Transients and Asymptotics of Natural Gradient Learning

Magnus Rattray and David Saad
Neural Computing Research Group, Aston University, Birmingham B4 7ET, UK.

Abstract

We analyse natural gradient learning in a two-layer feed-forward neural network using a statistical mechanics framework which is appropriate for large input dimension. We find significant improvement over standard gradient descent in both the transient and asymptotic phases of learning.

1 Introduction

One of the most popular forms of neural network training is on-line learning, in which training examples (input-output pairs) are presented sequentially and independently at each learning iteration. Natural gradient learning was recently proposed by Amari as a more principled alternative to standard on-line gradient descent [1]. When learning to emulate a stochastic rule with some probabilistic model this learning algorithm has the desirable properties of asymptotic optimality, given a sufficiently rich model which is differentiable with respect to its parameters, and invariance to reparameterizations of our model distribution. This latter property is achieved by viewing the parameter space of the model as a Reimannian space in which local distance is defined by the KL-divergence [2, 3]. The Fisher information matrix then plays the role of a Reimannian metric in this space. The natural gradient learning rule is obtained by pre-multiplying the standard gradient with the inverse of this matrix. In practice, we require knowledge of the input distribution in order to determine the Fisher information matrix. Yang and Amari discuss methods of pre-processing training examples to obtain a whitened Gaussian process for the inputs [3]. If this is possible, then when the input dimension N is large compared to the number of hidden units K, inversion of the Fisher information for two-layer feed-forward networks requires only $O(N^2)$ operations, providing an efficient and practical algorithm.

We quantify the benefits of natural gradient learning using a recent statistical mechanics description of the learning process which is appropriate when $N \gg K$ [4]–[7]. This formalism allows us to compare performance with standard gradient descent in both the transient and asymptotic phases of learning, and to obtain generic results in terms of task complexity and non-linearity. We show that trapping time in an unstable fixed point which dominates the training time is significantly reduced by using natural gradient learning and exhibits a slower power law increase as task complexity grows. We also find that asymptotic performance is greatly improved.

2 Natural gradient learning

Consider a mapping from an input space $\boldsymbol{\xi} \in \Re^N$ onto a scalar $\phi_{\mathbf{J}}(\boldsymbol{\xi}) = \sum_{i=1}^{K} g\,(\mathbf{J}_i^{\mathrm{T}}\boldsymbol{\xi})$, which defines a soft committee machine (we call this the 'student' network), where we choose $g(x) \equiv \mathrm{erf}(x/\sqrt{2})$ to be the activation function of the hidden units, $\mathbf{J} \equiv \{\mathbf{J}_i\}_{1 \leq i \leq K}$ is the set of input to hidden weights and the hidden to output weights are set to one. We can then define a Gaussian noise model for output ζ_{m} given input $\boldsymbol{\xi}$ which is parameterized by $\mathbf{J}$,

$$p_{\mathbf{J}}(\zeta_{\mathrm{m}}|\boldsymbol{\xi}) = \frac{1}{\sqrt{2\pi\sigma_{\mathrm{m}}^2}} \exp\left(\frac{-(\zeta_{\mathrm{m}} - \phi_{\mathbf{J}}(\boldsymbol{\xi}))^2}{2\sigma_{\mathrm{m}}^2}\right) . \tag{1}$$

Let $(\boldsymbol{\xi}^\mu, \zeta^\mu)$ be the μth input-output pair in a sequence of training examples. The training error at each learning iteration is taken to be proportional to the log-likelihood of the current example under our noise model, $\epsilon_{\mathbf{J}}(\zeta^\mu, \boldsymbol{\xi}^\mu) \equiv \frac{1}{2}(\zeta^\mu - \phi_{\mathbf{J}}(\boldsymbol{\xi}^\mu))^2$ and the most basic learning algorithm is to adapt the student weights in the negative gradient direction of this error at each iteration. However, such an algorithm is not consistent with our probabilistic interpretation of the problem, since it depends on our particular choice of model parameterization. A more principled learning algorithm can be defined by viewing the manifold of models as a Reimannian space in which local distance is defined by the KL-divergence [3]. The Fisher information matrix $\mathbf{G} = [G_{i\alpha,k\beta}]$ (where $1 \leq i,k \leq K$ and $1 \leq \alpha,\beta \leq N$) defines the appropriate metric in this space [1],

