Bayesian Invariant Measurements of Generalisation for Discrete Distributions

Huaiyu Zhu and Richard Rohwer

NCRG/4351

Technical Report NCRG/4351

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

Bayesian Invariant Measurements of Generalisation for Discrete Distributions

Huaiyu Zhu Richard Rohwer
Department of Computer Science and Applied Mathematics
Aston University, Aston Triangle, Birmingham B4 7ET

August 31, 1995

Abstract

Neural network learning rules can be viewed as statistical estimators. They should be studied in Bayesian framework even if they are not Bayesian estimators. Generalisation should be measured by the divergence between the true distribution and the estimated distribution. Information divergences are invariant measurements of the divergence between two distributions.

The posterior average information divergence is used to measure the generalisation ability of a network. The optimal estimators for multinomial distributions with Dirichlet priors are studied in detail. This confirms that the definition is compatible with intuition. The results also show that many commonly used methods can be put under this unified framework, by assume special priors and special divergences.

Contents

1 Introduction **2**

2 Example: The Binomial Distribution with Beta Prior **4**

2.1 The statistical model 4

2.2 Bayesian methods 5

2.3 Beta distribution prior 5

2.4 Information divergence 6

2.5 Optimal D_1 estimator for binomial distribution 7

2.6 Optimal D_0 estimator for binomial distribution 9

3 Generalisation Measure for Discrete Distributions **10**

3.1 Kullback-Leibler distance and information divergence 10

3.2 Generalisation measure for estimators and estimates 11

3.3 Optimal estimators and estimates 12

4 Optimal Estimators for Multinomial Distribution **13**

4.1 Multinomial distribution with Dirichlet prior 13

4.2 δ-Optimal estimator for multinomial distribution 15

5 Conclusions and Discussions **15**

A Properties of Gamma and Beta Functions **16**

B Multivariate Beta Function and Dirichlet Distribution **18**

B.1 Multinomial coefficients 18

B.2 Multivariate Beta Function 18

B.3 Normalised multivariate Beta function 20

B.4 Partial increments of order δ 20

1 Introduction

Most NN learning rules can be considered as estimating an unknown probability distribution based on finite data taken from that distribution. Fundamental to any of such methods is a notion of optimal estimation: given two estimations obtained by two different methods from the same data, on what ground shall we evaluate their relative merits?

Let us discuss this in more details. A neural network can be viewed as a parameterised statistical model [Whi89], where the parameters, ie. weights w, are to be estimated from training data $z \in Z$. Each weight vector w will decide a unique member q from the family $\mathcal{Q}$ of probability distributions over Z representable under a given network architecture. A learning rule τ is therefore an adaptive method of point estimation of probability distributions from finite data.

In statistical terms, a neural network corresponds to a statistical manifold $\mathcal{Q}$. The weight vector $w \in W$ correspond to coordinates on $\mathcal{Q}$. A learning rule $\tau : Z \to \mathcal{Q}$ corresponds to an estimator. A trained model $q \in \mathcal{Q}$ is an estimate.

Obviously, there are in general infinitely many different parameters values w which could have given rise to the same set of training data z. It is well established that Bayesian

methods can be used to derive posterior distributions of parameters, $\Pr(w|z)$, given a prior on the parameters $\Pr(w)$ and the likelihood function of the data, $\Pr(z|w)$. The fundamental problem to be addressed here is to decide a unique set of parameters $\widehat{w}$ which represents optimal generalisation in a certain sense. From now on we shall talk about the estimate q instead of the weight w.

We stipulate that generalisation should be measured by the performance of the estimator on independent data drawn from identical distribution, averaged over the prior distribution of training problems.

It is obvious that generalisation should be tested on independent data, otherwise to repeat what is in the training data might be a good strategy to get good score. It is not widely recognised, but is as important, that it should be tested on data drawn from identical distribution. Otherwise a biased (in a broad sense) estimator might get a better score on average. Furthermore, the very idea of a learning rule requires that the performance should be measured on average over all the possible problems this rule is to be applied to.

The next question to be settled is the meaning of "good performance". It should in some way measure the "divergence" between the true but unknown distribution p and the model distribution q. This is well studied in information geometry (See [Ama85, Ama87] and references given therein). It is known that there is a family of "information divergences" which measures the difference between any two probability distributions. These divergences are unique in many important properties they enjoy, including invariance under reparameterisation and one-one transformations in the sample space. This gives a family of generalisation measures for neural network training problem or any statistical estimation problem where a point estimation is required from finite data.

In other words, we shall consider Bayesian decision theory with the information divergences as loss functions.

In this report, the first of three installments, we shall concentrate on finite sample spaces. Explicit formulas of the optimal estimates will be derived for multinomial distribution with Dirichlet priors; they have close relations with well established statistical estimators. They are sufficient statistics. The technical reason will be apparent when the space of all positive measures, not just that of probability measures, is considered [ZR95b]. The case for continuous distributions will be studied in [ZR95a].

Since the model space representable by a certain neural network class usually does not contain the intended probability distribution, there is a further problem of approximating the "optimal estimations" defined by any of the criteria considered here. The problem of "approximation" will be discussed elsewhere [ZR95b]. In other words, in this note we generally assume that the space of probability distributions contains all the possible probability distributions of the "world".

In §2, we derive the results for binomial distributions with Beta prior. This serves as an illustration of what our more general results will look like, and motivates the more involved mathematical derivations for the more general results.

In §3, some important results of Bayesian methods and information geometry are collected and recasted in a form convenient for our requirements. The generalisation measure based thereon is defined and its optimal estimate is derived in a general form.

The family of multinomial distributions with its natural conjugate prior, Dirichlet distribution, is analysed in §4.

Discussions and conclusions are in §5.

2 Example: The Binomial Distribution with Beta Prior

2.1 The statistical model

Consider an imaginary "coin-flipping machine", which has a lever on a scale labelled $[0, 1]$. For each position $p \in \mathcal{P} = [0, 1]$ of the lever, the machine will toss coins in a sequence $z = [z_k : \ k = 1, 2, \ldots]$ with identical independent distribution $\Pr\{z_k = 1\} = p$, where $z_k \in Z = \{0, 1\}$ is the result of the kth toss and "head/tail" is represented by "1/0".

