.1}$$

The first inequality becomes equality only when $p$ and $q$ are proportional to each other. The second inequality becomes equality only when $p$ and $q$ have the same mass.

In view of the $\delta$-representations, the $\delta$-divergence is a generalisation of squared distance from Hilbert spaces to a dual pair of Banach spaces,

$$D_\delta(p,q) = \delta\theta(p)\cdot\eta(p) + (1-\delta)\theta(q)\cdot\eta(q) - \theta(p)\cdot\eta(q). \tag{A.2}$$

For $\delta = 1/2$, the space $L_{1/\delta} = L_2$ is a self-dual Banach space, hence a Hilbert space. Correspondingly the connection $\overset{1/2}{\nabla}$ is a self-dual connection, hence the Levi-Civita connection. The 1/2-divergence is the Hellinger distance,

$$D_{1/2}(p,q) = \frac{1}{2}\|\theta(p)-\theta(q)\|_2^2 = \frac{1}{2}\|\eta(p)-\eta(q)\|_2^2 = 2\int\left(\sqrt{p}-\sqrt{q}\right)^2, \tag{A.3}$$

since $\theta(p)=\eta(p)$, and $\theta(q)=\eta(q)$.

For $\delta=1$, we obtain the cross entropy (Kullback-Leibler-divergence), given by

$$D_1(p,q) = \int\left(q - p + p\log\frac{p}{q}\right). \tag{A.4}$$

Similarly, $D_0$ is the reverse cross entropy.

The space $\widetilde{\mathcal{P}}$ can be identified with $L_{1/\delta}$ through the one-one correspondence $p \leftrightarrow \overset{\delta}{l}(p)$. This is possible only through the use of an (arbitrarily chosen) base measure $\mu$. A property of $\widetilde{\mathcal{P}}$ defined through the $\delta$-coordinate is called invariant if it remains the same when a different $\mu$ is chosen.

It is easily seen that $\overset{\delta}{l}(\mathcal{P})$ is the positive sector of a sphere centred at the origin, with radius $1/\delta$, in the sense of the Banach norm $\|\cdot\|_{1/\delta}$. The affine geometry of $\widetilde{\mathcal{P}}$ defined by the $\delta$-connection corresponds to the natural affine geometry of $L_{1/\delta}$ as a linear space, which is also induced by any inner product on $L_{1/\delta}$. [10] The space of probability distributions $\mathcal{P}$

[10] Although this is only possible for some subspace of $L_{1/\delta}$, we shall not worry too much about this point. In any case the discussion below will always be valid for $\mathcal{D}$, the space of infinitely smooth functions with compact support, which is dense in any function space [Aub79]. Therefore by a continuity argument most properties under discussion will hold for some substantial subspace of $L_{1/\delta}$ satisfying some general regularity conditions. From an application's point of view, we are not really concerned with geometry in infinite dimensional spaces. What we are really interested in is whether a given property is valid for all the finite dimensional subspaces. To define a property on an infinite dimensional space is a way to ensure that the definitions in overlapping subspaces are compatible with each other.

is a smooth surface in this inner product space. It is not necessarily a sector of a sphere, because the inner product is not given by the $(1/\delta)$-norm unless $\delta = 1/2$.

The inner product introduces an (infinite dimensional) Riemannian geometry on $\tilde{\mathcal{P}}$. It induces a Riemannian geometry on $\mathcal{P}$, which is not flat unless $\delta = 1$, since $\mathcal{P}$ is not an affine subspace of $\tilde{\mathcal{P}}$. This induced geometry depends both on the inner product $A$ chosen for the space $\tilde{\mathcal{P}}$, and the base measure $\mu$ used to define the $\delta$-coordinate $\overset{\delta}{l}$. However, all the different inner products on $\tilde{\mathcal{P}}$ thus defined induce the same affine geometry on $\tilde{\mathcal{P}}$, that defined by the $\delta$-connections.

