Online learning with ensembles

R. Urbanczik
Neural Computing Research Group, Aston University, Aston Triangle, Birmingham B4 7ET, United Kingdom
(Received 8 February 2000)

Supervised online learning with an ensemble of students randomized by the choice of initial conditions is analyzed. For the case of the perceptron learning rule, asymptotically the same improvement in the generalization error of the ensemble compared to the performance of a single student is found as in Gibbs learning. For more optimized learning rules, however, using an ensemble yields no improvement. This is explained by showing that for any learning rule f a transform $\tilde{f}$ exists, such that a single student using $\tilde{f}$ has the same generalization behavior as an ensemble of f students.

PACS number(s): 87.10.+e

Online learning, where each training example is presented just once to the student, has proved to be a very successful paradigm in the study of neural networks using methods from statistical mechanics [1]. On the one hand, it makes it possible to rigorously [2] analyze a wide range of learning algorithms. On the other hand, online algorithms can in some cases yield a performance that equals that of the Bayes optimal inference procedure, e.g., asymptotically, when the probability of the data is a smooth function of the parameters of the network [3].

Some problems, however, do remain. For nonsmooth cases, which arise, e.g., in classification tasks, the Bayes optimal procedure yields a superior generalization performance, even asymptotically, to that of online algorithms [4,5]. Also, even for smooth problems, the online dynamics often has suboptimal stationary points arising from symmetries in the network architecture. Then the sample size needed to reach the asymptotic regime will scale faster than linearly with the number of free parameters if no prior knowledge is built into the initial conditions of the dynamics [6].

It thus seems of interest to ask which extensions of the online framework make sense. The findings quoted above indicate that, using a reasonable update rule, it is not possible to store all of the information contained in a training example into a single weight vector. Thus one should study learning systems that have a larger state space than just a single weight vector. Here we shall consider using an ensemble of students randomized by the choice of initial condition. Focusing on classification problems, we first analyze realizable learning in a perceptron. So the learning dynamics is based on a training set of αN input/output pairs (ξ^μ, τ^μ), $\xi^\mu \in \mathbb{R}^N$, and $\tau^\mu = \mathrm{sgn}(B^T \xi^\mu)$, where B is the unknown N-dimensional weight vector defining the teacher. For convenience we assume $|B| = 1$. The ensemble consists of K students and at time step μ the i-th student is characterized by a weight vector $J_i^\mu \in \mathbb{R}^N$. When classifying a new input ξ one may then use the majority vote of the K students instead of relying on the output $\mathrm{sgn}(J_i^{\mu T} \xi)$ of just a single student.

The dynamics of the i-th student takes the form

$$J_i^{\mu+1} = J_i^\mu + \xi^\mu N^{-1} f(\mu/N, |J_i^\mu|, B^T \xi^\mu, J_i^{\mu T} \xi^\mu), \qquad (1)$$

and the choice of the real valued function f defines the learning rule. Reasonably, f may only depend on the third argument $B^T \xi^\mu$ via its sign τ^μ, but it is not helpful to make this explicit in the notation. Note that all of the members of the ensemble learn from the same training examples, and these are presented in the same order.

Assuming that the components of the example inputs are independent random variables picked from the normal distribution on $\mathbb{R}$, the state of the ensemble can be described by the order parameters $R_i(\alpha) = B^T J_i^{\alpha N}$ and $Q_{ij}(\alpha) = J_i^{\alpha N T} J_j^{\alpha N}$. For a reasonable choice of f [2], the order parameters will be nonfluctuating for large N and satisfy the following differential equations:

$$\dot{R}_i = \langle y f_i^\alpha \rangle_{x_i, y},$$

$$\dot{Q}_{ij} = \langle x_i f_j^\alpha + x_j f_i^\alpha + f_i^\alpha f_j^\alpha \rangle_{x_i, x_j, y}, \qquad (2)$$