Denote $Z^N(m) := \left\{z \in Z^N : \ |z^N| = m\right\}$, where we use $|A|$ to denote the number of elements in a finite set A. Then $|Z^N(m)| = C(m, n)$. Denote $m := |z| = \sum_i z_i$, $n := N - m$. The mathematical description of the above setup is a Bernoulli experiment with parameter $p \in \mathcal{P}$, data $z^N \in Z^N$, and likelihood function

$$\Pr(z^N|p) = p^m(1-p)^n.$$

The output count m is a sufficient statistics (when N is known) with binomial distribution

$$\Pr(m|p) = C(m, n)p^m(1-p)^n.$$

The learning task we shall consider is the following: Given a finite sample $z^N = [z_1, \ldots, z_N]$ generated from an unknown p, compute $q = \tau(z) \in \mathcal{Q} = [0, 1]$ which is a "good estimation" of p in some sense. The following terms will be used throughout this document.

$z^N \in Z^N$ ..... z^N is the "training data", or "sample". It is a random variable. Z is the sample space.

N ..... the "data size".

$p \in \mathcal{P}$ ..... p is the "state of world", the "true parameter". It is the (unknown) distribution of z. $\mathcal{P}$ is the "world" or "world model". It is the set of all the possible world states p.

$q \in \mathcal{Q}$ ..... q is the "estimate", or the "trained model". It is a distribution intended to be a good approximation of p in some sense. $\mathcal{Q}$ is the "model", the set of all the estimates q. It is usually a subset of $\mathcal{P}$.

$\tau : Z^N \to \mathcal{Q}$ ... τ is the "learning method", or "training method", or "estimator". It maps samples to estimates. The intention is that $q = \tau(z^N)$ will be a good approximate of p, which will be better as N tends to infinity.

2.2 Bayesian methods

It is obvious that there are infinitely many p which could have possibly produced z^N, whatever N is. To make a probabilistic statement about p, one must assume a prior $\Pr(p)$ which describes the distribution of the position of the lever before one sees any data. We shall not go into the argument as why prior is necessary for the evaluation of learning rules [Zhu95], but it is worth pointing out that many methods which appear to have assumed no prior in fact fit in this framework by assuming a particular prior. This includes the maximum likelihood method.

For a given prior $\Pr(p)$, the Bayesian formula for posterior $\Pr(p|z^N)$ is

$$\Pr(z^N) = \int_p \Pr(p)\Pr(z^N|p), \qquad \Pr(p|z^N) = \Pr(z^N|p)\Pr(p)/\Pr(z^N). \tag{2.1}$$

It is important to note that the posterior is a distribution of the unknown worlds, not an estimate. Statistical problems in general, and neural network training problems in particular, require a specific estimate q of the unknown world p. Obviously, there are infinitely many possible choices, so the question "which is the optimal" remains to be answered.

This question cannot be dismissed by insisting on giving $\Pr(p|z^N)$ as the final answer, since in that case we are faced with the new problem of representing $\Pr(p|z^N)$ explicitly, a task immensely more difficult than representing q directly. If, on the other hand, we were to represent $\Pr(p|z^N)$ approximately by some parameterised model, we are faced with the same types of question, albeit on a much more complicated level.

2.3 Beta distribution prior

Although Bayes formula will give a posterior for any prior, even when the prior is not a proper distribution, there are certain families of distributions which have nice mathematical and computational properties as priors for a given family of distributions. They are such that the posterior is also a member of the family. Among them the natural conjugate priors [RS68, DeG70, BT73, Ber85] are such that the prior can be conveniently as representing previous empirical knowledge. It is to be noted that not all knowledge is necessarily empirical knowledge. However, it is usually beneficial to examine the consequence of a statistical method on the natural conjugate prior first.

The natural conjugate prior of binomial distribution is the Beta distribution

$$\Pr(p) = \frac{p^{a-1}(1-p)^{b-1}}{B(a,b)}. \tag{2.2}$$

The likelihood function is

$$\Pr(z^N|p) = p^m(1-p)^n. \qquad \Pr(m|p) = C(m,n)p^m(1-p)^n. \tag{2.3}$$

The data distribution is

$$\Pr(z^N) = \frac{B(a+m,b+n)}{B(a,b)}. \qquad \Pr(m) = \frac{C(m,n)B(a+m,b+n)}{B(a,b)}. \tag{2.4}$$

The posterior is

$$\Pr(p|z^N) = \Pr(p|m) = \frac{p^{a+m-1}(1-p)^{b+n-1}}{B(a+m,b+n)}. \tag{2.5}$$

It is quite obvious that $[a,b]$ is sufficient statistics for the prior, $[m,n]$ is sufficient statistics for the likelihood, and $[a+m,b+n]$ is sufficient statistics for the posterior.

In the rest of this document we shall suppress explicit notations for the sample size N. This means that the notation z denote a sample of size N, instead of size 1, for example.

2.4 Information divergence

What is one going to do with a Bayes posterior $\Pr(p|z)$, which is a distribution of distributions? This question is usually not systematically studied in the majority of Bayes methods, except in decision theories where an externally imposed loss function is assumed. As we shall see later, most Bayes methods can be regarded as using a "representative distribution" q as the final answer. Several examples from recent neural networks literature will illustrate this point. The evidence method of D. MacKay [Mac92] uses an approximation. D. Wolpert [Wol93] takes the maximum posterior distribution. R. Neal [Nea93] uses Monte Carlo simulations which is equivalent of sampling from posterior marginal distribution.

Now we come to the second important theme of this paper: to find an invariant measure of "divergence" $D(p,q)$ between the two distributions p and q, and demanding q to be closest to p averaged over the posterior $\Pr(p|z)$. In this section we shall only consider the most commonly used "divergences", the Kullback-Leibler divergence (also called cross entropy). In later sections we shall consider all the invariant divergences, in the sense that it is invariant under parameterisation of both p and z.

