Accepted for presentation at ICONIP'96

Using Ancillary Statistics in On-Line Learning Algorithms

Huaiyu Zhu and Richard Rohwer
Dept of Computer Science and Applied Mathematics
Aston University, Birmingham B4 7ET, UK
Email: H.Zhu@aston.ac.uk, R.J.Rohwer@aston.ac.uk

Abstract— Neural networks are usually curved statistical models. They do not have finite dimensional sufficient statistics, so on-line learning on the model itself inevitably loses information. In this paper we propose a new scheme for training curved models, inspired by the ideas of ancillary statistics and adaptive critics. At each point estimate an auxiliary flat model (exponential family) is built to locally accommodate both the usual statistic (tangent to the model) and an ancillary statistic (normal to the model). The auxiliary model plays a role in determining credit assignment analogous to that played by an adaptive critic in solving temporal problems. The method is illustrated with the Cauchy model and the algorithm is proved to be asymptotically efficient.

1 Introduction

Neural network (NN) training algorithms are essentially statistical estimators since they map random samples to some general rules or distributions underlying these samples. The main difference between NN models and classical statistical models is that NN models are in general non-linear, which in statistical terms means non-exponential family models. We shall call them curved models. This creates two problems, local minima and loss of information. The former is well known in NN community and will not be addressed here. The latter is less well known and often confused with the former. This is the issue considered here.

For a deterministic optimisation problem, including a stochastic problem in batch learning mode, the optimal choice of steplength is determined by the Hessian. For a flat statistical model, the optimal choice of steplength is determined by the variance, or equivalently by the sample size. Here we deal with on-line stochastic training of a curved model, so we must take into account the interplay between these two effects.

We have previously shown [14, 13, 12] that any statistical inference problem can be decomposed, at least in theory, into two problems: computing an ideal estimate in a flat model and projecting it onto the curved model. The first step does not involve curvature, while the second step is deterministic. The trouble is that the ideal estimate is usually infinite dimensional, so computing it is tantamount to retaining the whole data set.

It was an old idea of R. A. Fisher [5] that by keeping a finite dimensional ancillary statistic we ought to be able to construct an asymptotically efficient algorithm. This amounts to locally expanding the model to a flat model of higher dimension, spanned by the tangents and normals of the original model. The estimate in the tangent direction corresponds to the usual statistics, while that in the normal direction is called an ancillary statistic. The estimate is projected onto the model so a new auxiliary model can be constructed. The process is iterated until convergence. Fisher showed that if one starts from a consistent estimator, a single extra step will give an efficient estimator. This is still not an on-line method.

One of the best known on-line learning algorithms is the adaptive critic algorithm for learning in temporal problems. It can be explained as an auxiliary statistical model, although this is generally not recognised due to the special form of the curved model, composed of iterated conditional distributions in a Markov chain. The "moving target" method [10] can be interpreted as an adaptive critic method for structural credit assignment. The moving targets correspond to the point estimate in the auxiliary model, which unfortunately has to be infinite dimensional since the auxiliary model is fixed.

In this paper we combine all these ideas together to construct a method which is on-line, finite dimensional and asymptotically efficient, even when applied to models without a finite dimensional sufficient statistic. As far as we are aware, this is the first example of this kind in either the statistical literature or the neural network literature. The basic idea is to let the finite dimensional auxiliary model move with the current estimate, thus forming a moving frame along the model [7], and to transfer the moving target from one frame to its successor without projecting onto the main model.

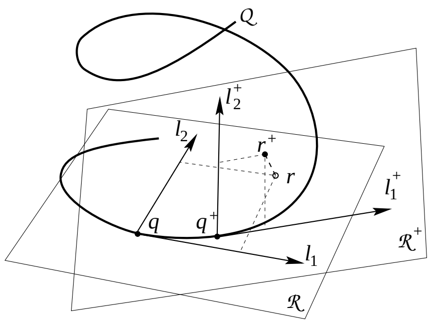
In this paper we shall illustrate this method on the Cauchy model, a one dimensional model whose minimum sufficient statistic is infinite dimensional. This makes it a relatively simple model for illustrating the non-trivial aspects of more general NN models.

2 Statistical background and outline of the algorithm

Consider a sample space X and the space of probability distributions $\mathcal{P}$ on X. If X is of infinite size $\mathcal{P}$ forms an infinite-dimensional manifold. A statistical model is a finite dimensional submanifold $\mathcal{Q} \subseteq \mathcal{P}$. We shall also consider the space $\widetilde{\mathcal{P}}$ of finite measures on X. It is also an infinite dimensional manifold, containing $\mathcal{P}$ as a smooth submanifold. See [4, 1, 2, 6, 14, 12].