$$f_i^\alpha \equiv f(\alpha, Q_{ii}^{1/2}, y, x_i),$$

where y and the x_i are zero mean Gaussian random variables with covariances $\langle x_i y \rangle = R_i$ and $\langle x_i x_j \rangle = Q_{ij}$. We shall only consider the case where the initial values J_i^0 are picked independently from the uniform distribution on a sphere with radius $\sqrt{P(0)}$. Then for large N the initial conditions for Eq. (2) are $R_i(0) = Q_{ij}(0) = 0$ for $i \neq j$ and $Q_{ii}(0) = P(0)$. These conditions are invariant under permutations of the site indices i and this also holds for the system of differential equations (2). Thus this site symmetry will be preserved for all time and we need only consider the three order parameters $R(\alpha) = R_i(\alpha)$, $P(\alpha) = Q_{ii}(\alpha)$, and $Q(\alpha) = Q_{ij}(\alpha)$ for $i \neq j$. Since the length of the students is of little interest, it will often be convenient to consider the normalized overlaps $r(\alpha) = R(\alpha)/\sqrt{P(\alpha)}$ and $q(\alpha) = Q(\alpha)/P(\alpha)$.

A new input ξ, picked from the same distribution as the training inputs, will be classified by the ensemble using a majority vote, that is, by

$$\sigma(\xi) = \mathrm{sgn}\left[\sum_{i=1}^{K} \mathrm{sgn}(J_i^{\alpha N T} \xi) \right]. \qquad (3)$$

As an alternative to using a majority vote, one might consider constructing a new classifier by averaging the weight vectors of the students, setting $\overline{J^{\alpha N}} = K^{-1} \Sigma_i J_i^{\alpha N}$. As in the

Gibbs theory [7], a simple application of the law of large numbers yields proof that the two classifiers are equivalent in the large K limit if $q(\alpha)=\mathcal{O}(1)$; that is, $\sigma(\xi)=\mathrm{sgn}(\overline{J}^{\alpha N}\xi)$ for almost all inputs. In the sequel we shall only consider the large K limit, assuming that $K \ll N$ so that the fluctuations in the site symmetry of the initial conditions can be ignored. The generalization error ϵ_e of the ensemble, which is the probability of misclassifying ξ, is then given by $\epsilon_e=\epsilon[r(\alpha)/\sqrt{q(\alpha)}]$ where

$$\epsilon(x)=\frac{1}{\pi}\arccos x. \tag{4}$$

Similarly, the generalization error of a single student in the ensemble is $\epsilon_s=\epsilon(r(\alpha))$.

We shall first consider a soft version of the perceptron learning rule:

$$f=\eta|J_i^\mu|H\left(\tau^\mu\frac{k}{\sqrt{1-k^2}}\frac{J_i^{\mu T}\xi^\mu}{|J_i^\mu|}\right)\tau^\mu, \tag{5}$$

where $H(x)=\frac{1}{2}\mathrm{erfc}(x/\sqrt{2})$ and η is a time dependent learning rate. For $k=0$ this reduces to Hebbian learning, whereas $k=1$ yields the perceptron learning rule. Note, however, that the $|J_i^\mu|$ prefactor makes the dynamics invariant with respect to the scaling of the student weight vectors. From Eq. (2) one obtains for the order parameters

$$\dot{r}=\frac{\eta}{\sqrt{2\pi}}(1-r^2)-\frac{\eta^2}{2}r[\epsilon(kr)-\tfrac{1}{2}\epsilon(k^2)],$$

$$\dot{q}=\frac{2\eta}{\sqrt{2\pi}}r(1-q)+\eta^2\left((1-q)\epsilon(kr)-\tfrac{1}{2}\epsilon(k^2q)+\frac{q}{2}\epsilon(k^2)\right). \tag{6}$$

