Statistical Mechanics of Ensemble Learning

Anders Krogh*
NORDITA, Blegdamsvej 17, 2100 Copenhagen, Denmark

Peter Sollich**
Department of Physics, University of Edinburgh, Kings Buildings, Mayfield Road, Edinburgh EH9 3JZ, U.K.

Physical Review E, 55:811–825, 1997; NORDITA preprint 96/24 S

Within the context of learning a rule from examples, we study the general characteristics of learning with ensembles. The generalization performance achieved by a simple model ensemble of linear students is calculated exactly in the thermodynamic limit of a large number of input components, and shows a surprisingly rich behaviour. Our main findings are: For learning in large ensembles, it is advantageous to use under-regularized students, which actually over-fit the training data. Globally optimal generalization performance can be obtained by choosing the training set sizes of the students optimally. For smaller ensembles, optimization of the ensemble weights can yield significant improvements in ensemble generalization performance, in particular if the individual students are subject to noise in the training process. Choosing students with a wide range of regularization parameters makes this improvement robust against changes in the unknown level of corruption of the training data.

PACS numbers: 87.10.+e, 05.90.+m, 02.50.Wp

I. INTRODUCTION

The methods of statistical mechanics have been successfully applied to the study of neural networks and other systems which can learn rules from examples (for reviews see, *e.g.*, Refs. [1,2]). The main issue is normally the question of *generalization*: Given a set of training examples, *i.e.*, pairs of inputs and corresponding outputs produced according to some underlying but unknown rule (the 'teacher' or 'target'), one wants to generate—by a suitable training algorithm—a predictor (the 'student') which generalizes, *i.e.*, makes accurate predictions for the outputs corresponding to inputs not contained in the training set.

More recently, it has emerged that generalization performance can often be improved by training not just one predictor, but rather using an ensemble, *i.e.*, a collection of a (finite) number of predictors, all trained for the same task. This idea of improving generalization performance by combining the predictions of many different predictors has been investigated extensively in statistics, see, *e.g.*, Refs. [3–5]. Within the context of neural network learning, ensembles have also been studied by several groups, see for instance Refs. [6–9]. Usually the predictors in the ensemble are trained independently and then their predictions are combined. This combination can be done by majority (in classification) or by simple averaging (in regression), but one can also use a *weighted* combination of the predictor. We focus on the latter method in the following. Other schemes for combining predictors exists, like mixtures of experts [10] where the weighting of the ensemble members is highly non-linear, and boosting [11,12] in which the training data are partitioned among the individual predictors in a way that optimizes the ensemble performance.

Ten copies of the same weather forecast obviously contain exactly the same amount of information as just one copy. By obtaining ten *different* forecasts, however, it may actually be possible to predict tomorrow's weather more accurately, even if the forecasts are all based on the same satellite data. The same is true quite generally for ensemble learning; only if the predictors in an ensemble are different is there something to be gained from using an ensemble. This obvious insight was quantified in Ref. [9] by a relation stating that the generalization error of a weighted combination of predictors in an ensemble is equal to the average error of the individual predictors minus the 'disagreement' among them, which we refer to as the ambiguity. For completeness, the derivation of this basic relation is reviewed in the next section. In Ref. [9], a combination of the ensemble idea and the method of cross-validation was also suggested. It is implemented by training each student only on a subset of the available data and 'holding out' the remaining examples for testing its performance. There are several reasons why this approach is useful. Firstly, one can obtain an *unbiased* estimate of the ensemble generalization error, even though the ensemble as a whole is trained on all available data. Secondly, by training the individual students on different subsets of the training data, they are made more 'diverse', and so it should be possible to reduce the ensemble error by increasing the ambiguity more than the errors of the individual students. Thirdly, the ambiguity can be estimated from the distribution of inputs (without the corresponding target outputs) alone, which can easily be sampled in many practical applications. By estimating the ambiguity accurately, the optimal weight for each student in the ensemble can then be determined to a similar degree of precision.

The method outlined above raises several interesting questions. Firstly, it would be interesting to see under which circumstances one can actually improve the ensemble generalization performance by training each student only on a subset of the available data. A second question is how large a fraction of the data set should be held out to obtain the lowest ensemble generalization error. Finally, one would like to know whether it is useful to have students differ, for example, in the amounts of regularization they use, or whether it is more advantageous to have an ensemble of identically regularized students. In this paper we investigate these and other questions quantitatively. By turning to the simplest of all models for the students, the linear perceptron, we obtain analytical results for the generalization performance of the ensemble as a function of noise in the data, noise in the training process, the amount of regularization on the students and, finally, the size of the training sets of the individual students and their overlaps. The behaviour that we find for this simple system is surprisingly rich and sufficiently non-trivial to allow general conclusions to be drawn. We believe that these conclusions will, at least to some extent, also hold for more complex, nonlinear learning systems.