$$G_{i\alpha,k\beta} = \left\langle \frac{\partial \log p_{\mathbf{J}}(\zeta_{\mathrm{m}}|\boldsymbol{\xi})}{\partial J_{i\alpha}} \, \frac{\partial \log p_{\mathbf{J}}(\zeta_{\mathrm{m}}|\boldsymbol{\xi})}{\partial J_{k\beta}} \right\rangle_{\{\zeta_{\mathrm{m}},\xi\}} . \tag{2}$$

The brackets denote an average over ζ_{m}, according to equation (1), followed by an average over the input distribution. The natural gradient direction is found by premultiplying the training error gradient by the inverse of this matrix.

Amari has determined the Fisher information matrix for a general two-layer network. For our particular choice of activation function and with components of $\boldsymbol{\xi}$ selected independently each iteration from a zero-mean Gaussian distribution with unit variance, we find $\mathbf{G} = \mathbf{A}/\sigma_{\mathrm{m}}^2$, where

$$\mathbf{A}_{ik} = \frac{2}{\pi\sqrt{\Delta}}\left[\mathbf{I} - \frac{1}{\Delta}\left((1+Q_{kk})\mathbf{J}_i\mathbf{J}_i^{\mathrm{T}} + (1+Q_{ii})\mathbf{J}_k\mathbf{J}_k^{\mathrm{T}} - Q_{ik}(\mathbf{J}_i\mathbf{J}_k^{\mathrm{T}} + \mathbf{J}_k\mathbf{J}_i^{\mathrm{T}})\right)\right] \tag{3}$$

with $Q_{ik} \equiv \mathbf{J}_i^{\mathrm{T}}\mathbf{J}_k$ and $\Delta = (1+Q_{ii})(1+Q_{kk}) - Q_{ik}^2$.

3 Deriving the dynamics

We use a statistical mechanics description of the learning process which is exact in the limit of large N and provides an accurate model of mean behaviour for realistic values of N [4, 5]. The training example outputs are generated by a 'teacher' network corrupted by Gaussian noise,

$$p_{\mathbf{B}}(\zeta^\mu|\boldsymbol{\xi}^\mu) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(\zeta^\mu - \phi_{\mathbf{B}}(\boldsymbol{\xi}^\mu))^2}{2\sigma^2}\right) . \tag{4}$$

Here, $\phi_{\mathbf{B}}(\boldsymbol{\xi}^{\mu}) = \sum_{n=1}^{M} g\,(\mathbf{B}_n^{\mathrm{T}}\boldsymbol{\xi})$ defines a teacher which may differ in complexity from the student network introduced in the previous section. Due to the flexibility of this teacher mapping [9] we can represent a variety of learning scenarios within this theoretical framework. The weight update at each iteration of natural gradient learning is given by,

$$\mathbf{J}_i^{\mu+1} = \mathbf{J}_i^{\mu} + \frac{\eta}{N}\sum_{j=1}^{K} \mathbf{A}_{ij}^{-1}\delta_j^{\mu}\boldsymbol{\xi}^{\mu}\ , \tag{5}$$

where $\delta_i^{\mu} \equiv g'(\mathbf{J}_i^{\mathrm{T}}\boldsymbol{\xi}^{\mu})[\sum_{n=1}^{M} g(\mathbf{B}_n^{\mathrm{T}}\boldsymbol{\xi}^{\mu}) - \sum_{j=1}^{K} g(\mathbf{J}_j^{\mathrm{T}}\boldsymbol{\xi}^{\mu}) + \rho^{\mu}]$ and ρ^{μ} is zero-mean Gaussian noise of variance σ^2. The learning rate η is divided by the input dimension for convenience. Notice that knowledge of the noise variance is not required to execute this algorithm.