We first consider the perceptron learning rule, i.e., $k=1$. In the limit $r,q\to 1$ one finds $\dot{r}\sim\eta\sqrt{2/\pi}(1-r)-\eta^2\epsilon(r)/2$ and $\dot{q}\sim\eta\sqrt{2/\pi}(1-q)-\eta^2\epsilon(q)/2$; that is, r and q satisfy the same differential equation. If the learning rate schedule is such that this limit is reached, this means that $(1-r)/(1-q)$ will approach 1 for large α. Hence asymptotically, $\epsilon_e\sim\epsilon(\sqrt{r(\alpha)})$, and the same improvement by a factor of $1/\sqrt{2}$ in the generalization error of the ensemble compared to single student performance is found as in Gibbs learning. [Interestingly, the same asymptotic relationship between ϵ_e and ϵ_s also holds for the Adatron learning rule $f=-\Theta(-\tau^\mu J_i^{\mu T}\xi^\mu)J_i^{\mu T}\xi^\mu$.] The optimal asymptote of the learning rate schedule is $\eta\sim 2\sqrt{2\pi}/\alpha$ and this yields an $\epsilon_e\sim(2\sqrt{2})/\pi\alpha\approx 0.90/\alpha$ decay of the ensemble generalization error. This is very close to the $0.88/\alpha$ decay found for the optimal single student algorithm [5].

We next consider improving the performance by tuning k. From Eq. (6) one easily sees that single student performance is optimized when $k=r$. Asymptotically, this may be achieved by setting $k\sim 1-4/\alpha^2$ and choosing the optimal learning schedule that is asymptotically the same as the one for the standard perceptron learning rule. Then already a single student achieves $\epsilon_s\sim(2\sqrt{2})/\pi\alpha$, which is the same large α behavior as the ensemble in the $k=1$ case. Unfortunately r and q now have a different asymptote and one finds $1-q\ll 1-r$. So for all practical purposes the ensemble collapses to a single point, and for large α to leading order $\epsilon_e\sim\epsilon_s$.

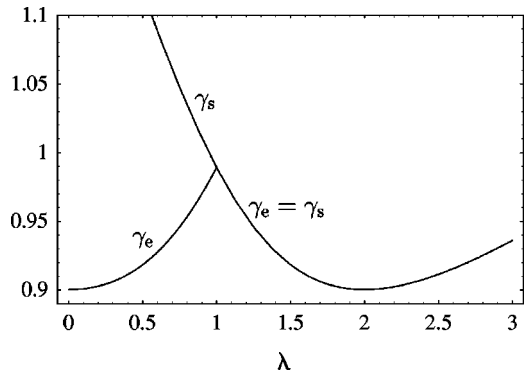

FIG. 1. Asymptotes of the soft perceptron learning rule. The generalization error of the ensemble decays as $\epsilon_e\sim\gamma_e/\alpha$, and for a single student $\epsilon_s\sim\gamma_s/\alpha$. The dependence of γ_e and γ_s on the parameter λ that controls the softness of the learning rule via $k\sim 1-(\lambda/\alpha)^2$, is shown in the plot. The learning rate schedule is $\eta\sim 2\sqrt{2\pi}/\alpha$. For all values of λ, this schedule optimizes both single student and ensemble performance. For $\lambda>1$ the students in the ensemble correlate quickly with increasing α, and using an ensemble asymptotically yields no improvement over single student performance.

It is of course not clear whether optimizing single student performance is a good idea, and we thus analyze more generic schedules, setting $k\sim 1-(\lambda/\alpha)^2$. Figure 1 then, however, shows that the two cases considered above are optimal for ensemble and, respectively, single student performance.