For the case of an ensemble of unregularized linear students, two limiting cases of our analysis have previously been studied in Ref. [13]: The limit in which all the students are trained on the full data set and the one where all training sets are mutually non-overlapping. The main contribution of the present paper is that we are able to treat the case of intermediate training set sizes and overlaps exactly, yielding novel insights into ensemble learning. Furthermore, our analysis also allows us to study the effect of noise in the training algorithm, the influence of having different regularizations for the students in the ensemble, and the performance improvements that can be gained by optimizing the weights with which individual students contribute to the ensemble predictions. A short account of some of this work has appeared in Ref. [14].

II. GENERAL FEATURES OF ENSEMBLE LEARNING

A. Ensemble generalization error and ambiguity

Let us consider the task of predicting a rule (teacher) given by a target function f_0 mapping inputs $\mathbf{x} \in R^N$ to outputs $y \in R$. We assume that we can only obtain noisy samples of this mapping, and denote the resulting stochastic target function $y(\mathbf{x})$. Assume now that an ensemble of K independent predictors $f_k(\mathbf{x})$ of $y(\mathbf{x})$ is available. Weighted averages over this ensemble will be denoted by an overbar. The final output of the ensemble, for example, is given by

$$\overline{f}(\mathbf{x}) = \sum_k \omega_k f_k(\mathbf{x}).$$

We can think of weight ω_k as our belief in predictor k, and therefore constrain the weights to be positive and to sum to one.

We define the *ambiguity* on input $\mathbf{x}$ of a single member of the ensemble as $a_k(\mathbf{x}) = (f_k(\mathbf{x}) - \overline{f}(\mathbf{x}))^2$. The *ensemble ambiguity* on input $\mathbf{x}$

$$\overline{a}(\mathbf{x}) = \sum_k \omega_k a_k(\mathbf{x}) = \sum_k \omega_k (f_k(\mathbf{x}) - \overline{f}(\mathbf{x}))^2 \qquad (1)$$

quantifies the disagreement among the predictors on input $\mathbf{x}$; it is simply the variance of their outputs around the weighted ensemble mean. The quadratic errors of predictor k and of the ensemble are

$$\epsilon_k(\mathbf{x}) = (y(\mathbf{x}) - f_k(\mathbf{x}))^2$$
$$\epsilon(\mathbf{x}) = (y(\mathbf{x}) - \overline{f}(\mathbf{x}))^2$$

respectively. Adding and subtracting $y(\mathbf{x})$ in (1) yields after a few manipulations

$$\epsilon(\mathbf{x}) = \overline{\epsilon}(\mathbf{x}) - \overline{a}(\mathbf{x}), \qquad (2)$$

where it was used that the weights ω_k sum to one, and we have defined the average error of the individual predictors, $\overline{\epsilon}(\mathbf{x}) = \sum_k \omega_k \epsilon_k(\mathbf{x})$.

Let us now assume that the input $\mathbf{x}$ is sampled randomly from a probability distribution $P(\mathbf{x})$. The above formulas can be averaged over this distribution and the corresponding (stochastic) target outputs $y(\mathbf{x})$. If we define the average of $\epsilon(\mathbf{x})$ to be the *ensemble generalization error* ϵ, then we obtain by averaging (2)

$$\epsilon = \overline{\epsilon} - \overline{a}. \qquad (3)$$

The first term on the right is the weighted average of the generalization errors of the individual predictors ($\overline{\epsilon} = \sum_k \omega_k \epsilon_k$), while the second is the (average) ensemble ambiguity

$$\overline{a} = \sum_k \omega_k a_k = \sum_k \omega_k \left\langle (f_k(\mathbf{x}) - \overline{f}(\mathbf{x}))^2 \right\rangle_{\mathbf{x}}. \qquad (4)$$

The general relation (3), which has been previously derived in Ref. [9], shows clearly that the more the predictors differ, the lower the ensemble error will be, provided the individual errors remain constant. We want the predictors to disagree! Another important feature of eq. (3) is that it decomposes the generalization error into a term that depends only on the generalization errors of the individual predictors and another term that contains *all correlations* between the predictors. Furthermore, as eq. (4) shows, the correlation term $\overline{a}$ can be estimated entirely from *unlabeled data*, *i.e.*, no knowledge is required of the actual target function. The term 'unlabeled example' is borrowed from classification problems, and in this context it means an input $\mathbf{x}$ for which the value of the target output $y(\mathbf{x})$ is unknown.

Parenthetically, we note that our definition of the generalization error includes a contribution arising from the stochasticity of the target outputs alone (namely, the variance of $y(\mathbf{x})$, averaged over $\mathbf{x}$). Eq. (3) also holds when this irrelevant constant is dropped on both sides, and this is indeed what we shall do in our explicit calculations of the generalization error below. We also observe from eq. (3) that the generalization error of the ensemble is always smaller than the (weighted) average error of the individual predictors, $\epsilon < \bar{\epsilon}$. In particular for uniform weights

$$\epsilon \leq \frac{1}{K}\sum_k \epsilon_k$$

which has been noted by several authors, see, *e.g.*, Ref. [7].