The Kullback divergence between two distributions $p, q \in [0,1]$ is

$$K(p,q) = p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q}. \tag{2.6}$$

It is one instance of the family of α-divergences in information geometry [Che72, Ama85]. For technical reasons we find it more convenient to use $\delta = (1-\alpha)/2$, following [Hou82,

Kas84]. Denoting $p_1 = p, p_2 = 1 - p$, the δ-divergence is defined $\forall \delta \in [0,1]$:

$$D_\delta(p,q) := \frac{1}{\delta(1-\delta)} \left(1 - \sum_i p_i^\delta q_i^{1-\delta} \right), \qquad \forall \delta \in (0,1). \tag{2.7}$$

$$D_0(p,q) := \lim_{\delta \to 0} D_\delta(p,q) = K(q,p). \tag{2.8}$$

$$D_1(p,q) := \lim_{\delta \to 1} D_\delta(p,q) = K(p,q). \tag{2.9}$$

It is obvious that the δ-divergence is independent of the way the distribution is parameterised. It is often more convenient to use other parameterisations in computations, such as $\log(p/(1-p))$, for example. The invariance of the divergence means that the distribution which minimises the expected divergence from the true distribution conditional on the observations is independent of the parameterisation. However, there is a family of parameterisations which are quite convenient in subsequent developments. These are called δ-coordinates and are defined by

$$l^\delta(p) := p^\delta/\delta, \qquad l^0(p) = \log p. \tag{2.10}$$

The 1-coordinate is called mixture coordinate, while the 0-coordinate called the exponential coordinate [Ama85].

Given a sample z and prior $\Pr(p)$, the δ-(optimal) estimate is defined as the distribution q such that $D_\delta(p,q)$ is minimal on average over the posterior distribution $\Pr(p|z)$.

2.5 Optimal D_1 estimator for binomial distribution

We first consider the binomial case, with $\delta = 1$.

The learned model $q := \tau(z)$ is dependent upon the data. Suppose the world p is known, we want to find an optimal learning method to minimise the expected divergence

$$E_1(\tau|p) := \sum_z \Pr(z|p) D_1(p, \tau(z)). \tag{2.11}$$

There is obviously a unique solution to this problem, $\tau(z) = p$, with the absolute minimum $E_1(\tau|p) = 0$. The fact that the solution $q = \tau(z)$ is independent of the data z is not surprising, since we have assumed that p is known. This kind of learning rule is of no use since it is only good when it happens to hit upon the true state of the world, and if that happens there is nothing to be "learned".

Now in reality p is unknown with a distribution $\Pr(p)$. Therefore we seek to minimise the expected divergence for the whole learning rule,

$$E_1(\tau) := \int_p \Pr(p) E_1(\tau|p). \tag{2.12}$$

Such a τ is called the 1-(optimal) estimator. Using the Bayes theorem this can be rewritten as

$$E_1(\tau) = \sum_z \Pr(z) E_1(q|z), \tag{2.13}$$

where

$$E_1(q|z) := \int_p \Pr(p|z) D_1(p,q). \tag{2.14}$$

This shows that to minimise the prior expected divergence of the learning rule, $E_1(\tau)$, it is equivalent to minimise the posterior expected divergence of the estimate, $E_1(q|z)$, for each possible sample z, $\Pr(z) > 0$. Such a q is called the 1-(optimal) estimate based on data z. These (expected) divergences are differentiable with respect to q,

$$\partial_q D_1(p,q) = \frac{1-p}{1-q} - \frac{p}{q} = \frac{q-p}{(1-q)q}, \tag{2.15}$$

$$\partial_q E_1(q|z) = \int_p \Pr(p|z) \frac{q-p}{(1-q)q} = \frac{q - \langle p \rangle_z}{(1-q)q}, \tag{2.16}$$

where $\langle p \rangle_z$ is the expectation of p conditional on z, defined by

$$\langle p \rangle_z := \int_p \Pr(p|z) p. \tag{2.17}$$

Therefore the 1-estimate is given by $q = \langle p \rangle_z$. This also completely specifies the 1-optimal estimator up to a set of data with zero probability. As said earlier, q is in fact a distribution, instead of simply a real number. In more details, let z' be any test data, then the 1-estimate q is defined as a distribution

$$\Pr(z'|q) = \int_p \Pr(p|z) \Pr(z'|p) = \Pr(z'|z). \tag{2.18}$$

The right hand side is exactly the posterior marginal distribution, which is widely used in many Bayes methods, such as the Monte Carlo method [Nea93].

It is interesting to observe the relation between the 1-optimal estimate and the 1-coordinates: Let $q = \tau(z)$ be the estimate given by the learning rule τ. Then q is 1-optimal estimate if and only if $l^1(q)$ is the posterior expectation of $l^1(p)$ conditional on z. We shall see that a variant of this statement is true in general for any δ and any distribution family.

The formula for computing q is very simple for the Beta distribution prior $\Pr(p) \sim p^{a-1}(1-p)^{b-1}$. Denoting $C := a + b$, it is easy to show that

$$\langle p \rangle_z = \frac{a+m}{C+N}. \tag{2.19}$$

If one considers the prior as representing a "previous set of data" of size C with a 1's and b 0's, then the optimal estimate is simply the arithmetic mean of the "total data set". Since q is a sufficient statistic conditional on the "ancillary" $C+N$, the learning rule $\tau : z \to \tau(z)$ is adaptive in the sense that

$$q_+ = q + \frac{N}{C+N}\left(\frac{m}{N} - q\right). \tag{2.20}$$

2.6 Optimal D_0 estimator for binomial distribution

Now let us consider another important case, $\delta = 0$. The 0-divergence between two distributions $p, q \in [0,1]$ is

$$D_0(p,q) = K(q,p) = (1-q)\log\frac{1-q}{1-p} + q\log\frac{q}{p}. \tag{2.21}$$

Similar to the case of $\delta = 1$, we have the following corresponding definitions.