The Fisher information matrix can be inverted using the partitioning method described in [3] and each block is some additive combination of the identity matrix and outer products of the student weight vectors. Using the methods described in [4] it is then straightforward to derive equations of motion for a set of order parameters $\mathbf{J}_i^{\mathrm{T}}\mathbf{J}_k \equiv Q_{ik}$, $\mathbf{J}_i^{\mathrm{T}}\mathbf{B}_n \equiv R_{in}$, and $\mathbf{B}_n^{\mathrm{T}}\mathbf{B}_m \equiv T_{nm}$, measuring overlaps between student and teacher vectors. These order parameters are necessary and sufficient to determine the generalization error [4]. The equations of motion are coupled first order differential equations for the order parameters with respect to the normalized number of examples $\alpha = \mu/N$ and we can integrate them numerically in order to determine the evolution of the generalization error.

Although our equations of motion are sufficient to describe learning for arbitrary system size, the number of order parameters is $\frac{1}{2}K(K-1) + KM$ so that the numerical integration soon becomes rather cumbersome as K and M grow and analysis becomes difficult. To obtain generic results in terms of system size we therefore exploit symmetries which appear in the dynamics for isotropic tasks and structurally matched student and teacher ($K = M$ and $T = T\delta_{nm}$). In this case we define a four dimensional system via $Q_{ij} = Q\delta_{ij} + C(1-\delta_{ij})$ and $R_{in} = R\delta_{in} + S(1-\delta_{in})$ which can be used to study the dynamics for arbitrary K and T (here, δ_{ij} denotes the Kronecker delta). In [10] we show how the Fisher information matrix can be inverted for this reduced dimensionality system. At the cost of some generality we therefore obtain a much simplified dynamical system which is amenable to analysis.

As for standard gradient descent [4], the dynamics is characterized by two major phases of learning. Initially, the order parameters are trapped in an unstable fixed point, the symmetric phase, in which the generalization error remains at a constant non-zero value and the student-teacher overlaps are virtually indistinguishable (see figure 1(a)). Eventually, small perturbations due to the random initial conditions lead to an escape from this phase and convergence towards zero generalization error[1]. If the teacher is corrupted by noise then the learning rate must be annealed at late times in order for the generalization error to decay. The fastest decay for natural gradient learning is achieved by setting $\eta = 1/\alpha$ and this leads to a inverse decay law for

[1] we define the generalization error to be the expected error in the absence of noise. The prediction error contains an additive contribution proportional to the noise variance

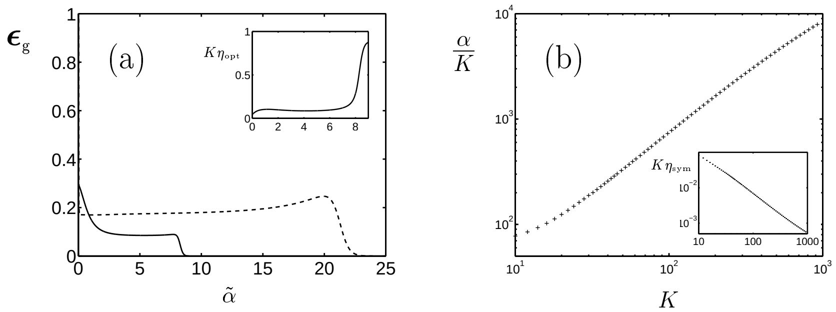

Figure 1: In (a) the generalization error is shown for optimal natural gradient learning (solid line) and optimal gradient descent (dashed line) for $K = 10$, $T = 1$ and zero noise (we define $\tilde{\alpha} = 10^{-2}\alpha$). The inset shows the optimal learning rate for natural gradient learning. In (b) the time required for optimal natural gradient learning to reach a generalization error of $10^{-4}K$ is shown as a function of K on a log-log scale. The inset shows the optimal learning rate within the symmetric phase. In both (a) and (b) we used initial conditions $R = 10^{-3}$, $Q = U[0, 0.5]$ and $S = C = 0$.