B. Bias and variance

All our observations up to this point do not depend on how the predictors f_k are obtained. In the rest of this paper, we assume that the f_k are generated on the basis of a training set consisting of p examples of the target function, $(\mathbf{x}^\mu, y^\mu)$, $\mu = 1\ldots p$, where $y^\mu = f_0(\mathbf{x}^\mu) + \eta^\mu$ with η^μ being zero mean additive noise. In this context, it is natural to refer to the f_k as students and to focus on the *average* ensemble generalization error as the main quantity of interest. The average is taken over all training sets, *i.e.*, over all sets of training inputs $\mathbf{x}^\mu$, randomly and independently sampled from $P(\mathbf{x})$, and the corresponding noisy training outputs y^μ. Decomposing the ensemble output $\overline{f}(\mathbf{x})$ into its average over all training sets, $\langle\overline{f}(\mathbf{x})\rangle_{\mathbf{x}^\mu,y^\mu}$ and the deviation $\Delta\overline{f}(\mathbf{x})$ from this average, one can write the average ensemble generalization error as

$$\begin{aligned}
\langle\epsilon\rangle_{\mathbf{x}^\mu,y^\mu} &= \left\langle\left\langle\left(y(\mathbf{x})-\overline{f}(\mathbf{x})\right)^2\right\rangle_{\mathbf{x},y}\right\rangle_{\mathbf{x}^\mu,y^\mu} \\
&= \left\langle\left(y(\mathbf{x})-\langle\overline{f}(\mathbf{x})\rangle_{\mathbf{x}^\mu,y^\mu}\right)^2\right\rangle_{\mathbf{x},y} + \left\langle\langle(\Delta\overline{f}(\mathbf{x}))^2\rangle_{\mathbf{x}^\mu,y^\mu}\right\rangle_{\mathbf{x}} \\
&= \left\langle B^2(\mathbf{x},y)\right\rangle_{\mathbf{x},y} + \langle V(\mathbf{x})\rangle_{\mathbf{x}}\,. \qquad (5)
\end{aligned}$$

The first and second term on the r.h.s. of (5) are normally referred to as (squared) bias and variance of the ensemble output (both averaged over the test input $\mathbf{x}$ and test output y), respectively [15]. Since the bias of the ensemble

$$\begin{aligned}
B(\mathbf{x},y) &= y(\mathbf{x}) - \langle\overline{f}(\mathbf{x})\rangle_{\mathbf{x}^\mu,y^\mu} \\
&= \sum_k \omega_k\left(y(\mathbf{x}) - \langle f_k(\mathbf{x})\rangle_{\mathbf{x}^\mu,y^\mu}\right)
\end{aligned}$$

is simply the average of the biases of the individual students, ensemble learning cannot normally be expected to yield a significant reduction in bias compared to learning with a single student. The variance of the ensemble output, on the other hand, is given by

$$\begin{aligned}
V(\mathbf{x}) &= \left\langle\left(\sum_k \omega_k \Delta f_k(\mathbf{x})\right)^2\right\rangle_{\mathbf{x}^\mu,y^\mu} \\
&= \sum_{kl}\omega_k\omega_l\,\langle\Delta f_k(\mathbf{x})\Delta f_l(\mathbf{x})\rangle_{\mathbf{x}^\mu,y^\mu}\,.
\end{aligned}$$

It is upper bounded by the average of the variances of the individual students:

$$V(\mathbf{x}) \leq \sum_k \omega_k\left\langle(\Delta f_k(\mathbf{x}))^2\right\rangle_{\mathbf{x}^\mu,y^\mu}\,. \qquad (6)$$

This bound is saturated when the fluctuations of the student outputs (as functions of hypothetical 'fluctuations' in the training set) are fully correlated and of equal variance, confirming again the intuition that the benefit of ensemble learning is small if all students are identical in the sense that they react very similarly to different training sets. In the opposite case where the fluctuations of the students are completely uncorrelated, one has

$$V(\mathbf{x}) = \sum_k \omega_k^2\left\langle(\Delta f_k(\mathbf{x}))^2\right\rangle_{\mathbf{x}^\mu,y^\mu} \qquad (7)$$

which for approximately uniform ensemble weights ($\omega_k \approx 1/K$) is significantly lower (by a factor of $O(1/K)$) than the average (6) of the individual variances. We expect, therefore, that ensemble learning is most useful in circumstances where the generalization errors of the individual students are dominated by variance rather than bias. This expectation will be confirmed by our results for a simple model system, to be described in the following sections.

C. Training on subsets

As pointed out in the introduction, the students in the ensemble need not be trained on all available training data. In fact, since training on different examples will generally increase the ambiguity, it is possible that training on subsets of the data will *improve* generalization performance. An additional advantage is that, by holding out a different part of the total data set for the purpose of testing each student, one can use the whole data set for training the ensemble and still get an unbiased estimate of the ensemble generalization error. Denoting this estimate by $\hat{\epsilon}$, one has simply

$$\hat{\epsilon} = \overline{\epsilon^{\text{test}}} - \hat{\overline{a}} \qquad (8)$$

where $\overline{\epsilon^{\text{test}}} = \sum_k \omega_k \epsilon_k^{\text{test}}$ is the average of the students' test errors, and $\hat{\overline{a}}$ is an estimate of the ensemble ambiguity, obtained from unlabeled examples as explained above [16].