It has been shown [14, 12] that in general a statistical inference problem can be specified by a prior $P(p)$ on $\mathcal{P}$, the information divergence $D_\delta(p, q)$, $\delta \in [0, 1]$, and the model $\mathcal{Q}$. For a given sample x, there exists a unique ideal estimate, called the δ-estimate, $\widehat{p} \in \widetilde{\mathcal{P}}$, given by $\widehat{p}^\delta = \int_p P(p|x) p^\delta$. The optimal estimate in the model, $\widehat{q} \in \mathcal{Q}$, is given by the δ-projection of $\widehat{p}$ onto $\mathcal{Q}$, which minimises $D_\delta(\widehat{p}, q)$. It is important to note that this only works if we allow $\widehat{p} \in \widetilde{\mathcal{P}}$ to be unnormalised.

Here we shall only consider maximum likelihood estimates (MLE), which are equivalent to 1-estimates with a 0-uniform prior. In this case the ideal estimate $\widehat{p}$ is simply the empirical distribution, and the optimal estimate $\widehat{q} \in \mathcal{Q}$ is solution to $\mathrm{Min}_{q \in \mathcal{Q}} K(\widehat{p}, q)$, where the generalised KL-divergence is given by $D_1(p, q) = K(p, q) := \int \left(q - p + p \log \frac{p}{q} \right)$. If both p, q are normalised, ie., if $\int p = \int q = 1$, then we get the usual $K(p, q) = \int p \log(p/q)$.



$q \in \mathcal{Q}$ is the current estimate. l_1 and l_2 are the tangent and normal of $\mathcal{Q}$ at q. $r \in \mathcal{R}$ is the current auxiliary estimate. The entities with superscript + correspond to the updated version.

Figure 1: Schematic illustration of the adaptive critic method

To avoid complicated notation, in this paper we shall only consider one-dimensional models $\mathcal{Q}$, ie. smooth curves $\mathcal{Q} \in \widetilde{\mathcal{P}}$. As observed by Fisher [5], the reason behind the information loss of curved models is the turning of the tangent in the log-likelihood space when the estimate moves. To hold all the information in the sample relevant for statistical estimation on model $\mathcal{Q}$, we need a flat model $\mathcal{R} \subseteq \widetilde{\mathcal{P}}$ spanned by $\mathcal{Q}$. If the smallest such $\mathcal{R}$ is infinite dimensional, then it is impossible to do so exactly without keeping an increasing amount of data. This is the case for the Cauchy distribution model. Locally, however, a curve only turns in the direction of its normal. The basic idea of our proposed algorithm is to construct an $\mathcal{R}$ to hold information both in the tangent and normal directions of the model. This guards against the information loss caused by the turning of tangent within the osculating plane. The residual information loss is due to non-zero torsion, the fact that the osculating space itself is also turning. We shall show that this only gives a higher order term so the algorithm is asymptotically efficient. The outline of the algorithm is as follows (cf. Figure 1).

1. Find an estimate $q \in \mathcal{Q}$ parameterised by μ.

2. Construct tangent $l_1 := \partial_\mu \log q$ and normal $l_2 := \partial_\mu^2 \log q$ of the model $\mathcal{Q}$ at point q. Define the auxiliary model as the exponential family (not normalised) spanned by $[l_1, l_2]$,

$$\mathcal{R} := \left\{ r : r = q \exp(\theta_1 l_1 + \theta_2 l_2),\ \theta \in \mathbb{R}^2 \right\}, \tag{2.1}$$

3. Update the MLE $r \in \mathcal{R}$ parameterised by $[\theta_1, \theta_2]$, in light of new data.

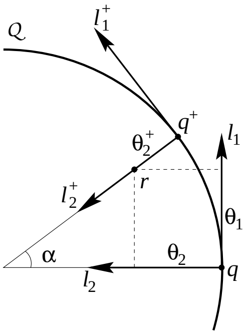
4. Compute new estimate $q^+ \in \mathcal{Q}$ by projecting r onto $\mathcal{Q}$. Compute the new frame $[l_1^+, l_2^+]$. Project r onto $r^+ \in \mathcal{R}^+$, the new auxiliary model spanned by $[l_1^+, l_2^+]$. This is accomplished by computing $[\theta_1^+, \theta_2^+]$.