If the canonical inner product on $\tilde{\mathcal{P}}$, defined by the norm

$$\left(\int \overset{\delta}{l}(p)^2\right)^{1/2} \tag{A.5}$$

is used, the induced geometry is called $\delta$-geometry. The $\delta$-metric is not invariant unless $\delta = 1/2$, but the $\delta$-affine connections are invariant. Therefore, the statement that $\delta$-connections are not metric [Ama85] should be interpreted as meaning that they are not induced by the only invariant metric, the $1/2$-metric.

It is interesting to know the expressions for $\delta$-geometry in the $\epsilon$-coordinate system, where $\epsilon \neq \delta$. This will clarify the above discussions. To avoid discussion of regularity conditions, we shall only consider the case where $Z$ is finite.

Let $\theta^i := \overset{\delta}{l}_i$, $\eta_i := \overset{1/\delta}{l}{}_i$, and $\xi^a := \overset{\epsilon}{l}_a$. Then, by definition,

$$\partial_a^i = p_a^{\delta-\epsilon}\mathbf{1}_a^i, \qquad \partial_{ab}{}^i = (\delta-\epsilon)p_a^{\delta-2\epsilon}\mathbf{1}_{ab}{}^i, \tag{A.6}$$

$$\overset{\delta}{g}_{ij} = \mathbf{1}_{ij}, \qquad \overset{\delta}{?}{}_{ijk} = 0. \tag{A.7}$$

Therefore

$$\overset{\delta}{g}_{ab} = \partial_a^i\partial_b^j\overset{\delta}{g}_{ij} = p_a^{2(\delta-\epsilon)}\mathbf{1}_{ab}, \tag{A.8}$$

$$\overset{\delta}{?}{}_{abc} = \partial_{ab}{}^i\overset{\delta}{g}_{ij}\partial_c^j = (\delta-\epsilon)p_a^{2\delta-3\epsilon}\mathbf{1}_{abc}, \tag{A.9}$$

$$\overset{\delta}{?}{}_{ab}{}^c = (\delta-\epsilon)p_a^{-\epsilon}\mathbf{1}_{ab}{}^c. \tag{A.10}$$

This specifies the $\delta$-geometry in the $\epsilon$-coordinates. The above formulas are of course compatible with the Eguchi relations

$$\overset{1/2}{g}{}_{ab} = \partial_a\theta^i\partial_b\eta_i = p_a^{1-2\epsilon}\mathbf{1}_{ab}, \tag{A.11}$$

$$\overset{1/2}{g}{}_{cd}\overset{\delta}{?}{}^d_{ab} = \partial_{ab}\theta^i\partial_c\eta_i = (\delta-\epsilon)p_a^{1-3\epsilon}\mathbf{1}_{abc}. \tag{A.12}$$

The Christoffel symbols for different $\delta$ are proportional to each other. This does not mean that they are induced by metrics which are proportional to each other, since Christoffel symbols are not tensors. [11] Indeed, the proportionality coefficients $\delta - \epsilon$ depends on the

[11] We would like to thank Prof. S. Amari for clarifying this potentially confusing point.

particular coordinate system through the presence of $\epsilon$.

In multinomial models, the volume differential element associated with $\overset{\delta}{g}$ is

$$\sqrt{\det[\overset{\delta}{g}_{ab}]} = \prod_i p_i^{\delta-\epsilon}. \tag{A.13}$$

This defines a density on $\widetilde{\mathcal{P}}$ which can be called the $\delta$-uniform density. In the coordinate system with $\epsilon = 1$, it is easily seen that this is exactly the Dirichlet prior with parameter $\delta\mathbf{1} = [\delta, \ldots, \delta]$,

$$\sqrt{\det[\overset{\delta}{g}_{ab}]} = p^{\delta\mathbf{1}-\mathbf{1}} \sim D(p|\delta\mathbf{1}). \tag{A.14}$$

These densities, especially for $\delta = 1/2$, are called "non-informative priors" [Jef61, DeG70, BT73, Ber85]. Although the $\delta$-metric is not invariant, the $\delta$-densities are proportional to each other in different coordinate systems. The $\delta$-uniformity is therefore invariant. This is of course very natural, considering that the inner product is invariant under the Lie group associated with an affine space. The non-informative priors were in fact first introduced by Jeffreys [Jef61] through invariance properties. See [Kas89] for more on the relation between non-informative priors and affine structure.