So far, we have not mentioned how to find the ensemble weights ω_k. Often uniform weights $\omega_k = 1/K$ are used, but it is tempting to optimize the weights in some way. In Refs. [7,8], the training set was used to perform the optimization, *i.e.*, the weights were chosen to minimize the ensemble training error. This can easily lead to substantial over-fitting, as we shall show below. It has therefore been suggested [9] to minimize the estimated generalization error (8) instead. If this is done, the estimate (8), evaluated at the optimized weights, is of course no longer unbiased; intuitively, however, we expect the resulting bias to be small for large ensembles. A quantitative analysis of this point is beyond the scope of our present analysis, since the fluctuations of the test errors around the corresponding generalization errors vanish in the thermodynamic limit considered below, making the estimate (8) not only unbiased but in fact exact. Note that since both the ensemble training error and the ensemble generalization error involve only terms linear and quadratic in the ensemble output (and hence in the ensemble weights ω_k), finding the corresponding optimal ω_k is simply a quadratic optimization problem, made nontrivial only by the constraints that the weights should be positive and sum to one.

III. ENSEMBLES OF LINEAR STUDENTS

A. Linear perceptron learning

In preparation for our analysis of learning with ensembles of linear students we now briefly review the case of a single linear student, also referred to as a 'linear perceptron'. Such a student implements the input-output mapping

$$f(\mathbf{x}) = \frac{1}{\sqrt{N}}\mathbf{w}^{\mathrm{T}}\mathbf{x}$$

parametrized in terms of an N-dimensional parameter vector $\mathbf{w}$ with real components; the scaling factor $1/\sqrt{N}$ is introduced for convenience, and $^{\mathrm{T}}$ denotes the transpose of a vector. The student's parameter vector $\mathbf{w}$ should of course not be confused with the ensemble weights ω_k. The most common method for training such a linear student (or parametric inference models in general) is minimization of the sum-of-squares training error

$$E^{\mathrm{t}} = \sum_{\mu}(y^{\mu} - f(\mathbf{x}^{\mu}))^2$$

where $\mu = 1 \ldots p$ numbers the training examples. To prevent the student from fitting noise in the training data, a weight decay term is often added, and one minimizes the energy function

$$E = E^{\mathrm{t}} + \lambda \mathbf{w}^2 \qquad (9)$$

instead. The size of the weight decay parameter λ determines how strongly large parameter vectors are penalized; large λ corresponds to a stronger *regularization* of the student. Within a Bayesian framework, λ can also be viewed as implementing prior knowledge about the type of task to be learned (see, *e.g.*, Refs. [17–19]). Finally, λ can loosely be interpreted as a soft constraint on the complexity of the mapping that the linear student can implement. In the context of learning with multi-layer feedforward networks, for example, large λ would thus correspond to a high cost for adding new hidden units, so that simple networks with few hidden units would be preferred.

In practice, the minimum of the energy function E is often located by gradient descent. For the linear student E is a quadratic function of the parameter vector $\mathbf{w}$, and therefore this procedure will necessarily find the global minimum of E. However, for more realistic, nonlinearly parametrized students, this will not necessarily be the case, and one may often end up in a local minimum of E. We crudely model the corresponding randomness in the training process by considering white noise added to the gradient descent updates of the parameter vector $\mathbf{w}$. In a continuous learning time approximation, $\mathbf{w}$ then obeys a Langevin equation which for large learning times leads to a Gibbs distribution of parameter vectors [20]. This distribution can be written as $P(\mathbf{w}) \propto \exp(-E/2T)$, where the 'temperature' T measures the amount of noise in the learning process [21]. We focus our analysis on the thermodynamic limit $N \to \infty$ at constant normalized number of training examples, $\alpha = p/N$. In this limit, quantities such as the training or generalization error become self-averaging, *i.e.*, their averages over all training sets become identical to their typical values for a particular training set. For linear students, it was shown in Ref. [22] that in general the exact average case results obtained in the thermodynamic limit are good approximations even for system sizes N as small as a few tens or hundreds, and we expect the same to hold for our analysis of ensemble learning with linear students.