$$E_0(\tau|p) := \sum_z \Pr(z|p) D_0(p, \tau(z)). \tag{2.22}$$

$$E_0(q|z) := \int_p \Pr(p|z) D_0(p,q). \tag{2.23}$$

$$E_0(\tau) := \int_p \Pr(p) E_0(\tau|p) = \sum_z \Pr(z) E_0(q|z). \tag{2.24}$$

The concepts of 0-(optimal) estimates and estimators are similarly defined. The gradients can also be similarly derived, as

$$\partial_q D_0(p,q) = \log\frac{q}{p} - \log\frac{1-q}{1-p} = \log\frac{q}{1-q} - \log\frac{p}{1-p}, \tag{2.25}$$

$$\partial_q E_0(q|z) = \int_p \Pr(p|z)\left(\log\frac{q}{1-q} - \log\frac{p}{1-p}\right) = \log\frac{q}{1-q} - \left\langle \log\frac{p}{1-p}\right\rangle_z. \tag{2.26}$$

Therefore the 0-estimate q is given by

$$\frac{q}{1-q} = \exp\left\langle \log\frac{p}{1-p}\right\rangle_z. \tag{2.27}$$

Let $l^{1/\delta}$ denote the inverse of the mapping l^δ. The above is equivalent to

$$q \sim l^{1/0}\left\langle l^0(p)\right\rangle_z. \tag{2.28}$$

The corresponding result for $\delta = 1$ can also be expressed in the same form

$$q \sim l^{1/1}\left\langle l^1(p)\right\rangle_z. \tag{2.29}$$

This can be generalised to any δ in the following sections.

For the Beta distribution prior $\Pr(p) \sim p^{a-1}(1-p)^{b-1}$, we have

$$\left\langle \log\frac{p}{1-p}\right\rangle_z = \int_p \frac{p^{a+m-1}(1-p)^{b+n-1}}{B(a+m, b+n)} \log\frac{p}{1-p} = \Psi(a+m) - \Psi(b+n), \tag{2.30}$$

where Ψ is the digamma function. See Appendix A for definition and properties. It is known (See, eg., [Fer67, p. 180], which cites [JF45].) that for sufficiently large m,

$\exp \Psi(m) \approx m - 1/2$. Therefore $q = \tau(z) \approx (a + m - 1/2)/(C + N - 1)$, which is asymptotically equivalent to the 1-optimal estimate.

This also leads to an adaptive method. Let $\theta = \log(q/(1-q))$. Suppose that sufficient statistic a, b is kept. Then the method is described by

$$z = 1 \implies a_+ = a + 1,\ b_+ = b,\ \theta_+ = \theta + 1/a_+, \tag{2.31}$$

$$z = 0 \implies a_+ = a,\ b_+ = b + 1,\ \theta_+ = \theta - 1/b_+. \tag{2.32}$$

3 Generalisation Measure for Discrete Distributions

Instead of continuing with more special examples, we now turn to the task of defining generalisation measure for all discrete distributions, and find out the formula for corresponding optimal estimators.

Note that discrete distributions are characterised by the fact that they are dominated by a measure with a countable support. In other words, sum can be used in place of integration.

3.1 Kullback-Leibler distance and information divergence

Let Z be a finite sample space and $|Z| = n$. Then it can be identified with $\mathbb{N}_n$. The space $\mathcal{P}$ of distributions on Z can be identified with the standard $(n-1)$-simplex

$$\Delta^{n-1} := \left\{ p : \sum_i p_i = 1, p_i \geq 0 \right\} \subset \mathbb{R}^n. \tag{3.1}$$

Definition 3.1 (Information divergence) Let $p, q \in \Delta^{n-1}$. The Kullback-Leibler divergence

$$K(p,q) := \sum_i p_i \log \frac{p_i}{q_i}. \tag{3.2}$$

Let $\delta \in (0,1)$. The δ-divergence is defined by

$$D_\delta(p,q) := \frac{1}{\delta(1-\delta)} \left(1 - \sum_i p_i^\delta q_i^{1-\delta} \right), \tag{3.3}$$

$$D_0(p,q) := \lim_{\delta \to 0} D_\delta(p,q) = K(q,p), \tag{3.4}$$

$$D_1(p,q) := \lim_{\delta \to 1} D_\delta(p,q) = K(p,q). \tag{3.5}$$

It is easy to verify that that

$$D_{1/2}(p,q) = 2 \sum_i \left(\sqrt{p_i} - \sqrt{q_i} \right)^2. \tag{3.6}$$

The family of δ-divergences was discovered several times in both information theory and statistics. It is essentially the same as various information measures, including "Renyi's

information" and "Chernoff distance". In practice, $\delta \in \{0, 1/3, 1/2, 2/3, 1\}$ have distinct statistical interpretations [Hou82, Kas84]. Among many other names, D_1 and D_0 are called the cross entropy (Kullback-Leibler divergence) and the reversed cross entropy, respectively. $D_{1/2}$ is the Hellinger distance. See [AD75, p.208] and [Ama85, Ama87] for more backgrounds and references.

It is proved by [Che72] that in the case of finite sample space, the δ-connections are the only family of invariant connections. It was conjectured in [Ama85] that this is also true for general sample space. The δ-divergence are intrinsically related to the δ-connections, both through the δ-affine coordinates, the Generalised Pythagorean Theorem, and through the Eguchi relations [Egu83]. The relation between δ-divergence and the $L_{1/\delta}$ space shed some light on the duality between δ and $(1-\delta)$. This will be discussed elsewhere [ZR95b]

3.2 Generalisation measure for estimators and estimates

Let Z be a sample space. Consider the information manifold $\mathcal{P}$, the space of probability distributions. Suppose the world is represented by an unknown distribution $p \in \mathcal{P}$, with a prior $\Pr(p)$. Suppose in an experiment a sample z is observed, with likelihood $\Pr(z|p)$. The posterior distribution $\Pr(p|z)$ can be obtained from the Bayes rule. A learning method is a mapping $\tau : z \to q = \tau(z)$, which maps each observed set of data z to a unique distribution $q \in \mathcal{P}$.