5. Transfer statistics from $\mathcal{R}$ to $\mathcal{R}^+$, and calculate the effective sample size.

6. Go back to step 3.

3 Details

The relevant geometry for a statistical model in IID training is the exponential geometry, a special case of information geometry [4, 1, 2, 6]. Roughly speaking, it is defined by the Fisher information metric, which specifies an inner product on the tangent space, and the exponential affine connection, which specifies that exponential families are to be considered flat submanifolds. Note that as we are considering geometry for the whole $\widetilde{\mathcal{P}}$, exponential families are not normalised, and the metric is defined by the correlation $\langle uv \rangle_q$, instead of the covariance $\langle u, v \rangle_q$. [12]



$q \in \mathcal{Q}$ is the current estimate, with tangent l_1 and normal l_2. $r \in \mathcal{R}$ is the current auxiliary point, with auxiliary coordinates θ_1 and θ_2. The new estimate on $\mathcal{Q}$ is q^+, with tangent l_1^+ and normal l_2^+. The auxiliary point r is unchanged and is represented in the new auxiliary coordinates as θ_1^+ and θ_2^+, where $\theta_1^+ = 0$.

Figure 2: Change of coordinates caused by curvature

Locally, a curve looks like a helix up to third order approximation. It is characterised by its metric κ_1^2, curvature κ_2 and torsion κ_3. For the Cauchy model, it can be calculated that $\kappa_1^2 = 1/2$, $\kappa_2^2 = 7/2$, $\kappa_3^2 = 20/7$. The value of κ_2 is the absolute curvature in $\widetilde{\mathcal{P}}$. The relative curvature in $\mathcal{P}$ was calculated by [5, 4] as ${\kappa'_2}^2 = 5/2$, using covariance in place of correlation. The difference ${\kappa_2}^2 - {\kappa'_2}^2$ is the square of normal curvature of $\mathcal{P}$ in $\widetilde{\mathcal{P}}$, which is unity for any direction in $\mathcal{P}$ [11]. The calculation of torsion for the Cauchy model appears to be original.

Doing statistics on the exponential family $\mathcal{R}$ can be easily accomplished by projecting $\widehat{p}$ to $r \in \mathcal{R}$ with Newton's method (denoting $\partial_i := \partial / \partial_{\theta_i}$)

$$[\Delta\theta_i] = -\left[\partial_i \partial_j K(\widehat{p}, r)\right]^{-1} \left[\partial_j K(\widehat{p}, r)\right] = \left[\int r l_i l_j\right]^{-1} \left[\int \widehat{p} l_j - \int r l_j\right], \tag{3.1}$$

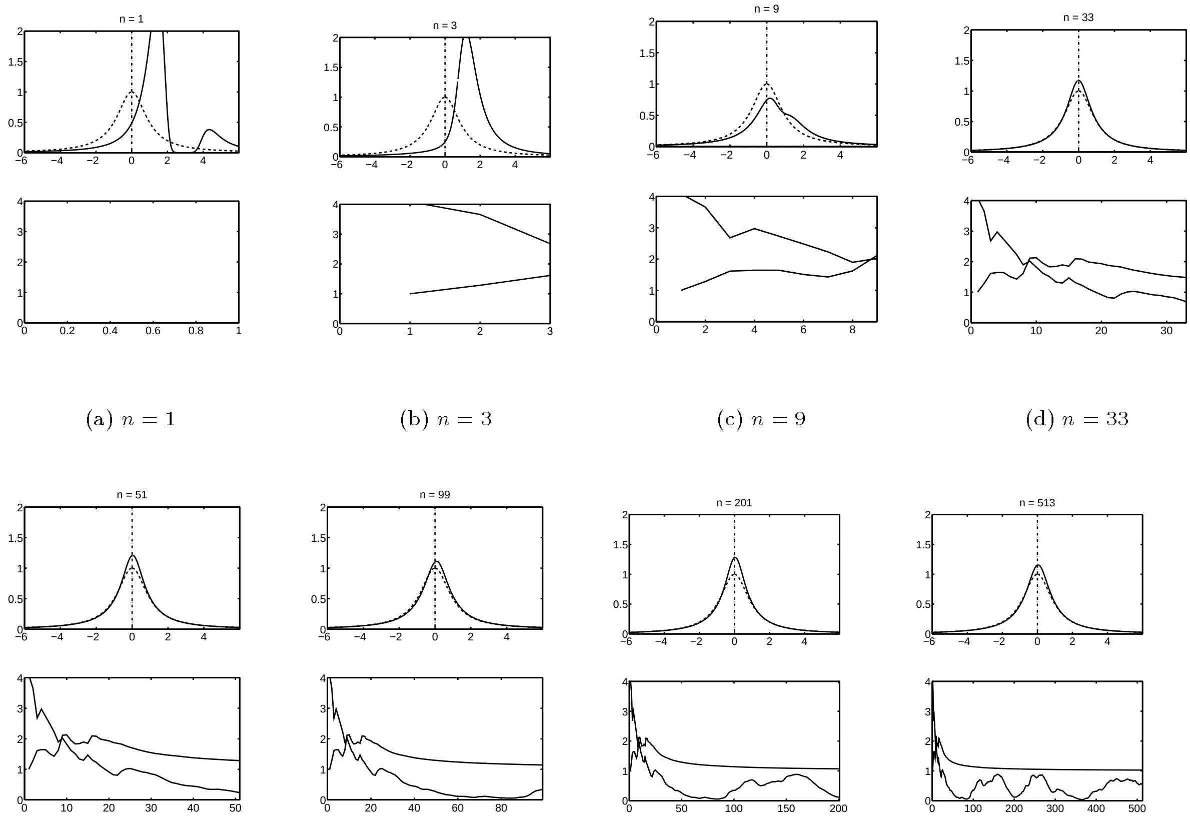
where $\int r l_i l_j$ and $\int r l_j$ are functions of θ and can be computed without knowing the sample, and $\int \widehat{p} l_j$ is simply the sample mean of l_j which can be accumulated easily.