Let us assume that the training inputs $\mathbf{x}^{\mu}$ are chosen randomly and independently from a Gaussian distribution $P(\mathbf{x}) \propto \exp(-\frac{1}{2}\mathbf{x}^2)$, and that training outputs are generated by a linear target function corrupted by noise, $y^{\mu} = \mathbf{w}_0^{\mathrm{T}}\mathbf{x}^{\mu}/\sqrt{N} + \eta^{\mu}$ where η^{μ} is zero mean additive noise with variance σ^2. Fixing the length of the target parameter vector to $\mathbf{w}_0^2 = N$ for simplicity, the resulting generalization error of a linear student with weight decay λ and learning noise T can be written as [23]

$$\epsilon = (\sigma^2 + T)G + \lambda(\sigma^2 - \lambda)\frac{\partial G}{\partial \lambda}. \qquad (10)$$

On the r.h.s. of this equation we have dropped the term arising from the noise on the target function alone, which is simply σ^2; this convention will be followed throughout. The 'response function' G is defined as $G = \frac{1}{N}\mathrm{tr}\,\langle \mathbf{g} \rangle$, where $\mathbf{g}^{-1}$ is half the Hessian of the energy function E

defined in (9) and $\langle\cdot\rangle$ is an average over the training inputs $\mathbf{x}^\mu$. Explicitly, g can be expressed as

$$\mathrm{g}^{-1} = \lambda\mathbf{1} + \mathbf{A} \tag{11}$$

where $\mathbf{1}$ is the $N \times N$ unit matrix and

$$\mathbf{A} = \frac{1}{N}\sum_\mu \mathbf{x}^\mu(\mathbf{x}^\mu)^{\mathrm{T}} \tag{12}$$

is the correlation matrix of the training inputs. The response function can be calculated as the physically relevant solution of the equation [22,24]

$$1/G = \alpha/(1+G) + \lambda \tag{13}$$

which leads to

$$G = G(\alpha,\lambda) = \frac{1}{2\lambda}\left[1 - \alpha - \lambda + \sqrt{(1-\alpha-\lambda)^2 + 4\lambda}\right]. \tag{14}$$

An equation exactly analogous to (10) also holds when the training examples are generated by a noisy nonlinear perceptron target function. In this case σ^2 is replaced by an effective noise level which is the sum of the actual noise variance and the error of the best linear fit to the transfer function of the nonlinear target perceptron [13,25,26].

We conclude our review of learning with a single linear student by remarking that for any given number of training examples α and zero learning noise $T = 0$, the generalization error (10) is minimized when the weight decay is set to the value $\lambda = \sigma^2$ [23]. Assuming that the noise on the training outputs and the prior probability of teacher parameter vectors $\mathbf{w}_0$ are Gaussian, this corresponds to optimal learning in the sense of Ref. [27] and also to the Bayes optimal estimator (see, *e.g.*, Refs. [28,29]). The minimal value of the generalization error thus obtained is $\sigma^2 G(\alpha,\sigma^2)$. For $\lambda < \sigma^2$, the student is *under-regularized* and will therefore tend to *over-fit* noise in the training data; for $\lambda > \sigma^2$, on the other hand, *over-regularization* forces the student to fit the data less closely and puts more emphasis on the preference for short parameter vectors $\mathbf{w}$, as expressed in the weight decay term of the energy function (9). In terms of the bias-variance decomposition of the generalization error discussed in the previous section, under-regularization corresponds to small bias but large variance, since the student's predictions depend strongly on noise in the training data. For over-regularized students, on the other hand, the variance is small but the suboptimally large value of λ leads to a large bias. This difference between under- and over-regularization will help us understand the resulting ensemble performance, as discussed in more detail in Section IV.

B. Ensemble generalization error

We now consider an ensemble of K linear students with weight decays λ_k and learning noises T_k ($k = 1 \ldots K$). Each student has an ensemble weight ω_k and is trained on $N\alpha_k$ training examples, with students k and l sharing $N\alpha_{kl}$ training examples. As above, we consider noisy training data generated by a linear teacher (or a nonlinear perceptron teacher with effective noise variance σ^2). Details of the calculation of the resulting ensemble generalization error are relegated to Appendix A; in Appendix B, we show how the relevant averages over training inputs can be calculated using either diagrammatic methods [24] or differential equations derived from matrix identities [22]. The resulting ensemble generalization error is:

$$\epsilon = \sum_{kl} \omega_k\omega_l\epsilon_{kl} \tag{15}$$

where

$$\epsilon_{kl} = \frac{\rho_k\rho_l + \sigma^2(1-\rho_k)(1-\rho_l)\alpha_{kl}/(\alpha_k\alpha_l)}{1-(1-\rho_k)(1-\rho_l)\alpha_{kl}/(\alpha_k\alpha_l)} + \delta_{kl}T_kG_k. \tag{16}$$

Here G_k is defined as $G_k = G(\alpha_k,\lambda_k)$, and $\rho_k = \lambda_k G_k$. Rewriting the definition of ρ_k as $\rho_k = \left\langle \frac{1}{N}\,\mathrm{tr}\,\lambda_k(\lambda_k\mathbf{1} + \mathbf{A}_k)^{-1}\right\rangle$, where $\mathbf{A}_k$ is the correlation matrix of the training inputs on which student k is trained, ρ_k can be interpreted as the fraction of the N parameters of student k which are not well-determined by its training data (but rather by the weight decay regularization) [30]. The Kronecker delta in the last term of (16) arises because the learning noises for different students are uncorrelated. The generalization error and ambiguity of the individual students are

$$\epsilon_k = \epsilon_{kk} \qquad a_k = \epsilon_{kk} - 2\sum_l \omega_l\epsilon_{kl} + \sum_{lm}\omega_l\omega_m\epsilon_{lm}.$$

