To appear in Neural Processing Letters

Bayesian Invariant Measurements of Generalisation

Huaiyu Zhu and Richard Rohwer
Neural Computing Research Group
Dept of Computer Science and Applied Mathematics
Aston University, Birmingham B4 7ET, UK
Email: zhuh@aston.ac.uk, Fax: +44 121 333 6215

November 17, 1995

Abstract

The problem of evaluating different learning rules and other statistical estimators is analysed. A new general theory of statistical inference is developed by combining Bayesian decision theory with information geometry. It is coherent and invariant. For each sample a unique ideal estimate exists and is given by an average over the posterior. An optimal estimate within a model is given by a projection of the ideal estimate. The ideal estimate is a sufficient statistic of the posterior, so practical learning rules are functions of the ideal estimator. If the sole purpose of learning is to extract information from the data, the learning rule must also approximate the ideal estimator. This framework is applicable to both Bayesian and non-Bayesian methods, with arbitrary statistical models, and to supervised, unsupervised and reinforcement learning schemes.

Keywords: Information geometry, Bayesian decision theory, optimal learning rule, model effect, neural networks.

1 Introduction

We are concerned with the problem of evaluating and comparing neural network learning rules among themselves and with other statistical methods. To this end we shall outline a new theory of statistical inference by combining Bayesian decision theory with information geometry. We shall first present the main definitions and results in a mathematical form. Then we shall use a simple example to illustrate the meaning of these results. Finally we shall explain briefly why this framework is suitable for statistical inference in general, and neural network learning rules in particular, and discuss its implications.

2 Bayesian Decision Theory

Consider a sample space Z, the points of which are the states of the "visible neurons", for example. [1] To avoid measure-theoretic concerns, one may conveniently consider Z to be finite; the more general case requires more dedicated mathematics than space here allows. Denote by $\mathcal{P}$ the space of probability distributions on Z. It forms a differential manifold [1, 2], which may be infinite-dimensional if Z is infinite. Consider a statistical model on Z, such as a neural network, parameterised by $w \in W$. Each particular value of "weight" w corresponds to a distribution q, the totality of which forms a submanifold $\mathcal{Q} \subseteq \mathcal{P}$. One can think of w as coordinates of q.

A statistical estimator (learning rule) is a mapping $\tau : Z \to Q$. A learning problem is a distribution $p =: P(\cdot|p) \in \mathcal{P}$. The purpose of an estimator is to provide, for each training set z sampled from p, an estimate $q = \tau(z)$ which approximates p. This whole framework is schematically illustrated in Figure 1.

The normal practice of evaluating the performance of τ, such as by cross validation, is to compare $q = \tau(z)$ with some "test data" $z' \in Z$ generated from the same distribution p but independently of z. It is reasonable to hope that such evaluations get more accurate as the size of z' increases, and that at the infinite size limit one ends up evaluating $D(p, q)$, a measure of "divergence" between the true distribution p and the estimated distribution q. The "optimal learning rule" in this setting is obviously $\tau(z) \equiv p$, which learns nothing from the data z. This is also true if the ensemble of pos-

[1] We follow the standard practice of also using $z \in Z$ to refer to an IID sample of these states, without explicit reference to the sample size.

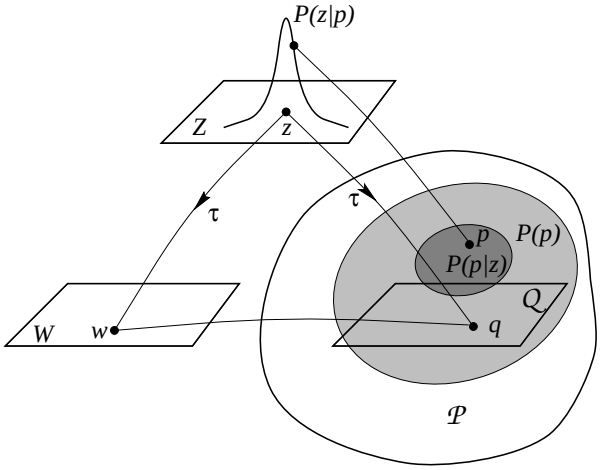

Figure 1: Relation between weight space, learning rule, and Bayes Theorem

sible data sets is considered, since in that case one is minimising

$$E(\tau|p) := \int_z P(z|p)D(p,\tau(z)). \tag{1}$$

Therefore it is clear that such procedures, although quite valid for evaluating the estimate q, fail completely to evaluate the estimator τ. The reason for the failure is that only one fixed problem p is considered, for which there is no way to distinguish "hard working" learning rules which learn from the data from "genius" learning rules which happen to have guessed correctly. Although in practice one does not normally know p, by evaluating learning rules in this way one is always (often unconsciously) requiring a good learning rule to ignore the data.