the generalization error. This choice saturates the Cramer-Rao bound and provides asymptotic performance equalling the best batch algorithm [1]. The benefit of natural gradient learning over standard gradient descent is two-fold: the symmetric phase is shortened significantly and better asymptotic performance is obtained. In the following two sections we consider each phase in turn.

3.1 Transient dynamics

Unfortunately, even for standard gradient descent an analytical study of the symmetric phase is only possible for small learning rates, which are far from optimal. Such an approach is not appropriate for realistic learning rates and often gives misleading results. It is also unclear how to proceed for natural gradient learning even in this limit, since the Fisher information is singular at the fixed point considered in [4]. In order to obtain generic results in this case we apply a recent method for obtaining globally optimal time-dependent learning parameters by variational maximization of the total reduction in generalization error [6]. We obtain the optimal learning rates for both gradient descent and natural gradient learning in order to compare optimal performance for both methods.

We note that the impact of output noise on the symmetric phase dynamics is not considered explicitly here. For low noise levels there is no noticeable effect on the length of the symmetric phase, or on the order parameters and generalization error within this phase. For larger noise levels the symmetric phase increases in length and the student norms increase, resulting in a larger generalization error. However, we feel that these are secondary effects and that most essential features of this phase are captured by the noiseless dynamics. This is not true for later stages of learning, where the inclusion of noise completely alters qualitative features of the dynamics.

In figure 1(a) we compare optimal performance for $K = 10$ and $T = 1$, which

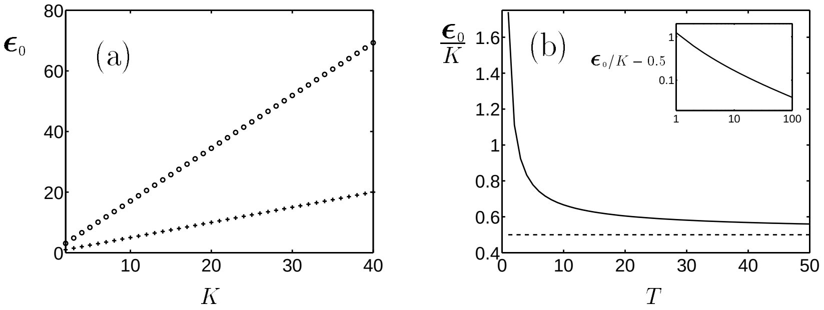

Figure 2: Prefactor for the asymptotic decay of the generalization error ($\epsilon_g = \sigma^2\epsilon_0/\alpha$): (a) shows the prefactor for $T = 1$ as a function of K for optimal gradient descent (circles) and natural gradient learning (crosses) while (b) shows how the prefactor for optimal gradient descent (large K) decays towards $K/2$ as T increases, which is the prefactor for natural gradient learning.

indicates a significant shortening of the symmetric phase for natural gradient learning (the inset shows the optimal learning rate). Figure 1(b) shows the time required for natural gradient learning to reach a generalization error of $10^{-4}K$ as a function of K (for $T = 1$). The learning time is dominated by the symmetric phase, so that these results provide a scaling law for the length of the symmetric phase in terms of task complexity. We find that the escape time for natural gradient learning scales as K^2, while the inset shows that the learning rate within the symmetric phase scales as K^{-2}. Scaling laws for gradient descent were determined in [7], showing a $K^{\frac{8}{3}}$ law for escape time and a learning rate scaling of $K^{-\frac{5}{3}}$ within the symmetric phase. The escape time for the adaptive gradient learning rule studied in [7] scales as $K^{\frac{5}{2}}$, which is also worse than for natural gradient learning.