The existence of invariant non-informative priors removes a major reason for reluctance to use the Bayes methods [Fis30, Fis34, Fis36]. This $\delta$-uniform prior is well defined for any finite dimensional $\delta$-affine submanifold of $\widetilde{\mathcal{P}}$. However, for higher dimensional cases many non-Bayesian statistics corresponds to Bayesian statistics with improper priors which are not non-informative [DSZ73, Aka80]. We do not know whether the concept of Haar measure is applicable to a non-locally compact group.

One should not insist on choosing the 1/2-uniform prior as the only non-informative prior, since it is readily observed that the maximum likelihood estimate for the multinomial distribution is the 1-optimal estimate with a 0-uniform prior. It is commonly accepted that the maximum likelihood estimate is among the best estimate based on information from the data alone. It is very likely that most of the desirable properties of the maximum likelihood estimate also hold for any $\delta$-optimal estimate with $\epsilon$-uniform prior, for any $\delta, \epsilon \in [0, 1]$.

The relation between the dual affine geometry of statistical manifolds and the duality between Banach space pairs opens a fascinating front for mathematical research. One obvious question to ask is whether any dual affine statistical manifolds in the sense of Lauritzen[Lau87] are associated with Banach space pairs, and whether the converse is true. In the theory of function spaces[Ada75, Aub79, Tri83], the index $\delta$ can be interpreted as a coefficient (multiplied on dimensionality) of orders of smoothness for the functions in a function space. It is another interesting question whether this interpretation has anything to do with $\delta$-divergence.

Another interesting issue worthy of further exploration is the relation between the $\delta$-optimal estimate and the need for ancillary statistics. It seems that the ancillary is used to indicate the distance from the $\delta$-optimal estimate in $\widetilde{\mathcal{P}}$ and its image on $\mathcal{P}$.

**Acknowledgements** This work was partially supported by EPSRC grant GR/J17814.

We would like to thank people in the Neural Computing Research Group for interesting discussions. In particular, we would like to thank C. Williams, C. Bishop, C. Qazaz, D. Lowe, I. Nabney for valuable comments, interesting suggestions, and stimulating discussions.

## References

[ABNK$^+$87] S. Amari, O. E. Barndoff-Nieldon, R. E. Kass, S. L. Lauritzen, and C. R Rao, editors. *Differential Geometry in Statistical Inference*, volume 10 of *IMS Lecture Notes Monograph.* Inst. Math. Stat., Hayward, CA, 1987.

[Ada75] R. A. Adams. *Sobolev Spaces.* Academic Press, New York, San Fransisco, London, 1975.

[AHS85] D. H. Ackley, G. E. Hinton, and T. J. Sejnowski. A learning algorithm for Boltzmann machines. *Cog. Sci.*, 9:147–169, 1985.

[Aka80] H. Akaike. The interpretation of improper prior distributions as limits of data dependent proper prior distributions. *J. Roy. Stat. Soc., B*, 42(1):46–52, 1980.

[Ama82] S. Amari. Differential geometry of curved exponential families—curvature and information loss. *Ann. Stat.*, 10(2):357–385, 1982.

[Ama85] S. Amari. *Differential-Geometrical Methods in Statistics*, volume 28 of *Springer Lecture Notes in Statistics.* Springer-Verlag, New York, 1985.

[Ama87] S. Amari. Differential geometrical theory of statistics. In Amari et al. [ABNK$^+$87], chapter 2, pages 19–94.

[Ama94] S. Amari. Information geometry of the EM and em algorithms for neural networks. Technical Report METR94-4, Univ. Tokyo, 1994.

[Ama95] S. Amari. The EM algorithm and information geometry in neural network learning. *Neural Computation*, 7(1):13–18, 1995.

[Ami89] D. J. Amit. *Modeling Brain Function: The World of Attractor Neural networks.* Cambridge University Press, Cambridge, 1989.

[AMR83] R. Abraham, J. E. Marsden, and T. Ratiu. *Manifolds, Tensor Analysis, and Applications.* Addison-Wesley, London, 1983.

[Aub79] J. P. Aubin. *Applied Functional Analysis.* J. Wiley & Sons, New York, 1979.