From these expressions one can again verify the general relation (3). In Sections IV and V, we shall explore the consequences of the general result (15, 16) first for the limit of a large ensemble $K \to \infty$, and then for more realistic ensemble sizes. We will concentrate on the case where the training set of each student is sampled randomly (without replacement) from the total available data set of size $N\alpha$. For the overlap of the training sets of students k and l ($k \neq l$) one then has $\alpha_{kl}/\alpha = (\alpha_k/\alpha)(\alpha_l/\alpha)$ up to fluctuations which vanish in the thermodynamic limit, and hence

$$\alpha_{kl} = \alpha_k\alpha_l/\alpha. \tag{17}$$

For finite ensembles one can construct training sets for which $\alpha_{kl} < \alpha_k\alpha_l/\alpha$. This results in a slightly smaller generalization error, but for simplicity we use (17).

C. Ensemble training error

We now give the analogue of the result (15, 16) for the error of the ensemble predictions on the *training set*. It has been suggested [7,8] that the ensemble weights ω_k should be chosen such that this so-called *ensemble training error* is minimized, which motivates our interest in this quantity. Since the ensemble training error is not an unbiased estimate of the generalization error, choosing the ensemble weights to minimize it may well lead to over-fitting. However, when some examples are held out for testing each student, the ensemble error on the training set contains contributions from both training and test errors of the individual students. (This shows that the term 'ensemble training error' is actually a slight misnomer in this context.) The test errors estimate the corresponding generalization errors without bias, and one would therefore expect the degradation of generalization performance from minimizing the ensemble training error rather than the estimated generalization error (8) to be relatively benign, as long as the test sets for the individual students are not too small.

The calculation of the ensemble training error, which is detailed in Appendices A and B, yields the result

$$\epsilon^{\mathrm{t}} = \left\langle \frac{1}{p} \sum_{\mu} \left(y^{\mu} - \sum_{k} \omega_k \bar{f}(\mathbf{x}^{\mu}) \right)^2 \right\rangle = \sum_{kl} \omega_k \omega_l \epsilon^{\mathrm{t}}_{kl}. \tag{18}$$

In the absence of learning noise (all $T_k = 0$), the $\epsilon^{\mathrm{t}}_{kl}$ are related to the corresponding coefficients ϵ_{kl} in the result for the ensemble generalization error (16) by

$$\epsilon^{\mathrm{t}}_{kl}\big|_{T_k=0} = \left(\epsilon_{kl}\big|_{T_k=0} + \sigma^2 \right) \times \left\{ 1 - \frac{1}{\alpha} \left[2 - \rho_k - \rho_l - (1-\rho_k)(1-\rho_l) \frac{\alpha_{kl}}{\alpha_k \alpha_l} \right] \right\} \tag{19}$$

Since the students can fit noise in the training data, the ensemble training error can of course be smaller than σ^2, and we have therefore retained the contribution from noise on the training examples in (19). This is why the ensemble training error is related to the ensemble generalization error *including* noise on the test examples, $\epsilon + \sigma^2$. Eq. (19) shows that, as expected, the training error is always smaller than the (noisy) generalization error. The same is also true in the presence of learning noise ($T_k > 0$), where one has

$$\epsilon^{\mathrm{t}}_{kl} = \epsilon^{\mathrm{t}}_{kl}\big|_{T_k=0} + \delta_{kl} T_k \left[\frac{\alpha_k}{\alpha} \frac{G_k}{1+G_k} + \left(1 - \frac{\alpha_k}{\alpha}\right) G_k \right] \tag{20}$$

compared to $\delta_{kl} T_k G_k$ for the generalization error. For the diagonal terms in (19, 20) one can show the intuitively reasonable result

$$\epsilon^{\mathrm{t}}_{kk} = \frac{\alpha_k}{\alpha} \epsilon^{\mathrm{t}}_k + \left(1 - \frac{\alpha_k}{\alpha}\right)(\epsilon_k + \sigma^2)$$

where ϵ^{t}_k and ϵ_k are the training and generalization errors of student k. This shows explicitly that the ensemble training error is a mixture of training and generalization errors.

IV. LARGE ENSEMBLE LIMIT

We now use our main result (15) to analyse the generalization performance of an ensemble with a large number K of students, in particular when the size of the training sets for the individual students are chosen optimally. If the ensemble weights ω_k are approximately uniform ($\omega_k \approx 1/K$), the ensemble generalization error is dominated by the off-diagonal elements of the matrix (ϵ_{kl}) in the limit of a large ensemble, $K \to \infty$. The diagonal elements can therefore be replaced with the corresponding expressions for the off-diagonal elements, yielding together with (17)

$$\epsilon \approx \sum_{kl} \omega_k \omega_l \frac{\rho_k \rho_l + \sigma^2 (1-\rho_k)(1-\rho_l)/\alpha}{1 - (1-\rho_k)(1-\rho_l)/\alpha}. \tag{21}$$

For the special case where all students are identical and are trained on training sets of identical size $\alpha_k = (1-c)\alpha$, we show the resulting ensemble generalization error in Figure 1A. The minimum at a nonzero value of c, which is the fraction of the total data set held out for testing each student, can clearly be seen. This confirms our intuition that when the students are trained on smaller, less overlapping training sets, the increase of the errors of the individual students can be more than offset by the corresponding increase in ambiguity.