Up to second order approximation, $\mathcal{Q}$ can be locally identified with the osculating circle, and $\mathcal{R}$ with the osculating plane. The new estimate q^+ and auxiliary estimate r^+ can be calculated by projecting r onto them (Figure 2).

$$\Delta\mu = \frac{\alpha}{K} = \frac{1}{K} \operatorname{atan}\left(\frac{\theta_1 K}{1 - \theta_2 K^2}\right), \tag{3.2}$$

$$\theta_1^+ = 0, \qquad \theta_2^+ = \frac{1}{K^2}\left(1 - \sqrt{(\theta_1 K)^2 + (1 - \theta_2 K^2)^2}\right), \tag{3.3}$$

where $K := \kappa_1\kappa_2$ is the rate the tangent turns relative to the parameter μ.



(a) $n = 1$ (b) $n = 3$ (c) $n = 9$ (d) $n = 33$

(e) $n = 51$ (f) $n = 99$ (g) $n = 201$ (h) $n = 513$

The upper part of each sub-plot depicts p (dashed line) and r (solid line). The lower part of each sub-plot depicts the expected squared error (smoother line) and the true squared error (rougher line), both multiplied by the sample size.

Figure 3: A typical run of the algorithm

The crucial point of this algorithm is using $\int r l_j^+$ in place of $\int \hat{p} l_j^+$ for the new auxiliary model. That is, the previous information as summarised by $r \in \mathcal{R}$ is projected onto $r^+ \in \mathcal{R}^+$, so that the algorithm is on-line. Because the osculating plane is also turning, there is inevitable information loss as r is replaced by r^+. This is determined by the angle by which the osculating plane turned, $\alpha = T\Delta\mu$, where $T := \kappa_1\kappa_3$ is the rate of direction change of the osculating plane in the μ coordinate. It can be shown that, asymptotically, the one-step efficiency ϵ, the proportion of information retained, is at least $\cos^2\alpha$. Since the direction change of the osculating plane is orthogonal to the model, the actual information loss is even less. Numerical experiments show that $\epsilon = \cos\alpha$ is more accurate. This is used to update $r^+ \in \mathcal{R}^+$. In any case the exact rate is irrelevant asymptotically, since all these reduce to $\epsilon = 1 - a\Delta s^2$, where a is a constant and s is the arc-length parameter. The overall information loss after n samples can be shown to be $a \log n + O(1)$, even if we have used $\epsilon = 1 - b\Delta s$, with $b \neq a$. The asymptotic efficiency of the algorithm is therefore

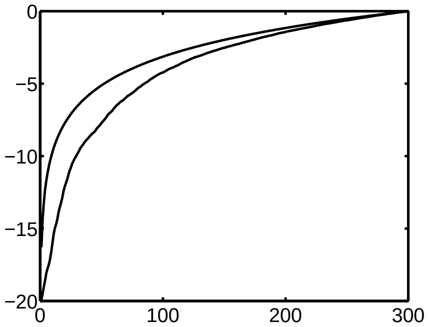
$$e_n = 1 - \frac{a \log n}{n} - O\left(\frac{1}{n}\right) \to 1, \qquad (n \to \infty). \tag{3.4}$$

The initial estimate can be obtained by using a good classical estimator on a small sample, such as the optimal L-estimator [8]. We find that a sample of size five is good enough in our case.