[Ber85] J. O. Berger. *Statistical Decision Theory and Bayesian Analysis.* Springer-Verlag, New York, 1985.

[BNCR86] O. E. Barndorff-Nielsen, D. R. Cox, and N. Reid. The role of differential geometry in statistical theory. *Intern. Stat. Rev.*, 54(1):83–96, 1986.

[BSA83] A. G. Barto, R. S. Sutton, and C. W. Anderson. Neuron-like elements that can solve difficult learning control problems. *IEEE Trans. Sys. Man and Cyber.*, 13:834–846, 1983.

[BSW90] A. G. Barto, R. S. Sutton, and C. J. C. H. Watkins. Sequential decision problems and neural networks. In D. S. Touretzky, editor, *Advances in Neural Information Processing Systems*, volume 2. Morgan Kaufmann, San Mateo, CA, 1990.

[BT73] G. E. P. Box and G. C. Tiao. *Bayesian Inference in Statistical Analysis.* J. Wiley & Sons, New York, 1973.

[CH74] D. R. Cox and D. V. Hinkley. *Theoretical Statistics.* Chapman and Hall, New York, 1974.

[Cox46] R. T. Cox. Probability, frequency and reasonable expectations. *Amer. J. Phys.*, 14:1–26, 1946.

[Daw79] A. P. Dawid. Conditional independence in statistical theory (with discussion). *J. Roy. Stat. Soc., B*, 41(1):1–31, 1979.

[DeG70] M. H. DeGroot. *Optimal Statistical Decisiosns.* McGraw-Hill, New York, 1970.

[DL77] A. P. Dempster and D. B. Laird, N. M. annd Rudin. Maximum likelihood from incomplete data via the em algorithm. *J. Roy. Stat. Soc., B*, 39:1–38, 1977.

[DSZ73] A. P. Dawid, M. Stone, and J. V. Zidek. Marginalization paradoxes in Bayesian and structural inference (with discussion). *J. Roy. Stat. Soc., B*, 35:189–233, 1973.

[Egu83] S. Eguchi. Second order efficiency of minimum contrast estimators in a curved exponential family. *Ann. Statist.*, 11:793–803, 1983.

[Fer67] T. S. Fersuson. *Mathematical Statistics : A Decision Theoretic Approach.* Academic Press, New York, 1967.

[Fis25] R. A. Fisher. Theory of statistical estimation. *Proc. Camb. Phi. Soc.*, 122:700–725, 1925. Reprinted in [Fis50].

[Fis30] R. A. Fisher. Inverse probability. *Proc. Camb. Phi. Soc.*, 26:528–535, 1930. Reprinted in [Fis50].

[Fis34] R. A. Fisher. Two new properties of mathematical likelihood. *Proc. Roy. Soc., A*, 144:285–307, 1934. Reprinted in [Fis50].

[Fis36] R. A. Fisher. Uncertain inference. *Proc. Amer. Acad. Arts Sci.*, 71(4):245–258, 1936. Reprinted in [Fis50].

[Fis50] R. A. Fisher. *Contributions to Mathematical Statistics.* J. Wiley & Sons, New York, 1950.

[GBD92] S. Geman, E. Bienenstock, and R. Doursat. Neural networks and the bias/variance dilemma. *Neural Computation*, 4(1):1–58, 1992.

[GG84] S. Geman and D. Geman. Stochastic relaxation, Gibbs distribution, and the Beyessian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 6:721–741, 1984.

[HCG93] S. J. Hanson, J. D. Cowan, and C. Lee Giles, editors. *Advances in Neural Information Processing Systems*, volume 5, San Mateo, CA, 1993. Morgan Kaufmann.

[HLP52] G. H. Hardy, J. E. Littlewood, and G. Polya. *Inequalities.* Cambridge University Press, 2 edition, 1952.

[Hou82] P. Hougaard. Parameterization of non-linear models. *J. Roy. Stat. Soc., B*, 44:244–252, 1982.

[HS49] P. R. Halmos and L. J. Savage. Application of the Radon-Nikodym theorem to the theory of sufficient statistics. *Ann. Math. Statist.*, pages 225–241, 1949.