Definition 3.2 (Generalisation error)

$$E_\delta(\tau|p) := \sum_z \Pr(z|p) D_\delta(p, \tau(z)). \qquad E_\delta(\tau) := \int_p \Pr(p) E_\delta(\tau|p). \tag{3.7}$$

$$E_\delta(q|z) := \int_p \Pr(p|z) D_\delta(p, q). \tag{3.8}$$

Note that both q and p are distributions in a common family, while τ itself is a mapping from a set of data to this family of distributions.

Corollary 3.1 *The performance of an estimator is the expected performance of estimates it gives.*

$$E_\delta(\tau) = \sum_z \Pr(z) E_\delta(\tau(z)|z). \tag{3.9}$$

These functionals measure the expected divergence of the learned distribution from the true distribution for a given estimator, which is invariant under reparameterisation.

Suppose it is known exactly that the world distribution is p, then using an estimator τ would result in an average δ-divergence of $E_\delta(\tau|p)$. So the optimal estimator is $\tau(z) = p$, independent of z. In general the true distribution p is unknown. Suppose we have a prior $\Pr(p)$ of possible worlds p. Then on average the generalisation error is $E_\delta(\tau)$.

Definition 3.3 (Optimal estimator) An estimator τ is called an δ-optimal estimator if it is the solution of the following optimisation problem

$$\operatorname*{Min}_{\tau(z)\in\mathcal{P}} E_\delta(\tau). \tag{3.10}$$

Definition 3.4 (Optimal estimate) A probability distribution p is called an δ-optimal estimator from data z if it is the solution of the following optimisation problem

$$\operatorname*{Min}_{q\in\mathcal{P}} E_\delta(q|z). \tag{3.11}$$

It is often argued that an optimal estimator should not be defined as one that gives the best average performance on all the possible data, but one that gives the the best performance on each data that actually occur [Lor90]. The relation between $E_\delta(\tau)$ and $E_\delta(p|z)$ shows that these two criteria are in fact equivalent.

Theorem 3.2 *An estimator is δ-optimal if and only if for any data, except a set of probability zero, the result of the estimator is a δ-optimal estimate.*

In other words, to minimise prior expected divergence is equivalent to minimise the posterior expected divergence for all possible sample. This means to find the optimal estimator we only need to find the estimate for each possible sample.

3.3 Optimal estimators and estimates

Suppose we have sample z and want to find out the δ-optimal estimate q, which can be regarded as a vector in $\mathbb{R}^n$ if we consider Δ^{n-1} as the standard simplex embedded in $\mathbb{R}^n$. Therefore the problem of finding the optimal q is a constrained minimisation problem, and can be solved by the Lagrange multiplier method. Define

$$F := E_\delta(q|z) - \lambda\left(\sum_{z'} q_{z'} - 1\right). \tag{3.12}$$

Intuitively, the data z is the one which actually observed, while z' is an arbitrary data whose probability the estimator must predict.

From equation (3.3), it can be derived that

$$\frac{\partial D_\delta(p,q)}{\partial q_{z'}} = -\frac{1}{\delta}\left(\frac{p_{z'}}{q_{z'}}\right)^\delta. \tag{3.13}$$

$$\frac{\partial E_\delta(q|z)}{\partial q_{z'}} = \int_p \Pr(p|z)\frac{\partial D_\delta(p,q)}{\partial q_{z'}} = -\frac{1}{\delta}\frac{\left\langle p^\delta_{z'}\right\rangle_z}{q^\delta_{z'}}. \tag{3.14}$$

Therefore, the δ-estimate q is given by

$$\frac{\partial F}{\partial q_{z'}} = -\frac{1}{\delta}\frac{\left\langle p^\delta_{z'}\right\rangle_z}{q^\delta_{z'}} - \lambda = 0 \iff q^\delta_{z'} \sim \left\langle p^\delta_{z'}\right\rangle_z.$$

The proportionality constant is the partition function. Translating back to distributions, q is the δ-estimate, which is itself a distributions, if and only if

$$\Pr(z'|q)^\delta \sim \int_p \Pr(p|z) \Pr(z'|p)^\delta. \tag{3.15}$$

For $\delta = 1$, this reduces to the posterior marginal distribution,

$$\Pr(z'|q) = \Pr(z'|z). \tag{3.16}$$

For $\delta = 0$, the result can be arrived at by taking limit,

$$\log \Pr(z'|q) = \langle \log \Pr(z'|p) \rangle_z - C, \tag{3.17}$$

where C is the logarithm of the partition function, which depends on z but is independent of z'.

Denote by $l^{1/\delta}$ the inverse mapping of l^δ. The above is summarised in the following theorem.

Theorem 3.3 (δ-optimal estimate of discrete distributions) *For discrete distributions, given observed data z, the δ-optimal estimate is given by*

$$q = \tau_\delta(z) \sim l^{1/\delta} \left(\left\langle l^\delta(p) \right\rangle_z \right) \tag{3.18}$$

The right hand side is called δ-average over the posterior $\Pr(p|z)$. [1] This means that the δ-optimal estimate can be obtained in three steps: use the δ-representation of p, average over the posterior, and renormalise.

4 Optimal Estimators for Multinomial Distribution

In this section we shall derive explicit formula for the δ-optimal estimators for the multinomial distribution.

4.1 Multinomial distribution with Dirichlet prior

Consider a multinomial distribution with n possible outcomes. Denote the total number of experiments as N. Let z_i^k be the ith component of the result of kth experiment. In each experiment exactly one event occurs, $|z^k| = 1$. This means $z \in Z^N$, where $Z := \mathbb{N}^n(1)$.