The optimal training set sizes α_k can in fact be calculated analytically. Setting the derivatives of the generalization error (21) w.r.t. α_k to zero, one obtains the conditions

$$\rho_k = \lambda_k G_k = \sigma^2 G(\alpha, \sigma^2) \equiv \rho \qquad (k = 1 \ldots K). \tag{22}$$

Using (13), the solution for the optimal training set sizes (c_k denotes the fraction of the total data set used for testing student k) is obtained as

$$c_k \equiv 1 - \frac{\alpha_k}{\alpha} = \frac{1 - \lambda_k/\sigma^2}{1 + G(\alpha, \sigma^2)}. \tag{23}$$

The corresponding generalization error is simply $\epsilon = \rho + O(1/K)$, which as explained in Section III A is the minimal generalization error that can be obtained. We can thus conclude that *a large ensemble with optimally chosen training set sizes can achieve globally optimal generalization performance*. However, we see from (23) that, since $c_k \geq 0$ by definition, optimal generalization performance can only be obtained by choosing optimal training set sizes if all the weight decays λ_k are smaller than σ^2,

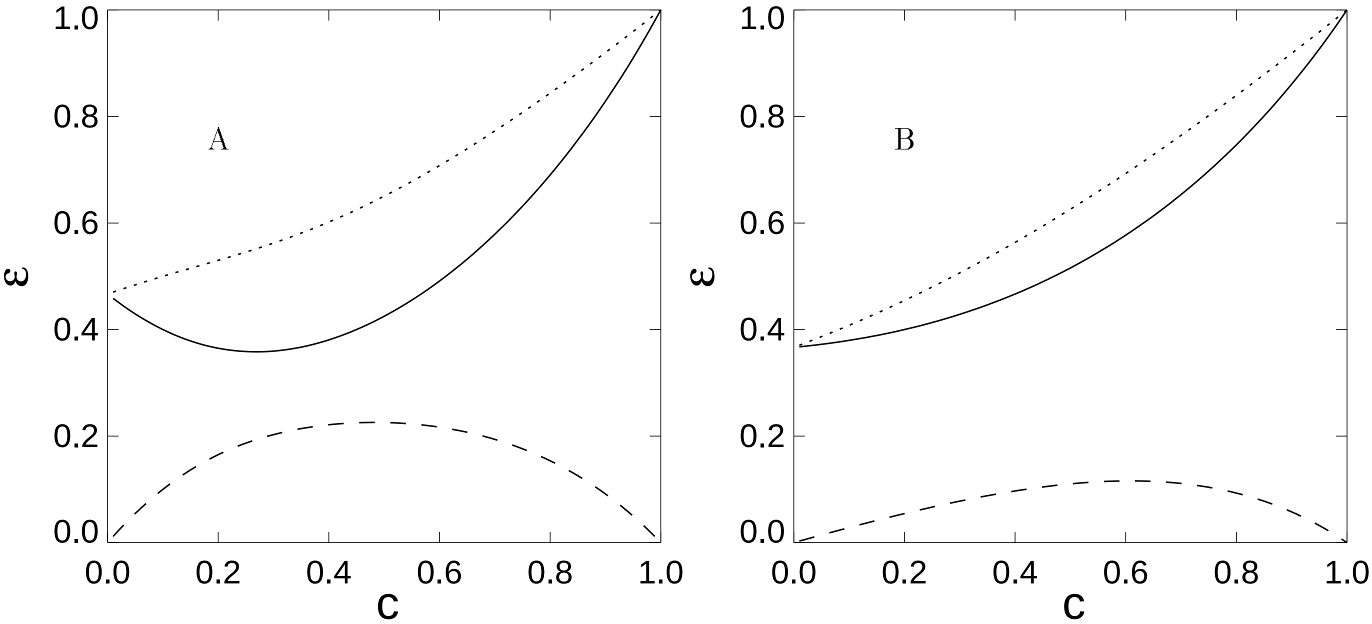

FIG. 1. Generalization errors and ambiguity for an infinite ensemble of identical students. The solid line is the ensemble generalization error, ϵ, the dotted line shows the error of the individual students, $\overline{\epsilon}$, and the ambiguity $\overline{a}$ is represented by the dashed line. Plot A shows the case of under-regularized students ($\lambda = 0.05$, $\sigma^2 = 0.2$). Note that there is an optimal c for which the generalization error of the ensemble has a minimum. This minimum exists whenever $\lambda < \sigma^2$. When the students are over-regularized as in plot B ($\lambda = 0.3$ at the same noise level $\sigma^2 = 0.2$), the minimum disappears. For both plots $\alpha = 1$.

