Accepted for oral presentation in ICNN'97. Area: Learning and Memory
ftp://ftp.santafe.edu/pub/zhuh/learn.ps.Z

Bayesian Geometric Theory of Learning Algorithms

Huaiyu Zhu[*]
Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA

Abstract

The problem of objective evaluation of learning algorithms is analyzed under the principles of coherence and covariance. The theory of Bayesian information geometry satisfies these principles and encompasses most of the commonly used learning criteria. Implications to learning theory are discussed.

1. Introduction

One fundamental problem in the study of learning algorithms is to find a scheme to compare them. Thousands of different learning algorithms are in use today. Some are simply variants with different parameter settings, others come from distinct assumptions claimed to be incompatible with each other. This poses obvious problems for an unfortunate user who has to choose one algorithm to suit his need from this "embarrassment of abundance". If he uses an "adaptive algorithm" to automatically sieve through many algorithms and choose the optimal one, the whole procedure is still another algorithm. Is it guaranteed to be better? Can we improve our result indefinitely by combining algorithms and investing in more computing resources? Most would agree that there must be an upper limit if our data are limited. But which factors will determine such a limit? Furthermore, if we put in more computing resources how will the result improve? Indeed, some may think that quite the opposite will happen, as it is often noticed that an over-sized neural network trained for too long is bound to "overfit" the noise in the data and give poorer results.

These questions are also important to a researcher as they determine whether a particular avenue of "improving" a learning rule is worthwhile to pursue. In this paper we shall outline the recent development of a branch of theoretical statistics called **Bayesian information geometry**, which appears to be capable of answering most of these question for most algorithms. The emphasis here is not on the statistical theory itself but rather on its implications to learning theory.

2. Fundamental principles of evaluation

Before embarking upon a road to find a general scheme capable of evaluating learning algorithms, we first need to set forth some criteria for admissible schemes. In this paper we adopt a statistical language in which a learning algorithm is simply an **estimator**, ie. a mapping from **sample** (aka data set) to **estimates** (aka weights). More details will be given later. It appears that a minimum requirement for an "objective evaluation" of estimators must include the following two principles:

Coherence Optimality of estimator is evaluated on its behavior on all samples.

Covariance Optimality is independent of the naming of either the samples or the data generators.

The coherence principle essentially says that our scheme should be able to compare rules which are based on other rules. Consider this example. Suppose a learning algorithm A says that if the data is a then estimate the weight as w_1. However once we actually observed a, we might be tempted to say that "since I now have new information (data a) I can make a better choice, and it is better to use rule B which would give w_2 as estimate". The coherence principle says that in this situation our overall behavior is equivalent to a rule C which should still be compared with A and B under the same scheme.

The covariance principle may look trivial to some while quite unacceptable to others. For example, one may have a linear least square problem which will be rather intractable if subject to an arbitrary nonlinear transformation. Others may consider it the most treasured property of neural networks to be used as black-boxes. The covariance principle simply says that if we consider a particular parameterization as important, we should explicitly single it out; Likewise, if we allow different parameterizations we should keep track of the transform so that the new optimal solution translates back to the optimal solution in the original problem. For intuitive illustration recall the transformation of the equation of a circle from Cartesian to polar coordinates.

Obviously these two principles already weed out most of the criteria often used to compare learning rules, but is

[*] Email: zhuh@santafe.edu

there anything left? Remarkably the answer is yes, and it appears that this answer, Bayesian information geometry, may be able to encompass most of the other criteria often used. That is, those criteria may not be coherent and covariant, but under special conditions usually implicitly assumed they correspond to special formulations in Bayesian information geometry.

3. Mathematical background

Some formal notations are unavoidable but we try to keep them to minimum.

A learning rule maps some **sample** (training data) $z \in Z$ to some **parameter** (weight) $w \in W$. The sample is assumed to be taken from a **true distribution** $p \in \mathcal{P}$, where $\mathcal{P}$ is the space of all probability distributions over the **sample space** Z. In a neural network with input x and output y the whole data set of x's and y's is z. The parameter w of a trained network represents an **estimate** $q = P(\cdot|w)$. The set $\mathcal{Q} \subseteq \mathcal{P}$ of all the possible estimates is the **computational model**. So a learning rule τ is an **estimator** mapping samples to estimates: $q = \tau(z)$.

In Bayesian information geometry we need to specify two things before using an estimator: a **prior** $P(p)$ which is a distribution over the space of all the true distributions, and an **information deviation** $D(p, q)$ which measures how much information is lost if we have arrived at an estimate q while the true distribution is p. The prior is also called a **statistical model**. Once these are specified, we then evaluate the estimator τ and estimate $q = \tau(z)$ by the **posterior average deviation**

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

$$E(q|z) := \int_{p} P(p|z) D(p, q). \tag{2}$$

By minimizing them the **optimal estimators** and **optimal estimates** are defined. It is well known in **Bayesian decision theory** that an estimator is optimal iff it gives optimal estimates for all the data. Therefore the coherence principle is satisfied.

