Natural Gradient Descent for On-Line Learning

Magnus Rattray
Computer Science Department, University of Manchester, Manchester M13 9PL, United Kingdom

David Saad
Neural Computing Research Group, Aston University, Birmingham B4 7ET, United Kingdom

Shun-ichi Amari
Laboratory for Information Synthesis, RIKEN Brain Science Institute, Saitama, Japan
(Received 15 May 1998)

Natural gradient descent is an on-line variable-metric optimization algorithm which utilizes an underlying Riemannian parameter space. We analyze the dynamics of natural gradient descent beyond the asymptotic regime by employing an exact statistical mechanics description of learning in two-layer feed-forward neural networks. For a realizable learning scenario we find significant improvements over standard gradient descent for both the transient and asymptotic stages of learning, with a slower power law increase in learning time as task complexity grows. [S0031-9007(98)07950-2]

PACS numbers: 87.10.+e, 02.50.–r, 05.20.–y

Optimization of a parametrized probability distribution to approximate some unknown underlying distribution instanced by a set of examples is an important interdisciplinary problem which recurs in various settings. On-line optimization, which is carried out by modifying the parameters iteratively in response to a stream of examples, is of particular importance as it is arguably the method of choice for large systems and nonstationary tasks. One of the most popular on-line optimization methods is stochastic gradient descent on some cost function. Over the years, many attempts have been made to devise practical alternative methods which will provide an improved performance over the entire learning session (i.e., not only asymptotically).

Natural gradient descent (NGD) was recently proposed by Amari as a principled alternative to standard on-line gradient descent (GD) [1]. When learning to emulate a stochastic rule with some probabilistic model, e.g., a feed-forward neural network, NGD has the desirable properties of asymptotic optimality, given a sufficiently rich model which is differentiable with respect to its parameters, and invariance to reparametrization of our model distribution. These properties are achieved by viewing the parameter space of the model as a Riemannian space in which local distance is defined by the Kullback-Leibler divergence [2,3]. The Fisher information matrix provides the appropriate metric in this space. If the training error is defined as the negative log likelihood of the data under our probabilistic model, then the direction of steepest descent in this Riemannian space is found by premultiplying the error gradient with the inverse of the Fisher information matrix; this defines the NGD learning direction. In practice we require knowledge of the input distribution in order to determine the Fisher information matrix. Yang and Amari discuss methods of preprocessing 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 matrix for a two-layer feed-forward network requires only $O(N^2)$ operations, providing an efficient and practical algorithm.

In this Letter we investigate the properties of NGD beyond the asymptotic regime using a nonlinear model which may be regarded as a probabilistic two-layer feed-forward neural network. Such models are of significant importance given their ability to approximate any continuous input-output mapping to an arbitrary degree of accuracy [4]. We quantify the benefits of NGD using a recent statistical mechanics framework which allows an exact solution to the dynamics as $N \to \infty$ for fixed K [5]. In particular, we consider a realizable learning scenario and apply a site-symmetric ansatz which simplifies the dynamical equations and allows us to obtain results for arbitrary task complexity and nonlinearity. For this special case we show that trapping time in an unstable fixed point which often dominates the training time is significantly reduced by using NGD and exhibits a slower power law increase as task complexity grows. We also find that asymptotic performance is greatly improved, with the generalization performance of NGD equaling the universal asymptotics for batch learning [6].

We consider a mapping from an input space $\boldsymbol{\xi} \in \mathbb{R}^N$ 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 $g(x)$ is some sigmoid activation function for the hidden units, $\mathbf{J} \equiv \{\mathbf{J}_i\}_{1 \le i \le K}$ is the set of input to hidden weights, and the hidden to output weights are set to one. We choose a Gaussian noise model for output ζ given input $\boldsymbol{\xi}$ which is parametrized by $\mathbf{J}$,

$$p_{\mathbf{J}}(\zeta \mid \boldsymbol{\xi}) = \frac{1}{\sqrt{2\pi\sigma_{\mathrm{m}}^2}} \exp\left\{ \frac{-[\zeta - \phi_{\mathbf{J}}(\boldsymbol{\xi})]^2}{2\sigma_{\mathrm{m}}^2} \right\}. \quad (1)$$

The Fisher information matrix is then defined,

$$\mathbf{G} = \int d\boldsymbol{\xi}\, p(\boldsymbol{\xi}) \int d\zeta\, p_{\mathbf{J}}(\zeta \mid \boldsymbol{\xi}) [\nabla_{\mathbf{J}} \log p_{\mathbf{J}}(\zeta \mid \boldsymbol{\xi})] \times [\nabla_{\mathbf{J}} \log p_{\mathbf{J}}(\zeta \mid \boldsymbol{\xi})]^{\mathrm{T}}. \quad (2)$$