The number of event i occurs in N experiments is denoted $m_i := |z_i|$. So $m := [m_1, \ldots, m_n] \in \mathbb{N}^n(M)$, and $N = |m| = |z| = \sum_i |z_i| = \sum_k |z^k|$. We have, $\forall z \in (\mathbb{N}^n(1))^N$, $m \in \mathbb{N}^n(N)$.

The natural conjugate prior for the multinomial distribution is the Dirichlet distribution, which generalises the Beta distribution on the unit interval to the standard simplex in

[1] Also called weighted Hölder δ-means [HLP52]. This usage is essentially the same as [AD75].

any finite dimensional space [DeG70, Ber85, Car77]. Most of relevant properties are summarised in Appendix B.

Then the likelihood function is

$$\Pr(z|p) = p^m := \prod_i p_i^{m_i}, \qquad \Pr(m|p) = C(m)p^m. \tag{4.1}$$

The Dirichlet prior is $\forall a \in \mathbb{R}_+^n$

$$\Pr(p) = \frac{p^{a-1}}{B(a)} = D(p|a), \tag{4.2}$$

The joint distribution of data and parameter is

$$\Pr(z,p) = \frac{p^{a+m-1}}{B(a)} = p^m D(p|a), \qquad \Pr(m,p) = C(m)\frac{p^{a+m-1}}{B(a)}. \tag{4.3}$$

The prior marginal distribution of data is

$$\Pr(z) = \frac{B(a+m)}{B(a)}, \qquad \Pr(m) = C(m)\frac{B(a+m)}{B(a)}. \tag{4.4}$$

The posterior is

$$\Pr(p|z) = \Pr(p|m) = \frac{p^{a+m-1}}{B(a+m)} = D(p|a+m). \tag{4.5}$$

The posterior marginal data distribution is

$$\Pr(z'|z) = \frac{B(a+m+m')}{B(a+m)}. \tag{4.6}$$

Since m is a sufficient statistic for z, it is only necessary to derive formulas for m instead of z. It is easy to see that these formulas reduce to corresponding ones for binomial distribution when $n = 2$.

Corollary 4.1 *The data distribution is multivariate hypergeometric distribution(?)*

$$\Pr(m) = \frac{C(m, a-1)}{C(|m|, |a|-1)}. \tag{4.7}$$

4.2 δ-Optimal estimator for multinomial distribution

See Appendix B for the notations used in this section. It is straight forward to derive, from the general formula for δ-optimal estimator and the posterior of multinomial distribution that

Theorem 4.2 *Let $\delta \in (0,1]$. The δ-optimal estimate $q = \tau_\delta(z)$ for multinomial distribution $M(m|p)$ with Dirichlet prior $B(p|a)$, data z with statistic $m_i := |z_i|$, is given by*

$$(q_i)^\delta \sim L_i^\delta(a+m) = \frac{(a_i+m_i)_\delta}{(|a+m|)_\delta} \sim (a_i+m_i)_\delta. \tag{4.8}$$

In particular, for $\delta = 1$,

$$q_k = B(a+m+\delta_k)/B(m+a) = (m_k+a_k)/|m+a|. \tag{4.9}$$

This can be intuitively interpreted as the arithmetic mean of the "combined data set" composed of the observed data with sufficient statistics m and a set of "previous data" with sufficient statistics a.

For the case of $\delta = 0$, we have the following theorem.

Theorem 4.3 *The 0-optimal estimate $q = \tau_0(m)$ for multinomial distribution $M(m|p)$ with Dirichlet prior $B(p|a)$ is given by*

$$\log q_i + C = L_i^0(a+m) = \Psi(a_i+m_i) - \Psi(|a+m|), \tag{4.10}$$

where Ψ is the logarithmic derivative of? function (also called the digamma function), and C is a constant. Equivalently,

$$q_i \sim \exp \Psi(a_i+m_i). \tag{4.11}$$

These formulas specifies the δ-estimate for multinomial distributions with Dirichlet priors. They also specifies the δ-estimators uniquely up to a set of data with zero probability.

5 Conclusions and Discussions

We have combined the Bayesian decision theory with information geometry to provide a theory for the evaluation of statistical estimators, defining a measure of generalisation which enables selection from the Bayes posterior a unique representation which is optimal in the sense of information geometry. It is shown that the δ-optimal estimates are characterised by the fact that their δ-coordinates proportional to posterior expectation of the δ-coordinates of the true distributions.

It is coherent in the sense that the optimal estimator is characterised by the fact that it gives the optimal estimates for almost all the the data. It is invariant under transformations both in the sample space and in the parameter space.

We have argued that the result of statistical estimator should be a point estimate. Although the δ-optimal estimates are points in the posterior, no information is lost when they are used as representatives of the posterior, since each of them is a sufficient statistic. The 1-optimal estimate is the posterior marginal distribution, which is the distribution effectively used in the Monte Carlo simulate methods [Nea93].

The Dirichlet prior $\Pr(p|a\mathbf{1})$ is a-uniform distribution over Δ^{n-1}, ie. uniform distribution over a-coordinates, and can be regarded as non-informative priors. The 1-estimate with 0-uniform prior coincide with the maximum likelihood estimator, which act as a representative "data point" in Amari's theory of information geometry for exponential families.

As far as we are aware, this is the first attempt to combine the Bayesian framework and information geometry. Detailed formula for multinomial distributions provide first hand, intuitively accessible knowledge about the consequences of this theory. A more general approach is pursued in [ZR95b].

The major contribution of these explorations is to show that it is possible to define generalisation in a way which is both coherent and invariant, thereby overcoming a major obstacle to Bayesian methods of inference. This therefore act as a reference point for the comparison of all the learning methods.

A Properties of Gamma and Beta Functions

Most of the materials here are standard. The main purpose of this appendix is to fix notations.