In the above formulation it is also obvious that everything is invariant to parameterization since we have not used any parameterization at all. But we still need to specify $D(p, q)$ in a covariant way while retaining the meaning of “information deviation”. For this purpose we adopt the following δ**-deviation** where $\delta \in (0, 1)$,

$$D_\delta(p, q) := \int \frac{\delta p + (1-\delta) q - p^\delta q^{1-\delta}}{\delta(1-\delta)}. \tag{3}$$

The deviations D_0 and D_1 are defined as limits as $\delta \to 0$ and $\delta \to 1$:

$$D_1(p, q) = D_0(q, p) = \int \left(q - p + p \log \frac{p}{q} \right). \tag{4}$$

In fact, this definition is also applicable to finite measures, ie., things like probability distributions except that they may integrate to a finite number other than unity. The set of finite measures will be denoted $\widetilde{\mathcal{P}}$ and will be used extensively. It is out of space here to explain the statistical rationale behind these definitions, but the properties and results to be given shortly will hopefully at least lend some intuitive support. Interested readers may consult [1, 6, 7, 5, 3].

The following simple properties are obvious and they resemble properties of squared distance:

$$D_\delta(p, q) \geq 0. \tag{5}$$

$$D_\delta(p, q) = 0 \iff p = q. \tag{6}$$

$$\forall a \in \mathbb{R}_+ : \quad D_\delta(ap, aq) = a D_\delta(p, q). \tag{7}$$

$$D_\delta(p, q) = D_{1-\delta}(q, p). \tag{8}$$

$$D_{1/2}(p, q) = 2 \int (\sqrt{p} - \sqrt{q})^2. \tag{9}$$

Also $\forall p, q \in \mathcal{P}$:

$$D_\delta(p, q) = \frac{1 - \int p^\delta q^{1-\delta}}{\delta(1-\delta)}, \tag{10}$$

$$D_1(p, q) = \int p \log \frac{p}{q}. \tag{11}$$

D_1 is known as the **Kullback-Leibler deviation** or the **cross entropy** and $D_{1/2}$ is known as **Hellinger distance**. Let $P_{y|x}$ be a statistical transform (**Markov morphism** or kernel) transforming a family of measures $\{p_x, q_x, \ldots\}$ to $\{p_y, q_y, \ldots\}$. Then $D_\delta(p_x, q_x) \geq D_\delta(p_y, q_y)$; the equality holds if and only if $P_{y|x}$ is sufficient [4]. This means that D_δ captures all the information and nothing else.

Information deviation enables us to treat the set of finite measures as a well behaved space rather than simply a point set, similarly to the way functions may be considered as forming Banach spaces with the help of norms.

4. Ideal estimate, error decomposition, and projection

In order to describe the kind of results obtainable from information geometry we need several further concepts. First, we need to accept that things like p^δ may be regarded as elements of a Banach space (complete normed linear space) consisting of the δth power of finite measures [2]. For illustration, let the sample space Z be the ordinary real line $\mathbb{R}$, and let dz be the Lebesgue measure on Z. Suppose we use density function $f = p/dz$ (corresponding to $p = f\,dz$) then p^δ corresponds to $f^\delta \in L_{1/\delta}(dz)$, the space of $(1/\delta)$th power Lebesgue integrable functions. (For $\delta = 1/2$ we get the Hilbert space $L_2(dz)$.) However, we must keep in mind that properties of p^δ are essentially independent of carrier measures such as dz.

Next, we define δ-**straightness**. A curve p_t in $\widetilde{\mathcal{P}}$ is call a δ-**geodesic** if p_t^δ is a straight line (in the Banach space). The concepts such as δ-**flat** manifold, δ-**convex** set are likewise defined.

Now if we have a distribution $P(p)$ over $\mathcal{P}$ (it is not a member of $\mathcal{P}$ but a distribution of members in $\mathcal{P}$), we can define the δ-**average** of p as

$$a_\delta(p) = \begin{cases} \left\langle p^\delta \right\rangle^{1/\delta}, & \delta \in (0,1] \\ \exp \left\langle \log p \right\rangle, & \delta = 0, \end{cases} \tag{12}$$

where $\langle\cdot\rangle$ denote expectation under the distribution $P(p)$. Intuitively, p is a random member of $\mathcal{P}$ while $a_\delta(p)$ is a fixed member of $\widetilde{\mathcal{P}}$. Note that $a_\delta(p)$ may not be a member of $\mathcal{P}$ which is not δ-convex, just as the average of points on a sphere may be somewhere in the interior of the ball.