[Jef61] H. Jeffreys. *Theory of Probability.* Clarendon Press, Oxford, 1961. First edition in 1939.

[Kas84] R. E. Kass. Canonical parameterization and zero parameter effects curvature. *J. Roy. Stat. Soc., B*, 46:86–92, 1984.

[Kas89] R. E. Kass. The geometry of asymptotic inference (with discussion). *Statistical Science*, 4(3):188–234, 1989.

[KGV83] C. Kirkpatrick, D. Gelat, Jr., and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220:671–680, 1983.

[Kol56] A. N. Kolmogorov. *Foundations of the Theory of Probability.* Chelsea Publishing Co., New York, 1956. Translation of *Grundbegriffe der Wahrscheinlichkeitsrechnung, 1933*, with added bibliography, edited by N. Morrison.

[KS79] M. Kendall and A. Stuart. *The Advanced Theory of Statistics: Inference and Relationship*, volume 2. Griffin, London, 4 edition, 1979.

[Kul59] S. Kullback. *Information Theory and Statistics.* J. Wiley & Sons, New York, 1959.

[Lau87] S. L. Lauritzen. Statistical manifolds. In Amari et al. [ABNK$^+$87], chapter 4, pages 163–216.

[LS72] D. V. Lindley and A. F. M. Smith. Bayes estimation for the linear model. *J. Roy. Stat. Soc., B*, 34:1–41, 1972.

[Mac92] D. J. C. MacKay. *Bayesian Methods for Adaptive Models.* PhD thesis, California Institute of Technology, Pasadena, CA, 1992.

[Mac93] D. J. C. MacKay. Hyperparameters: Optimise, or integrate out? Technical report, Cambridge, 1993.

[MRR$^+$53] N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller. Equations of state calculations by fast computing machines. *J. Chem. Phys.*, 21:1087, 1953.

[Nea93] R. M. Neal. Bayesian learning via stochastic dynamics. In Hanson et al. [HCG93], pages 475–482.

[Sha48] C. E. Shannon. A mathematical theory of communication, I & II. *Bell Syst. Tech. J.*, 27:379–423, 623–656, 1948.

[Tri83] H. Triebel. *Theory of Function Spaces.* Monographs in Mathematics, 78. Birkhäuser Verlag, Basel, 1983.

[Whi89] H. White. Learning in artificial neural networks: A statistical perspective. *Neural Computation*, 1(4):425–464, 1989.

[Wol93] D. H. Wolpert. On the use of evidence in neural neworks. In Hanson et al. [HCG93], pages 539–546.

[Yor92] J. York. Use of the Gibbs samplers in expert systems. *Artif. Intell.*, 56:115–130, 1992. Addendum, 397-398.

[Zac71] S. Zacks. *The Theory of Statistical Inference.* Wiley Series in Probability and Mathematical Statistics. J. Wiley & Sons, New York, 1971.

[Zhu93] H. Zhu. *Neural Networks and Adaptive Computers: Theory and Methods of Stochastic Adaptive Computations.* PhD thesis, Dept. of Stat. & Comp. Math., Liverpool University, 1993. `ftp://archive.cis.ohio-state.edu/pub/neuroprose/Thesis/zhu.thesis.ps.Z`.

[Zhu95] H. Zhu. Can we get something out of nothing? Manuscript, 1995.

[ZR95a] H. Zhu and R. Rohwer. Bayesian invariant measurements of generalisation for continuous distributions. Technical Report NCRG/4352, Dept. Comp. Sci. & Appl. Math., Aston University, August 1995. ftp://cs.aston.ac.uk/neural/zhuh/continuous.ps.Z.

[ZR95b] H. Zhu and R. Rohwer. Bayesian invariant measurements of generalisation for discrete distributions. Technical Report NCRG/4351, Dept. Comp. Sci. & Appl. Math., Aston University, August 1995. ftp://cs.aston.ac.uk/neural/zhuh/discrete.ps.Z.

[ZR95c] H. Zhu and R. Rohwer. Measurements of generalisation based on information geometry. Presented at the 1st Mathematics of Neural Networks and Applications Conference (MANNA), Oxford, July 1995.