The Gamma function is defined as

$$\text{(A.1)} \qquad ?\,(a) := \int_0^\infty e^{-t} t^{a-1} dt = \int_0^1 (-\log u)^{a-1} du.$$

The Beta function is defined as

$$\text{(A.2)} \qquad B(a,b) := \int_0^1 p^{a-1}(1-p)^{b-1} dp.$$

It has the well known Gamma representation

$$\text{(A.3)} \qquad B(a,b) = \frac{?\,(a)?\,(b)}{?\,(a+b)}.$$

The Psi function, also called the digamma function, is logarithmic derivative of Gamma function, defined as

$$\text{(A.4)} \qquad \Psi(a) := \Pr(\log ?\,(a)) = \partial ?\,(a)/?\,(a).$$

It has the following interesting representations

$$\Psi(x)+\gamma=\sum_{k=0}^{\infty}\left(\frac{1}{1+k}-\frac{1}{x+k}\right) \tag{A.5}$$
$$=\int_0^{\infty}\frac{e^{-t}-e^{-xt}}{1-e^{-t}}dt=\int_0^1\frac{1-u^{x-1}}{1-u}du,\qquad \forall x\in\mathbb{R}_+.$$

$$\Psi(x)-\Psi(y)=\sum_{k=0}^{\infty}\left(\frac{1}{z+k}-\frac{1}{x+k}\right) \tag{A.6}$$
$$=\int_0^1\frac{u^{z-1}-u^{x-1}}{1-u}du,\qquad \forall x,y\in\mathbb{R}_+.$$

$$\Psi(n)=-\gamma+\sum_{k=1}^{n-1}\frac{1}{k},\qquad \forall n\in\mathbb{N}_+. \tag{A.7}$$

$$\Psi(0)=\pm\infty,\qquad \Psi(1)=-\gamma,\qquad \gamma\approx 0.577215. \tag{A.8}$$

$$\exp\Psi(x)\approx x-1/2,\qquad \forall x\gg 1. \tag{A.9}$$

Notation. The Pochhammer symbol $(a)_b$ and the Appell symbol (a,b) are defined as

$$(a)_b:=(a,b):=?\,(a+b)/?\,(b). \tag{A.10}$$

We do not use the Appell symbol. In particular,

$$(a)_1=a,\qquad (a)_n=a(a+1)\cdots(a+n-1),\qquad (1)_n=n!. \tag{A.11}$$

Notation. We use the following notation for binomial coefficients

$$C(m,n):=C^m_{m+n}=\binom{n+m}{m} \tag{A.12}$$
$$=\frac{1}{(n+m+1)B(m+1,n+1)}=\frac{(m+1)_n}{(1)_n}.$$

Notation. The Beta distribution with parameter a,b is given by the pdf

$$D(p|a,b):=\frac{p^{a-1}(1-p)^{b-1}}{B(a,b)}. \tag{A.13}$$

Its maximum is located at $p=(b-1)/(a+b-2)$. Its δ-moment is

$$\int_p D(p|a,b)p^{\delta}=\frac{B(a+\delta,b)}{B(a,b)}=\frac{(a)_{\delta}}{(a+b)_{\delta}}. \tag{A.14}$$

Theorem A.1 *Let $a,b\in\mathbb{R}_+$.*

$$L(a,b):=\partial_a\log B(a,b)=\int_0^1 dpD(p|a,b)\log p=\Psi(a)-\Psi(a+b). \tag{A.15}$$

$$T(a,b):=\int_0^1 dpD(p|a,b)\log\frac{p}{1-p}=\Psi(a)-\Psi(b). \tag{A.16}$$

Proof: From the integral representation of Beta function we obtain the integral representation of $L(a,b)$, and from the Gamma representation of the Beta function we obtain the Psi representation of $L(a,b)$. This proves the identity for $L(a,b)$. The identity for $T(a,b)$ is obtained by noticing that $T(a,b) = L(a,b) - L(b,a)$. □

B Multivariate Beta Function and Dirichlet Distribution

Many of the results presented here are standard [DeG70, Car77], but we shall present them in a set of concise notation.

B.1 Multinomial coefficients

Notation. Let $a, b \in \mathbb{R}^n_+$, $p \in \Delta^{n-1}$, $m \in \mathbb{N}^n$:

$$p^a := \prod_i p_i^{a_i}, \qquad m! := \prod_i m_i!, \tag{B.1}$$

$$(a)_b := \prod_i (a_i)_{b_i}, \qquad ?\,(a) := \prod_i ?\,(a_i). \tag{B.2}$$

Definition B.1 The multinomial coefficients are defined $\forall m, k \in \mathbb{N}^n$:

$$C(m) := \frac{|m|!}{m!}. \qquad C(m,k) := \prod_i C(m_i, k_i). \tag{B.3}$$

Notation. Let $n, M \in \mathbb{N}$. Then

$$\mathbb{N}^n(M) := \{m \in \mathbb{N}^n : \ |m| = M\}\,; \tag{B.4}$$

$$\mathbb{N}_M := \{m \in \mathbb{N} : \ m \le M\}\,; \tag{B.5}$$

$$\mathbb{N}^n_M := \mathbb{N}_M \times \cdots \times \mathbb{N}_M = \{m \in \mathbb{N}^n : \ m_i \le M\}\,; \tag{B.6}$$

Lemma B.1 $|\mathbb{N}^n(N)| = C(n-1, N)$.

Theorem B.2 (Multinomial expansion) *Let $x \in \mathbb{R}^n$, $N \in \mathbb{N}$. Then*

$$\left(\sum_i x_i\right)^N = \sum_{m \in \mathbb{N}^n(N)} C(m) x^m. \tag{B.7}$$

B.2 Multivariate Beta Function

Definition B.2 The multivariate Beta function is defined $\forall a \in \mathbb{R}^n_+$:

$$B(a) := \int_{\Delta^{n-1}} dp\; p^{a-1}, \tag{B.8}$$

where $dp := dp_1 \cdots dp_{n-1}$, with $p_n = 1 - (p_1 + \cdots + p_{n-1})$.