3.2 Asymptotic dynamics

In the presence of output noise the learning rate must be annealed in order to achieve zero generalization error asymptotically and it is well known that natural gradient learning is asymptotically optimal with $\eta = 1/\alpha$ [1]. We apply recent analytical results for the annealing dynamics of gradient descent [8] in order to compare the asymptotic generalization error for natural gradient learning with the result for gradient descent. We find that the asymptotic result for natural gradient learning takes a very simple form: $\epsilon_g \sim K\sigma^2/2\alpha$ [10]. In figure 2 we compare the prefactor of the generalization error decay for natural gradient learning and optimal gradient descent ($\epsilon_g = \sigma^2\epsilon_0/\alpha$). Figure 2(a) shows the result for $T = 1$ as a function of K, indicating a linear scaling for both methods (there are slight deviations for gradient descent). In figure 2(b) we compare the decay prefactors for each method as a function of T, showing how the difference diverges as T is reduced. This can be explained by examining the asymptotic expression for the Fisher information matrix [10]. For large T the diagonals of this matrix are $O(1/\sqrt{T})$ and equal (for large N) while all other terms are at most $O(1/T)$, so that the Fisher information is effectively proportional to the identity matrix in this limit and Natural gradient learning is asymptotically equivalent to gradient descent. However, for small T the diagonals are $O(T^2)$ while the off-diagonals remain finite, so that the Fisher information is dominated by off-diagonals in this limit.

4 Conclusion

We have analysed natural gradient learning under a statistical mechanics framework which is exact in the limit of large input dimension. We find significant improvements over standard gradient descent in both the transient and asymptotic stages of learning, with improved scaling of learning time against task complexity. The major drawback with using Natural gradient learning is that the input distribution is required in order to determine the Fisher information matrix exactly, and for non-Gaussian inputs it is unclear whether inversion can be carried out efficiently. Efficient averaging and inversion may be achieved using the matrix momentum algorithm suggested by Orr and Leen [11] and we are currently investigating this approach within the present framework.

Acknowledgement This work was supported by the EPSRC grant GR/L19232.

References

[1] S. Amari *Neural Computation* **10**(2) 251 (1998).

[2] H. Y. Yang, S. Amari *Advances in Neural Information Processing Systems* vol 10, ed M. I. Jordan, M. J. Kearns and S. A. Solla (Cambridge, MA: MIT Press, 1998).

[3] H. Y. Yang, S. Amari 'Natural Gradient Descent for Training Multi-Layer Perceptrons' submitted to *IEEE Transactions on Neural Networks* (1998).

[4] D. Saad, S. A. Solla, *Phys. Rev. Lett.* **74**, 4337 (1995); *Phys. Rev. E* **52** 4225 (1995).

[5] D. Barber, D. Saad, P. Sollich *Europhysics Letters* **34** 151 (1996).

[6] D. Saad, M. Rattray *Phys. Rev. Lett.* **79** 2578 (1997).

[7] A. H. L. West, D. Saad *Phys. Rev.* E **56** 3426 (1997).

[8] T. K. Leen, B. Schottky, D. Saad *Advances in Neural Information Processing Systems* vol 10, ed M. I. Jordan, M. J. Kearns and S. A. Solla (Cambridge, MA: MIT Press, 1998).

[9] C. Cybenko *Math. Control Signals and Systems* **2** 303 (1989).

[10] M. Rattray, D. Saad, S. A. Solla, S. Amari 'Natural gradient descent for on-line learning' (in preparation, 1998).

[11] G. B. Orr, T. K. Leen *Advances in Neural Information Processing Systems* vol 9, ed M. C. Mozer, M. I. Jordan and T. Petsche (Cambridge, MA: MIT Press, 1997) p 606