Let $P(p)$ be a prior over $\widetilde{\mathcal{P}}$, $z \in Z$, and denote the expectation under the posterior $P(p|z)$ as $\langle\cdot\rangle_z$. The δ-average over the posterior is called the δ-**ideal estimate** based on z.

Theorem 4.1 (Error decomposition) *Let $\widehat{p}$ be the δ-ideal estimate based on $z \in Z$. Then $\forall q \in \widetilde{\mathcal{P}}$:*

$$\left\langle D_\delta(p,q)\right\rangle_z = \left\langle D_\delta(p,\widehat{p})\right\rangle_z + D_\delta(\widehat{p},q), \tag{13}$$

where the **generalized posterior variance** *is given by*

$$\begin{aligned} &\left\langle D_\delta(p,\widehat{p})\right\rangle_z \\ &= \begin{cases} \left\langle \dfrac{\int p}{1-\delta} \right\rangle_z - \dfrac{\int \widehat{p}}{1-\delta}, & \delta \in [0,1) \\ \left\langle \displaystyle\int p \log p \right\rangle_z - \displaystyle\int \widehat{p} \log \widehat{p}, & \delta = 1. \end{cases} \end{aligned} \tag{14}$$

We may obtain some intuitive appreciation of (12)–(14) by comparing them with following familiar formulas, noting that there are now a family of deviations indexed by $\delta \in [0,1]$ instead of one single distance.

$$\widehat{x} = \langle x \rangle, \tag{15}$$

$$\left\langle \|x-a\|^2 \right\rangle = \left\langle \|x-\widehat{x}\|^2 \right\rangle + \|\widehat{x}-a\|^2, \tag{16}$$

$$\left\langle \|x-\widehat{x}\|^2 \right\rangle = \left\langle \|x\|^2 \right\rangle - \|\widehat{x}\|^2. \tag{17}$$

Now the ideal estimate is essentially a **point estimate** out of the posterior, albeit in the space $\widetilde{\mathcal{P}}$, while the posterior is a *distribution* of the true distributions, how good is this estimate? Under some mild conditions it can be shown that the ideal estimates are **sufficient statistics** of the posterior, ie. from which one can in principle reconstruct the whole posterior without the data. In other words, if we choose our estimate $\widehat{p}$ by minimizing the posterior mean information deviation then it will extract all the available information. This justifies calling D_δ as *information* deviation and the error decomposition theorem justifies calling $\widehat{p}$ as *ideal* estimate. There are reasons to believe that no other definition of information deviation will support the concept of ideal estimate.

As may be expected, in general the ideal estimate is difficult to compute and represent, as it usually lies in an infinite dimensional space. In certain cases it is exactly the **empirical distribution**. So what can we do with a model, such as a neural network, which is usually finite dimensional?

Theorem 4.2 (Projection) *Let $\mathcal{Q} \subseteq \widetilde{\mathcal{P}}$, and let $\widehat{p}$ be the δ-ideal estimate. Then the δ-***optimal estimate** *$\widehat{q} \in \mathcal{Q}$ is obtained by minimizing $D_\delta(\widehat{p},q)$ for $q \in \mathcal{Q}$. Furthermore, if $\mathcal{Q}$ is a submanifold, any (local) δ-optimal estimate $\widehat{q} \in \mathcal{Q}$ is a δ-***projection** *of $\widehat{p}$ onto $\mathcal{Q}$, ie. the δ-geodesic connecting $\widehat{p}$ and $\widehat{q}$ is* **orthogonal** *to any curve in $\mathcal{Q}$ which passes point $\widehat{q}$.*

The resemblance of this theorem to minimizing a squared distance on a linear model is obvious.

All the above discussions assume the prior to be unrelated to the model. Now we consider a special case, as is usually done in Bayesian methods, where the prior is constrained in a finite dimensional model.

Theorem 4.3 (Asymptotic Error) *Suppose the prior $P(p)$ is a smooth measure on a smooth finite dimensional submanifold $\mathcal{Q} \subseteq \widetilde{\mathcal{P}}$. Let $m = \dim \mathcal{Q}$. Then for most samples $z \in Z$ of large size n, the posterior mean deviation is $\left\langle D_\delta(p,\widehat{q})\right\rangle_z \approx m/2n$.*

This provides a universal learning curve for finite dimensional models. Although this is only a very special result in Bayesian information geometry, it nevertheless summarizes most of the results of learning curves in the literature. Numerous special instances may be found under various conditions, most of which boil down to the assumption of smooth priors on smooth manifolds (with smooth coordinates). Some of these results are given in the form of accumulated error being of the order $\frac{1}{2}m \log n$ due to the fact $\sum_n 1/n \approx \log n$.

5. Implications to Learning Theory

In essence the problem studied is a Bayesian decision problem in which the sole objective is to extract and retain information, but it also impacts upon other learning problems: Optimal decisions under other criteria may be based on the ideal estimate alone because it is sufficient.