With $g(x) = \mathrm{erf}(x/\sqrt{2})$ and an isotropic Gaussian input distribution $p(\boldsymbol{\xi}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$ we find a particularly simple expression for $\mathbf{G} \equiv \mathbf{A}/\sigma_{\mathrm{m}}^2$ with $\mathbf{A}$ in block form,

$$\mathbf{A}_{ij} = \frac{2}{\pi\sqrt{\Delta_{ij}}} \left\{ \mathbf{I} - \frac{1}{\Delta_{ij}} [(1 + Q_{jj})\mathbf{J}_i\mathbf{J}_i^{\mathrm{T}} + (1 + Q_{ii})\mathbf{J}_j\mathbf{J}_j^{\mathrm{T}} - Q_{ij}(\mathbf{J}_i\mathbf{J}_j^{\mathrm{T}} + \mathbf{J}_j\mathbf{J}_i^{\mathrm{T}})] \right\}, \quad (3)$$

where $\Delta_{ij} = (1 + Q_{ii})(1 + Q_{jj}) - Q_{ij}^2$ and $Q_{ij} \equiv \mathbf{J}_i^{\mathrm{T}}\mathbf{J}_j$.

In order to analyze the dynamics of NGD we need to specify the function being learned. A sequence of independently drawn inputs $\boldsymbol{\xi}^\mu$: $\mu = 1, 2, \ldots$ are labeled by a "teacher" network corrupted by Gaussian noise,

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

Here, $\phi_{\mathbf{B}}(\boldsymbol{\xi}^\mu) = \sum_{n=1}^{M} g(\mathbf{B}_n^{\mathrm{T}}\boldsymbol{\xi}^\mu)$ defines the teacher network which may differ in complexity from the student. Because of the flexibility of this mapping [4] we can represent a variety of learning scenarios within this framework. The weight update at each iteration of NGD is then given by

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

where $\delta_i^\mu \equiv g'(\mathbf{J}_i^{\mathrm{T}}\boldsymbol{\xi}^\mu)[\phi_{\mathbf{B}}(\boldsymbol{\xi}^\mu) - \phi_{\mathbf{J}}(\boldsymbol{\xi}^\mu) + \rho^\mu]$, ρ^μ is zero-mean Gaussian noise of variance σ^2, and the learning rate η is scaled by the input dimension. 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,

$$\mathbf{A}_{ij}^{-1} = \alpha_{ij}\mathbf{I} + \sum_{k=1}^{K}\sum_{l=1}^{K} \Gamma_{ij}^{kl}\mathbf{J}_k\mathbf{J}_l^{\mathrm{T}}. \quad (6)$$

Using the methods described in [5] we derive equations of motion for a set of macroscopic order parameters $\mathbf{J}_i^{\mathrm{T}}\mathbf{J}_j \equiv Q_{ij}$, $\mathbf{J}_i^{\mathrm{T}}\mathbf{B}_n \equiv R_{in}$, and $\mathbf{B}_n^{\mathrm{T}}\mathbf{B}_m \equiv T_{nm}$, measuring overlaps between student and teacher weight vectors. These order parameters are necessary and sufficient to determine the generalization error $\epsilon_g = \langle \frac{1}{2}[\phi_{\mathbf{J}}(\boldsymbol{\xi}) - \phi_{\mathbf{B}}(\boldsymbol{\xi})]^2 \rangle_\xi$ [5]. 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. 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 (further details will be given elsewhere [7]). As for GD, a typical learning scenario is often characterized by trapping in one or more unstable fixed points in which the generalization error remains at a constant nonzero value while the student adapts slowly because of an inherent symmetry between different permutations of its hidden nodes. Figure 1(a) shows an example of the generalization error evolution for NGD (solid line) and GD (dashed line). Notice that the approach to the plateau, or symmetric phase, is slower for NGD. We find that a very brief initial stage of GD (to $\alpha \simeq 1$) improves performance considerably. Eventually, small perturbations due to the random initial conditions lead to an escape from the symmetric phase and convergence towards zero generalization error. 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.