Definition B.3 (Dirichlet distribution) Let $a \in \mathbb{R}_+^n$. Then $\forall p \in \Delta^{n-1}$:

$$D(p|a) := \frac{p^{a-1}}{B(a)}. \tag{B.9}$$

Theorem B.3 (Symmetry) *The multivariate Beta function is symmetric with regard to its arguments.*

Theorem B.4 (Recursive formula) *Let $i \in \mathbb{N}_n$. Denote $I := \mathbb{N}_n \setminus i$. Then $\forall a \in \mathbb{R}_+^n$:*

$$B(a) = B(a_I)B(a_i, |a_I|). \tag{B.10}$$

Proof: Denote $q := p_I/|p_I| \in \Delta^{n-2}$, $b := a_I \in \mathbb{R}_+^{n-1}$. By definition of the Beta function and multivariate Beta function, we have

$$\begin{aligned} B(a) &= \int_{\Delta^{n-1}} p^{a-1} dp \\ &= \int_0^1 dp_i \, p_i^{a_i-1} \int_{(1-p_i)\Delta^{n-2}} dq \, q^{b-1} \\ &= \int_0^1 dp_i \, p_i^{a_i-1} (1-p_i)^{|b|-1} \int_{\Delta^{n-2}} dq \, q^{b-1} \\ &= B(a_i, |b|)B(b). \end{aligned}$$

In the above, the multiple integration is substituted by iterated integration. □

Theorem B.5 (Gamma representation)

$$B(a) = \frac{?(a)}{?(|a|)}. \tag{B.11}$$

Proof: Denoting $b = a_{\mathbb{N}_{n-1}}$, it follows the recursive formula that

$$B(a) = B(b)B(a_n, |b|) = \frac{?(b)}{?(|b|)} \frac{?(a_n)?(|b|)}{?(|a|)} = \frac{?(a)}{?(|a|)}.$$

□

Corollary B.6 *Let $\mathbf{1} := [1, \ldots, 1]^T \in \mathbb{R}^n$. Then*

$$B(\mathbf{1}) = \int_{\Delta^{n-1}} dp = \frac{1}{?(n)}. \tag{B.12}$$

B.3 Normalised multivariate Beta function

The Beta function is normalised in a certain sense since the measure of the unit interval is unity. This is not so for the multivariate Beta functions when $n > 2$. It is sometimes convenient to use a normalised version of Beta function, defined as $B'(a) := B(a)/B(\mathbf{1})$.

Theorem B.7 *The following holds* $\forall a \in \mathbb{R}_+^n, m \in \mathbb{N}^n$*:*

(B.13)
$$B(a+1)(|a|+1)_{n-1}C(a) = 1. \qquad B'(m+1)C(m)C(|m|, n-1) = 1.$$

For $n = 2$ *the these reduce to* $B(a+1) = B'(a+1) = 1/(|a|+1)C(a)$.

Theorem B.8 *The following holds:*

(B.14)
$$\sum_{m \in \mathbb{N}^n(N)} C(m)B(m+1) = \frac{1}{?\,(n)}. \qquad \sum_{m \in \mathbb{N}^n(N)} C(m)B'(m+1) = 1.$$

Proof: Two proofs are available. (1) As a corollary of the above thereon and $|\mathbb{N}^n(N)| = C(n-1, N)$. 2) Use multinomial expansion theorem on the definition of $B(m+1)$. □

B.4 Partial increments of order δ

Theorem B.9 *Let* $a, b \in \mathbb{R}_+^n$*. Then*

(B.15)
$$L_b(a) := \int_{\Delta^{n-1}} dp\, D(p|a)p^b = \frac{B(a+b)}{B(a)} = \frac{(a)_b}{(|a|)_{|b|}}.$$

Notation. Denote by $\mathbf{1}_i \in \mathbb{R}^n$ the ith unit vector.

Corollary B.10 *Let* $i \in \mathbb{N}_n$*,* $a \in \mathbb{R}_+^n$*. Let* $\delta \in (0, 1]$*. Then*

(B.16)
$$L_i^\delta(a) := \int_{\Delta^{n-1}} dp\, D(p|a)p_i^\delta = \frac{B(a+\delta\mathbf{1}_i)}{B(a)} = \frac{(a_i)_\delta}{(|a|)_\delta}.$$

Theorem B.11 *The following is true* $\forall i \in \mathbb{N}_n$*,* $a \in \mathbb{R}_+^n$

(B.17)
$$L_i^0(a) := \int_{\Delta^{n-1}} dp\, D(p|a) \log p_i = \Psi(a_i) - \Psi(|a|) = -\sum_{k=a_i}^{|a|-1} \frac{1}{k}.$$

Proof: Consider $\partial_i \log B(a)$. It equals the integral representation of $L_i^0(a)$ following the definition of $B(a)$, while equals the Psi function representation following the Gamma representation of $B(a)$. □

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 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.

[AD75] J. Aczél and Z. Daróczy. *On measures of information and their characteristics*. Academic Press, New York, 1975.

[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.

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

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

[Car77] C. Carlson, B. *Special Functions of Applied Mathematics*. Academic Press, New York, 1977.

[Che72] N. N. Chentsov. *Optimal Decision Rules and Optimal Inference*. Nauka, Moscow, 1972. In Russian. English translation AMS: Rhode Island, 1982.

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

[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.

[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.

[JF45] E. Janhke and Emde F. *Tables of Functions with Formulae and Curves*. Dover Publications, New York, 4 edition, 1945.

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

[Lor90] T. J. Loredo. From Laplace to supernova SN 1987A: Bayesian inference in astrophysics. In P. F. Fougère, editor, *Maximum Entropy and Bayesian Methods*, pages 81–142. Kluwer Academic Publishers, 1990.

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

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

[RS68] H. Raiffa and R. Schlaifer. *Applied Statistical Decision Theory*. MIT Press, Cambridge, Mass., 1968.

[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.

[Zhu95] H. Zhu. Why is prior necessary for evaluating learning rules? 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. Information geometric measurements of generalisation. Technical Report NCRG/4350, Dept. Comp. Sci. & Appl. Math., Aston University, August 1995. `ftp://cs.aston.ac.uk/neural/zhuh/generalisation.ps.Z`.