It may be noticed that we have used the word “statistical model” and “computational model” in the text, meaning the prior $P(p)$ and the model $\mathcal{Q}$, respectively. Their roles are quite distinct and it is not required that the prior be concentrated in $\mathcal{Q}$. With this setting the relation between **estimation** and **approximation** becomes immediately clear. The problem of **robustness** is also easily formalized: If the prior is spilled out of the model, how bad the optimal estimate within the model may be?

We are now in a position to resolve a notorious controversy about model selection: Should we choose a larger or smaller model? The answer depends on which model is under consideration. A "smaller" statistical model means a more concentrated prior and sharper results. We should not make it sharper than we have knowledge to, but we should always strive at this direction. On the other hand, a smaller computational model gives us less room to represent our knowledge so leads to poorer results. If computational model cannot contain the statistical model (usual situation), there will be an approximation error. If we have a computational model containing the convex hull of the support of the prior the best result is simply the ideal estimate.

So how do we explain the overfitting encountered in practice? The answer lies in the way priors are usually incorporated into algorithms. In most learning algorithm there are some "fiddle factors", such as learning rate, which are chosen by trial and error from many "typical" problems. In other words, they are chosen to fit a *prior distribution* of problems. However, since the statistical and computational models are usually not distinguished clearly, these factors also depend on the model size in a non-trivial way, such that choosing large model has the effect of choosing an algorithm with less prior knowledge. In infinite dimensional space (such as curve fitting) with finite data, if the prior is weak, even the ideal estimate will be poor, hence come the overfitting. In fact, a Gaussian measure with a "spherical covariance in the space L_2" is exactly the **white noise**, a sample of which cannot be fixed by any finite amount of data. On the other hand, once the prior preference of smoothness is fixed a larger $\mathcal{Q}$ always leads to better results. Details are given in [8] where it is also shown how to represent the prior knowledge that "a function is somewhat smooth" by specific **Gaussian random fields**.

This theory can also be reconciled with those theories without priors. On the one hand, many non-Bayesian methods are simply not optimal [9]. On the other hand, many good non-Bayesian algorithms are special cases of the ideal estimators with implicitly specified priors. For example, with the **0-uniform prior**, the 1-ideal estimate corresponds to the **empirical distribution**, while the 1-optimal estimate on any model is the **maximum likelihood estimate** [7].

We have used the results on linear Gaussian models as an analogy to illustrate our general results, but in fact the former is also a special case of the latter: The (0,1)-dual geometry reduces to Euclidean (Hilbert) geometry with the inner product defined by the inverse of the covariance. This encompasses all the **quadratic approximation** theories (**least mean squares**) in function spaces, with or without **regularisation** [3].

Some Bayesians may insist that the whole posterior instead of an estimate be used. Clearly the sufficiency of the ideal estimate reduces the force of whatever its supporting arguments. Furthermore, usually these methods actually use a Monte Carlo **simulated posterior**. That is, whenever a prediction is called for, it is made by a sampled weight value. This is exactly equivalent to sampling from the **posterior marginal distribution**, ie. the 1-ideal estimate.

In summary, if we have limited data but infinite computing power, the best thing is to figure out the ideal estimate; If our computing power is restricted by a certain model, the best thing is to approximate a projection of the ideal estimate. Of course, in practice we are still faced with the problem of achieving this while minimizing computational cost. The promise of Bayesian information geometry is that all these are now technical problems with clearly stated goals and without the usual philosophical controversies.

Acknowledgments

This work was partially supported by EPSRC grant GR/J17814 while I was in Aston NCRG. I am indebted to R. Rohwer and C. Williams for extensive discussions. I thank several other people there, especially C. Bishop, I. Nabney, D. Saad, and D. Lowe, for discussions. My work in SFI is supported by TXN, Inc.

References

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

[2] J. Neveu. *Mathematical Foundations of the Calculus of Probability*. Holden-Day, San Francisco, 1965. Translated from French, 1964, Masson.

[3] H. Zhu. On the mathematical foundation of learning algorithms. Submitted to *Machine Learning*, 1996.

[4] H. Zhu. On information and sufficiency. SFI working paper 97-02-014. Submitted to *Ann. Stat.*, 1997.

[5] H. Zhu and R. Rohwer. Bayesian geometric theory of statistical inference. Submitted to *Ann. Stat.*, 1995.

[6] H. Zhu and R. Rohwer. Bayesian invariant measurements of generalisation. *Neural Proc. Lett.*, 2(6):28–31, 1995.

[7] 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.

[8] H. Zhu and R. Rohwer. Bayesian regression filters and the issue of priors. *Neural Comp. Appl.*, 4(3):130–142, 1996.

[9] H. Zhu and R. Rohwer. No free lunch for cross validation. *Neural Computation*, 8(7):1421–1426, 1996.