Although our equations of motion are sufficient to describe learning for arbitrary system size, they number $\frac{1}{2}K(K + 1) + KM$ so that the numerical integration soon becomes rather cumbersome as K and M grow and analysis becomes difficult (the complexity of inverting the Fisher information matrix also grows with increasing K). 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 $T = T\delta_{nm}$). This site-symmetric ansatz is rigorously justified only for the special case of symmetric initial conditions and further investigations are required to determine the validity of this approximation in general for large values of K (fixed points other than those considered here have been reported [8] and it is

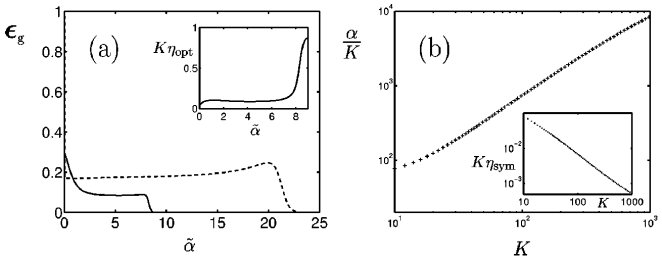

FIG. 1. In (a) the generalization error is shown for optimal NGD (solid line) and optimal GD (dashed line) for $K = 10$ (we define $\tilde{\alpha} = 10^{-2}\alpha$). The inset shows the optimal learning rate for NGD. In (b) the time required for optimal NGD to reach a generalization error of $10^{-4}K$ 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 $T = 1$, zero noise and initial conditions $R = 10^{-3}$, $Q = U[0, 0.5]$ and $S = C = 0$. A brief stage of GD is used before NGD is started.

unclear whether or not their basins of attraction are negligible). Simulations of the GD dynamics for K up to 10, with random initial conditions, show good correspondence with the symmetric system. In this case we define four new order parameters via $Q_{ij} = Q\delta_{ij} + C(1 - \delta_{ij})$ and $R_{in} = R\delta_{in} + S(1 - \delta_{in})$ (here, δ_{ij} denotes the Kronecker delta). For this simplified system symmetries suggest the following form for the tensor Γ in Eq. (6):

$$\Gamma_{ij}^{kl} = \gamma_1\delta_{ij}\delta_{ik}\delta_{il} + \gamma_2(\delta_{ik}\delta_{il} + \delta_{jk}\delta_{jl}) + \gamma_3(\delta_{ik}\delta_{jk} + \delta_{il}\delta_{ij}) + \gamma_4\delta_{kl}\delta_{ij} + \gamma_5\delta_{ik}\delta_{jl} + \gamma_6\delta_{jk}\delta_{il} + \gamma_7(\delta_{jl} + \delta_{ik}) + \gamma_8(\delta_{jk} + \delta_{il}) + \gamma_9\delta_{kl} + \gamma_{10}\delta_{ij} + \gamma_{11}\,. \quad (7)$$

The inversion of $\mathbf{A}$ then reduces to solving an 11-dimensional linear system to determine γ. At the cost of some loss in generality we therefore obtain a much simplified dynamical system with complexity independent of K.

In [5] an analysis of the transient dynamics for GD was carried through for small η, since this allowed an analytical expression for the symmetric fixed point for the idealized situation where $Q = C$ (in practice $Q > C$ even for very small learning rates). Such an analysis is not possible for NGD because the dynamics never approaches this fixed point (the Fisher information matrix becomes singular when $Q = C$). In any case, a small η analysis will be of limited value since it is "diffusion" terms in the dynamics (which are quadratic in η) which set the learning time scale and determine the optimal and maximal learning rate during the symmetric phase. In order to obtain generic results for larger learning rates we therefore apply a recent method for obtaining globally optimal time-dependent learning parameters within the present framework [9]. This is achieved by a functional optimization of the total change in generalization error with respect to the learning rate. We obtain the optimal learning rates for both GD and NGD in order to compare optimal performance for both methods. The maximal learning rate, above which the student weight vector norms diverge or the symmetric fixed point becomes stable, is typically of the same order.

Figure 1(a) compares the optimal performance of NGD (solid line) and GD (dashed line) for $K = 10$, $T = 1$ and zero noise, indicating a significant shortening of the symmetric phase for NGD (the inset shows the optimal learning rate for NGD). Figure 1(b) shows the time required for NGD 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 NGD scales as K^2, while the inset shows that the optimal learning rate within the symmetric phase approaches a decay of K^{-2}. Scaling laws for GD were determined in [10], showing a $K^{8/3}$ law for escape time and a learning rate scaling of $K^{-5/3}$ within the symmetric phase. The escape time for the adaptive GD rule studied in [10] scales as $K^{5/2}$, which is also worse than for NGD. The impact of output noise on the symmetric phase dynamics is not analyzed here although we can make some qualitative observations. 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. A quantitative study of these effects and their impact on the above scaling relations will be the subject of future work [7].