The above analysis of the soft perceptron rule suggests that while for some rules using an ensemble does significantly improve on single student performance, for more optimized rules this may no longer be the case. We shall now prove that the generalization error of the optimal single student learning rule is also a lower bound of the ensemble performance for any learning rule f. To achieve this, a learning rule $\tilde{f}$ will be given that for each pattern yields the ensemble average of f. Then a single student $\tilde{J}^\mu$ using $\tilde{f}$ will have generalization behavior equal to that of a large ensemble of students using f. The dynamics for $\tilde{J}^\mu$ may be written as

$$\tilde{J}^{\mu+1}=\tilde{J}^\mu+\xi^\mu N^{-1}\tilde{f}(\mu/N,B^T\xi^\mu,\tilde{J}^{\mu T}\xi^\mu), \tag{7}$$

where $\tilde{f}$ is the following integral transform of f:

$$\tilde{f}(\alpha,y,\tilde{x})=\langle f[\alpha,P(\alpha)^{1/2},y,\tilde{x}+(P(\alpha)-Q(\alpha))^{(1/2)}z]\rangle_z. \tag{8}$$

Here the distribution of z is normal. The entire procedure is quite intuitive: $\tilde{J}^\mu$ represents the center of mass of the ensemble and $\tilde{J}^{\mu T}\xi^\mu+(P(\alpha)-Q(\alpha))^{(1/2)}z$ is a guess for the value of the hidden field $J_i^{\mu T}\xi^\mu$ of one of the ensemble members. For large K the distribution of the last two quantities will be the same, and the ensemble average of f can be reliably predicted. Further, note that the class of soft perceptron rules (5) is invariant under the integral transform (8) since

$\langle H(a+bz)\rangle_z = H(a/\sqrt{1+b^2})$. This explains why optimizing single student and optimizing ensemble performance within this class yields the same generalization behavior.

To demonstrate that $\tilde{J}^\mu$ does indeed emulate the large ensemble, consider the order parameters $\tilde{R}(\alpha)=B^T\tilde{J}^{\alpha N}$ and $\tilde{Q}(\alpha)=|\tilde{J}^{\alpha N}|^2$. We shall start with $\tilde{J}^0=0$; thus $\tilde{R}(0)=R(0)=\tilde{Q}(0)=Q(0)=0$, and it will suffice to show that the pair $\tilde{R},\tilde{Q}$ satisfies an identical differential equation as the pair R,Q. From Eq. (2) we obtain for Q:

$$\dot{Q}=\langle 2x_i f(\alpha,P(\alpha)^{1/2},y,x_j) + f(\alpha,P(\alpha)^{1/2},y,x_i) f(\alpha,P(\alpha)^{1/2},y,x_j)\rangle_{y,x_i,x_j}, \quad (9)$$

where i and j are any two different indices. The Gaussians x_i and x_j may be rewritten in terms of normal random variables z_i, z_j and z, independent of each other and of y, as

$$x_i=\sqrt{P-Q}z_i+\sqrt{Q-R^2}z+Ry,$$
$$x_j=\sqrt{P-Q}z_j+\sqrt{Q-R^2}z+Ry. \quad (10)$$

Carrying out the integrations over z_i and z_j in Eq. (9) yields $\dot{Q}=\langle 2\tilde{x}\tilde{f}(\alpha,y,\tilde{x})+\tilde{f}(\alpha,y,\tilde{x})^2\rangle_{y,z}$, where $\tilde{x}\equiv\sqrt{Q-R^2}z+Ry$. The variance of $\tilde{x}$ is Q and its covariance with y is R. Applying Eq. (2) to $\tilde{J}^\mu$ yields $\dot{\tilde{Q}}=\langle 2\tilde{x}\tilde{f}(\alpha,y,\tilde{x})+\tilde{f}(\alpha,y,\tilde{x})^2\rangle_{y,\tilde{x}}$, where the variance of $\tilde{x}$ is $\tilde{Q}$ and its covariance with y is $\tilde{R}$. Thus Q and $\tilde{Q}$ satisfy the same differential equation and an analogous argument shows that the same holds for R and $\tilde{R}$.