The only proper way to evaluate the "learning ability" of an estimator is to evaluate its average performance over a distribution of problems, the prior $P(p)$, by

$$E(\tau) := \int_p P(p)E(\tau|p), \tag{2}$$

so that any good estimator has to extract some information from the training data. An estimator τ is said to be optimal if it minimises $E(\tau)$.

Given a prior $P(p)$, it is also possible to obtain the posterior $P(p|z)$ by way of the Bayes rule $P(p|z) = P(z|p)P(p)/P(z)$. One may also demand that a learning rule should give the best estimate for each individual data set based on that knowledge alone; ie., to minimise

$$E(q|z) := \int_p P(p|z)D(p,q), \tag{3}$$

for each z. An estimate q is said to be optimal based on data z if it minimises $E(q|z)$.

All this leads naturally into the domain of Bayesian decision theory [3, 4, 5], where $D(p,q)$ acts as the "loss function"; one of its most important contribution is the following well-known theorem.

Theorem 2.1 (Coherence) *An estimator τ is optimal if and only if for any data $z \in Z$, excluding a subset of zero probability, $\tau(z)$ is an optimal estimate based on z.*

3 Information Divergence

In the general framework of decision theory the "loss function" $D(p,q)$ is externally supplied. One may ask what is the appropriate "loss function" if the sole utility is to extract information from the data. An answer to this question comes from information geometry [1, 6, 7], which defines a family of related geometries and related divergencies indexed by $\alpha \in [-1,1]$ [1]. For technical convenience we shall use $\delta = (1-\alpha)/2 \in [0,1]$ instead of Amari's α. We shall also consider the space

$$\widetilde{\mathcal{P}} := \left\{ p \geq 0 : \int p < \infty \right\} \tag{4}$$

of normalisable positive measures on Z, in which $\mathcal{P}$ forms a submanifold defined by $\int p = 1$. The most natural "information divergence" between $p, q \in \widetilde{\mathcal{P}}$ is defined by [1, 8]

$$D_\delta(p,q) = \frac{1}{\delta(1-\delta)} \int \left(\delta p + (1-\delta)q - p^\delta q^{1-\delta} \right). \tag{5}$$

One of the most important criteria of information divergence is invariance. Well-known special cases include the (extended) cross entropy (KL-distance) [9, 8]

$$(6)\quad D_1(p,q) = K(p,q) = \int \left(q - p + p \log \frac{p}{q} \right),$$

the reversed cross entropy $D_0(p,q) = K(q,p)$, and the Hellinger distance [1]

$$(7)\qquad D_{1/2}(p,q) = 2 \int \left(\sqrt{p} - \sqrt{q} \right)^2 .$$

When p and q are close to each other, it is approximately the χ^2 distance [7, 8]

$$(8)\qquad D_\delta(p, p+\Delta p) \approx \frac{1}{2} \int \frac{\Delta p^2}{p} \approx \frac{1}{2} \int p \left(\Delta \log p \right)^2 .$$

The quadratic form of the χ^2 distance can be represented by the Riemannian metric given by the Fisher information matrix [10, 1, 7], although for infinite Z these are all infinite dimensional objects. Using δ-divergence as the loss function in a Bayesian decision framework leads to our main result [8]:

Theorem 3.1 (Ideal estimate) *Given a prior $P(p)$ over $\mathcal{P}$, let $z \in Z$. Then, with expectation conditional on z denoted as $\langle \cdot \rangle_z$,*

$$(9)\qquad \left\langle D_\delta(p,q) \right\rangle_z = \left\langle D_\delta(p,\widehat{p}) \right\rangle_z + D_\delta(\widehat{p},q),$$

where $\widehat{p}$ is called the δ-estimate given by the δ-average over the posterior

$$(10)\qquad \begin{cases} \widehat{p}^\delta = \left\langle p^\delta \right\rangle_z , & \delta \in (0,1], \\ \log \widehat{p} = \left\langle \log p \right\rangle_z , & \delta = 0. \end{cases}$$