In the presence of output noise the learning rate must be annealed in order to achieve zero generalization error asymptotically and it is known that NGD is asymptotically optimal, in terms of the covariances of the student-teacher weight deviations (the quadratic estimation error), with $\eta = 1/\alpha$, saturating the Cramer-Rao bound and equaling in performance even the best batch methods [1]. However, the quadratic estimation error has no direct interpretation in terms of generalization ability. In order to determine the asymptotic generalization error decay we apply recent analytical results for the annealing dynamics of GD [11] (again for realizable isotropic tasks) in order to compare the asymptotic generalization error for NGD with the result for GD. We find that the asymptotic result for NGD is indeed optimized by choosing $\eta = 1/\alpha$ and takes a very simple form: $\epsilon_g \sim K\sigma^2/2\alpha \;\forall\; T$. This equals the universal asymptotics for optimal maximum likelihood and Bayes estimators which depend only on the learning machine's number of degrees of freedom [6]. NGD is therefore asymptotically optimal in terms of both generalization error and quadratic estimation error. In Fig. 2 we compare the prefactor of the generalization error decay for NGD and optimal GD ($\epsilon_g = \sigma^2\epsilon_0/\alpha$). 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 GD). In Fig. 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. 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 NGD is asymptotically equivalent to GD. 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. However, it should be noted that the optimal learning rate decay prefactor for GD will not generally be known, so

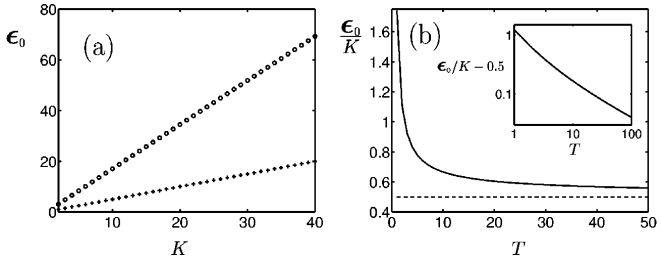

FIG. 2. Prefactor for the asymptotic decay of the generalization error: (a) shows the prefactor for $T = 1$ as a function of K for optimal GD (circles) and NGD (crosses) while (b) shows how the prefactor for optimal GD (large K) decays towards $K/2$ (which is the prefactor for NGD) as T increases. The inset to (b) shows the GD result on a log-log scale.

that these results provide only an upper bound for the performance of GD. In practice NGD can be expected to provide even greater improvement over GD.

To summarize, we have solved the dynamics of NGD using a statistical mechanics framework which is exact for large input dimension. Using a site-symmetric ansatz which resulted in a reduced dimensionality system allowed us to determine generic behavior in terms of task complexity K and nonlinearity T. An analysis of the transient was made possible by using a recent method for determining globally optimal learning parameters [9]. For the optimized system we find that trapping time in the symmetric phase scales as K^2, which is an improvement over both GD and the adaptive algorithm considered in [10]. Asymptotically we find that NGD saturates the universal bounds on generalization performance and provides a significant improvement even over optimized GD, especially for small T. We note here that the optimal learning rate during the symmetric phase decays more rapidly with K than one might expect *a priori* ($\eta_{\rm sym} \sim K^{-2}$) and in practice it may prove difficult to predict learning parameters giving optimal transient performance. An important task is therefore to determine robust learning strategies with minimal dependence on arbitrary parameters. NGD satisfies this criterion at late times, since it is known that $\eta \sim 1/\alpha$ is optimal asymptotically but is still dependent on good parameter choice at earlier times.

M. R. and D. S. were supported by EPSRC Grant No. GR/L19232.

[1] S. Amari, Neural Comput. **10**, 251 (1998).
[2] H. H. Yang and S. Amari, in *Advances in Neural Information Processing Systems*, edited by M. I. Jordan, M. J. Kearns, and S. A. Solla (MIT Press, Cambridge, MA, 1998), Vol. 10, p. 385.
[3] H. H. Yang and S. Amari (to be published).
[4] G. Cybenko, Math. Control Signals Systems **2**, 303 (1989).
[5] D. Saad and S. A. Solla, Phys. Rev. Lett. **74**, 4337 (1995); Phys. Rev. E **52**, 4225 (1995).
[6] S. Amari and N. Murata, Neural Comput. **5**, 140 (1993).
[7] M. Rattray and D. Saad (to be published).
[8] M. Biehl, P. Riegler, and C. Wöhler, J. Phys. A **29**, 4769 (1996).
[9] D. Saad and M. Rattray, Phys. Rev. Lett. **79**, 2578 (1997).
[10] A. H. L. West and D. Saad, Phys. Rev. E **56**, 3426 (1997).
[11] T. K. Leen, B. Schottky, and D. Saad, in *Advances in Neural Information Processing Systems* (Ref. [2]), p. 301.