Next consider more general architectures than the simple perceptron. It is easy to generalize the construction to the case of a tree committee machine: one just has to carry out an integration analogous to Eq. (8) for each branch of the tree. The case of the tree parity machine, however, is more involved since due to a gauge symmetry, students with differing weight vectors can implement the same function. Thus averaging the output of the ensemble members [Eq. (3)] may no longer be equivalent to averaging the weight vectors. Due to the permutation symmetry of the hidden units, the analogous problem arises in fully connected committee machines. But in both cases it is straightforward to break the symmetry in a formal way by adding a small deterministic drift term of the form $B^{(j)}\delta N^{-1}$ to the update equations (1) of each hidden unit j. Then for $\delta>0$ the same procedure as for the tree committee will yield an equivalent single student rule. In the end, one will of course want to take the limit $\delta\to 0$. This limit, however, in the case of a training set size that is on the order of the number of free parameters in a single student, yields pathological behavior on the single student level: For the parity machine only trivial generalization will result [8], and for the connected committee each student is stuck in a badly generalizing unspecialized state where only the mean of the weight vectors of the teacher is learned [6]. Thus for the parity machine an ensemble of such students will show no generalization, and for the connected committee it will show no specialization.

So the above procedure does not allow us to make any statement about the equivalence between ensemble and single student performance for the large training sets needed to achieve good generalization. It does, however, show that the pathological divergence of the training times, which results from the symmetries, cannot be overcome by the use of an ensemble.

Let us next consider whether there is a better way to encourage diversity in the ensemble than just choosing different initial conditions. The most promising approach seems to be to randomize the presentation order of the examples. So while all of the students use the same training set, they sample independently from this set. Now even if the sampling is done without replacement, the joint field distribution will no longer be Gaussian since correlations between students and inputs arise when calculating the overlap between different ensemble members. So it is far from clear how to emulate the ensemble behavior by using a single student. But for exactly the same reason the analysis of this scenario is very involved and will require the techniques (and approximations) used in the study of online learning when the training set is sampled with replacement [9,10]. Further, from a practical point of view, to permute the presentation order, the entire training set has to be stored somewhere, and one might ask whether such an algorithm should still be regarded as being online. Consequently, randomizing the order of presentation makes much more sense in the case where the training set is sampled with replacement.

The above results for the standard online scenario are entirely analogous to the findings in the case of batch learning. There, in a Bayesian framework, if one uses the suboptimal strategy of simply sampling the posterior, generalization performance is improved by using an ensemble and the large ensemble yields Bayes optimal performance [4]. However, it is possible to construct an optimal potential and also have the single student minimizing this potential display Bayes optimal generalization [11]. While in the online learning problems considered in this paper Bayes optimal generalization cannot be achieved with a single student, as in the batch case, an ensemble cannot improve on optimal single student performance. This is somewhat surprising since the key problem in online learning is to find a reasonable update rule that compresses the information about the teacher into the N-dimensional state vector of the student while sequentially scanning the training set. But when using an ensemble, a larger KN dimensional state space is available for storing the information. So the equivalence between optimal ensemble and single student performance shows that an ensemble is not a efficient way of using this larger state space. This raises the interesting question whether there are more efficient strategies of utilizing a large state space.

It is a pleasure to acknowledge helpful discussions with Manfred Opper and David Saad.

[1] *On-line Learning in Neural Networks*, edited by D. Saad (Cambridge University Press, Cambridge, 1998).
[2] G. Reents and R. Urbanczik, Phys. Rev. Lett. **80**, 5445 (1998).
[3] M. Opper, Phys. Rev. Lett. **77**, 4671 (1996).
[4] M. Opper and D. Haussler, Phys. Rev. Lett. **66**, 2677 (1991).
[5] O. Kinouchi and N. Caticha, J. Phys. A **25**, 6243 (1992).
[6] M. Biehl, P. Riegler, and C. Wöhler, J. Phys. A **29**, 4769 (1996).
[7] T. Watkin, Europhys. Lett. **21**, 871 (1993).
[8] R. Simonetti and N. Caticha, J. Phys. A **29**, 4859 (1996).
[9] A.C.C. Coolen and D. Saad (unpublished).
[10] A.C.C. Coolen and D. Saad (unpublished).
[11] O. Kinouchi and N. Caticha, Phys. Rev. E **54**, R54 (1996).