This reduces to the well-known "$MSE = VAR + BIAS^2$" formula for linear Gaussian models with $\delta \in \{0,1\}$, but the general form is applicable to any statistical model, which may be non-linear and non-Gaussian. In general the δ-estimate for $\delta \in [0,1)$ is not necessarily a probability distribution since it may not be normalised, but it will be shown that the optimal estimator within a probability model can be obtained simply by renormalisation.

It can also be shown that with natural conjugate priors [3], $\widehat{p}$ is a sufficient statistic for the commonly used statistical models, such as multinomials [11], Gaussians [12], and uniform distributions. We conjecture that this is always true, even for non-exponential families.

4 Optimal Estimators

For exponential families the δ-estimate can be represented simply by the sufficient statistics. For more general problems, however, we might have to be content with representing the estimate within a restricted model. As the δ-estimate $\widehat{p}$ is a member of $\widetilde{\mathcal{P}}$, all the inference problems within a model $\mathcal{Q}$ can be achieved by projecting $\widehat{p}$ onto $\mathcal{Q}$. Therefore the whole armoury of information geometry can be brought into force, which studies the asymptotic behaviour, the model deficiency, the curvature of the model and the estimator, etc. In particular, the following generalised Pythagorean theorem holds [1].

Theorem 4.1 (Optimal estimate) *Suppose that $\mathcal{Q}$ is a $(1-\delta)$-convex submanifold of $\widetilde{\mathcal{P}}$. Then*

$$(11)\qquad D_\delta(\widehat{p},q) = D_\delta(\widehat{p},\widehat{q}) + D_\delta(\widehat{q},q),$$

where $\widehat{q}$ is the δ-projection of $\widehat{p}$ onto $\mathcal{Q}$.

Theorem 3.1–4.1 are illustrated in Figure 2.

Theorem 4.2 *The optimal estimator within a probability model, ie. with $\int q = 1$, is given by renormalising the ideal estimator. That is, if $\mathcal{Q} = \mathcal{P}$, then*

$$(12)\qquad \widehat{q} = \widehat{p} \Big/ \int \widehat{p}.$$

Example 4.1 Consider the multinomial family of distributions $M(m|p)$ with a Dirichlet prior $D(p|a)$, where $m \in \mathbb{N}^n$, $p \in \Delta^{n-1} := \left\{ p \in \mathbb{R}^n_+ : \sum_i p_i = 1 \right\}$, $a \in \mathbb{R}^n_+$. The posterior is also a Dirichlet distribution $D(p|a+m)$. The δ-estimate is given by

$$(13)\qquad \widehat{p}_i^\delta = (a_i + m_i)_\delta / (|a+m|)_\delta,$$

where $|a| := \sum_i a_i$, $(a)_b := \Gamma(a+b)/\Gamma(a)$, and $\Gamma(a) := \prod_i \Gamma(a_i)$. In particular,

$$(14)\qquad \widehat{p}_i = (m_i + a_i)/|m+a|, \quad \delta = 1,$$

$$(15)\qquad \widehat{p}_i = \exp(\Psi(a_i + m_i) - \Psi(|a+m|)), \quad \delta = 0,$$

where Ψ is the the digamma function, the logarithmic derivative of Γ function. One can define the δ-uniform prior by way of the δ-affine connections [1, 8], and it can be shown that in this case it is $D(p|\delta \mathbf{1})$. Therefore the maximum likelihood estimator $\widehat{p}_i = m_i/|m|$ is the 1-estimator with 0-uniform prior.