4 Experiments and Discussion

One typical run of the algorithm is shown in Figure 3. Note that r is markedly non-Cauchy initially. The bimodal shape is caused by the tangent and the normal. The normalised expected squared error $\frac{n}{\kappa_1^2}(\mu - \mu_0)^2$ approaches 1 as $n \to \infty$ since the algorithm is asymptotically efficient. The actual squared error is χ_1^2 distributed (of which this figure only gives one sample), since the Cauchy distribution is locally one dimensional.

The expected and actual loss of information are plotted in Figure 4, showing that our calculation of asymptotic information loss is correct. It should be pointed out that there is an additional loss of a small sample used for the initial estimate.



The upper line is the expected loss of information $\kappa_3^2 \log n$; the lower line is the observed loss averaged over 20 runs. They are shifted and superimposed to show the asymptotic equivalence.

Figure 4: Expected and actual loss of information

The novelty of this algorithm is that the auxiliary information is retained and transferred at each step. Intuitively speaking, this means that "If you don't know which unimodal model to estimate, then use a multi-modal model". By allowing r to be outside $\mathcal{Q}$, or even outside $\mathcal{P}$, we are able to represent whatever information as simply the point r, which need not be outside $\widetilde{\mathcal{P}}$. Furthermore, locally or asymptotically, it is enough for r to be in a two dimensional exponential $\mathcal{R}$.

Recently Leen and Orr [9] proposed a stochastic search method to avoid inverting the stochastically updated Hessian. It appears that their method is intuitively equivalent to using (compare with (3.2))

$$\Delta\mu \approx \theta_1(1 + (\theta_2 K^2) + (\theta_2 K^2)^2 + \dots), \tag{4.1}$$

Since in our method the "denominator" is maintained this problem does not occur. It would be interesting to see how their method performs on the Cauchy model.

Most of the interesting curved models, including the Cauchy model and most tanh-type NNs such as the MLP, BM and Hopfield net, are mixture-of-exponential models. The best known method for such models is the EM algorithm, which has recently been given an information-geometric interpretation [3]. It would also be interesting to elucidate the relation between our method and the EM method. Our current understanding is that EM is more like Fisher's original algorithm which is not on-line.

For multilayer networks with an m-dimensional weight space, the tangent space is m-dimensional, but the normal space becomes m^2-dimensional. The method described here can still be applied, with more complicated differential-geometric notation. This will appear elsewhere. In many interesting cases using the diagonal of the normal tensor would be reasonably good so the algorithm requires keeping a $2m$-dimensional statistic. At present we do not know the asymptotic efficiency of such a diagonal approximation.

5 Conclusion

We have proposed and analysed an on-line training algorithm for curved models. It is asymptotically efficient even for models without finite dimensional sufficient statistics. This removes the need for any *ad hoc* adjustable parameters in training algorithms, such as a momentum term and step-length. The idea and performance of the algorithm is illustrated with the Cauchy model. It is expected that this method will have a significant impact in the area of on-line training of non-linear models. This also shows the importance of rigorous statistical theory, especially information geometry, in this active area.

Acknowledgement

We would like to thank people in the Aston NCRG for a discussion on asymptotic information loss, especially D. Saad, C. Williams and D. Barber for their comments. This work was supported by EPSRC grant GR/J17814.

References

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

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

[3] S. Amari. Information geometry of the EM and em algorithms for neural networks. Technical Report METR94-4, Univ. Tokyo, 1994. `ftp://archive.cis.ohio-state.edu/pub/neuroprose/amari.geometryofem.tar.Z`.

[4] B. Efron. Defining the curvature of a statistical problem (with applications to second order efficiency) (with discussion). *Ann. Statist.*, 3:1189–1242, 1975.

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

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

[7] W. Klingenberg. *A Course in Differential Geometry.* Graduate Texts in Mathematics, 51. Springer-Verlag, New York, 1978. Translated by David Hoffman.

[8] E. L. Lehmann. *Theory of Point Estimation.* J. Wiley, New York, 1983.

[9] G. B. Orr and T. K. Leen. Using curvature information for fast stochastic search. Manuscript, Oregan Grad. Inst. Sci. Tech. (Presened at post-NIPS'95 Workship), 1995.

[10] R. Rohwer. The "moving targets" training algorithm. In D. S. Touretzky, editor, *Advances in Neural Information Processing Systems*, volume 2, pages 558–565, San Mateo, CA, 1990. Morgan Kaufmann.

[11] H. Zhu. On the curvatures of information manifold. Manuscript, 1996.

[12] H. Zhu and R. Rohwer. A Bayesian geometric theory of statistical inference. Submitted to *Ann. Statist.*

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

[14] 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`.