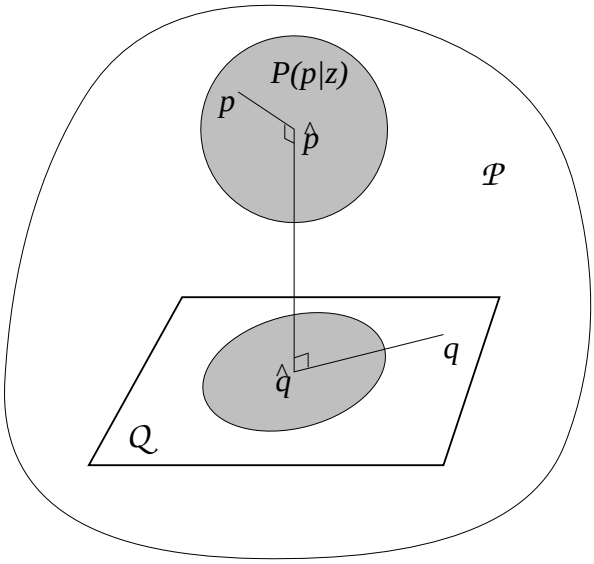

Figure 2: Decomposition of errors

5 Discussion and Summary

We outline here how various learning rules fit in this general framework.

A neural network (either deterministic or stochastic) can be regarded as a parameterised model $P(y|x,w)$ where x is input, y is output and w is weight [13]. With input distribution $P(x)$, it is also equivalent to $P(z|w)$, where $z := [x,y] \in Z := X \times Y$.

Bayesian methods which give the whole posterior $P(p|z)$ as an answer still fit in this framework, for either the posterior is analytically represented or it is simulated by a random process. In the former case one is faced with the same question as studied here. The latter case is equivalent to giving the posterior marginal distribution $P(z'|z)$ as the answer, which is exactly the 1-estimate $P(z'|\tau_1(z))$.

Many non-Bayesian estimators, such as the maximum likelihood estimator, can be regarded as generalised Bayesian estimators with improper priors [4, 5]. For multidimensional models there may be many different improper priors, not all of which are non-informative, giving rise to a variety of different non-Bayesian methods [14, 15].

The general conclusion from this study is that learning rules should be compared with each other only if the following three things are specified: (1) the prior $P(p)$, (2) the divergence $D(p,q)$, and (3) the model $\mathcal{Q}$. If the sole purpose of learning is to extract information from the data, then the information divergence should be used. These assumptions in turn guarantee the coherence and invariance of the evaluations, resolving the controversy in [16, 17, 18, 19]. Furthermore, any decision problem where external utility functions are given can be decomposed into two problems, the estimation problem which gives the optimal estimate, and the decision problem which is solely based on the optimal estimate.

Acknowledgement This work was partially supported by EPSRC grant GR/J17814. We would like to thank C. Williams for very stimulating discussions.

References

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

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

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

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

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

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

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

[8] H. Zhu and R. Rohwer. Information geometric measurements of generalisation. Technical Report NCRG/4350, Aston University, 1995. `ftp://cs.aston.ac.uk/neural/zhuh/generalisation.ps.Z`.

[9] S. Kullback and R. A. Leibler. On information and sufficiency. *Ann. Math. Statist.*, 22:79–86, 1951.

[10] A. W. F. Edwards. *Likelihood: An Account of the Statistical Concept of likelihood and Its Applications to Scientific Inference*. Cambridge University Press, Cambridge, 1972.

[11] H. Zhu and R. Rohwer. Bayesian invariant measurements of generalisation for discrete distributions. Technical Report NCRG/4351, Aston University, 1995. `ftp://cs.aston.ac.uk/neural/zhuh/discrete.ps.Z`.

[12] H. Zhu and R. Rohwer. Bayesian invariant measurements of generalisation for continuous distributions. Technical Report NCRG/4352, Aston University, 1995. `ftp://cs.aston.ac.uk /neural/zhuh/continuous.ps.Z`.

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

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

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

[16] D. J. C. MacKay. Bayesian interpolation. *Neural Computation*, 4(3):415–447, 1992.

[17] D. J. C. MacKay. A practical Bayesian framework for backpropagation networks. *Neural Computation*, 4(3):448–472, 1992.

[18] D. H. Wolpert. On the use of evidence in neural neworks. In S. J. Hanson, J. D. Cowan, and C. Lee Giles, editors, *Advances in Neural Information Processing Systems*, volume 5, pages 539–546, San Mateo, CA, 1993. Morgan Kaufmann.

[19] D. J. C. MacKay. Hyperparameters: Optimize, or integrate out? In G. Heidbreder, editor, *Maximum Entropy and Bayesian Methods, Santa Barbara 1993*, Dordrecht, 1995. Kluwer.