i.e., if the ensemble is under-regularized. This is exemplified, again for an ensemble of identical students, in Figure 1B which shows that for an over-regularized ensemble, the generalization error is a monotonic function of c, and never reaches the minimum generalization error. These results confirm our expectation that ensemble learning is most useful for reducing variance: The generalization error of under-regularized students is dominated by variance contributions which, as shown in Section II, can be significantly reduced by decorrelating the student outputs. This is achieved by training the students on non-identical training sets with small overlap, and in this way optimal generalization performance can be achieved (for optimal c). For over-regularized students, on the other hand, the generalization error is dominated by bias. Only the remaining small variance contribution can be reduced by using an ensemble, making it impossible to reach optimal performance.

The general conclusion which we draw from the above results is that *ensemble learning is most useful if the individual students are not already strongly regularized.* This means that for ensemble learning, over-fitting can actually have a positive effect by allowing full exploitation of the ensemble's potential for reducing variance. Using the correspondence between regularization and prior knowledge, we can also say that ensemble learning really comes into its own when only little prior knowledge about the task to be learned is available, which would normally lead to strong over-fitting when using a single student. Note that the large ensemble generalization error (21) has no contribution from the learning noise of the individual students. This property of ensemble learning, namely, the suppression of inherent randomness in the training process, will be explored in more detail in the next section.

An interesting consequence of (23) is that in order to obtain optimal generalization performance, more strongly regularized students should be trained on a larger fraction of the total data set. Using (22), this can also be interpreted in the sense that all students should have the same number of parameters which are well-determined by their respective training sets. This makes sense since one expects that in this case the fluctuations of all students caused by the randomness of the training examples will be of the same order, thus maximizing the overall ambiguity.

We now discuss the finite K corrections to the generalization error resulting from the (large K-optimal) choice (23) for the training set sizes, assuming that the ensemble is under-regularized, *i.e.*, $\lambda_k \leq \sigma^2$ for all k. For uniform weights ($\omega_k = 1/K$) one has $\sum_k \omega_k^2 = 1/K$, and in the general case we therefore define an effective ensemble size by $1/K_{\text{eff}} = \sum_k \omega_k^2$. Using (22) and (13), the ensemble generalization error can then be written in the form

$$\epsilon = \rho + \rho \sum_k \omega_k^2 \left[(1-\rho) \frac{\sigma^2 - \lambda_k}{\rho^2 + \lambda_k} + \frac{T_k}{\lambda_k} \right]$$

and bounded by

$$\epsilon \leq \rho \left[1 + \frac{1}{K_{\rm eff}} \left(\frac{\sigma^2 + T_{\rm max}}{\lambda_{\rm min}} - 1\right)\right]$$

where $\lambda_{\rm min}$ and $T_{\rm max}$ are the minimal weight decay and the maximal learning noise in the ensemble, respectively. The ensemble is thus large in the sense that optimal generalization performance can be achieved by tuning the training set sizes if

$$K_{\rm eff} \gg \left|\frac{\sigma^2 + T_{\rm max}}{\lambda_{\rm min}} - 1\right|.$$

This means that although it is useful not to over-regularize the students in the ensemble, one should definitely utilize whatever prior knowledge is available to provide some minimal regularization (corresponding to a nonzero value of $\lambda_{\rm min}$). Otherwise, prohibitively large ensemble sizes will be needed to achieve good generalization performance.

We conclude this section by discussing how the adaptation of the training set sizes could be performed in practice, confining ourselves to an ensemble of identically regularized students for simplicity, where only one parameter $c = c_k = 1 - \alpha_k/\alpha$ has to be adapted. If the ensemble is under-regularized one expects that the generalization error will have a minimum for some nonzero c as in Figure 1A. One could, therefore, start by training all students on a large fraction of the total data set (corresponding to $c \approx 0$), and then gradually and randomly remove training examples from the students' training sets. For each training set size, one could estimate the generalization error by the performance of the students on the examples on which they have not been trained according to eq. (8), and one would stop removing training examples when the generalization error stops decreasing. The resulting estimate of the generalization error will be slightly biased; however, it would seem that for a large enough ensemble and due to the random selection of training examples, the risk of obtaining a strongly biased estimate by, for example, systematically testing all students on too 'easy' training examples is rather small.

V. REALISTIC ENSEMBLE SIZES

We now discuss some new effects occurring in ensembles with 'realistic' numbers of students, which were not covered by the discussion of the large ensemble limit in the preceding section,

A. Effect of learning noise

We have seen that in an over-regularized ensemble, nothing can be gained by making the students more 'diverse' by training them on smaller, less overlapping training sets. One would also expect this kind of 'diversification' to be unnecessary or even counter-productive when the learning noise is high enough to provide sufficient inherent diversity of students. In the large ensemble limit, we saw that this effect is suppressed, but it does indeed occur for realistically sized ensembles. In Figure 2, we show the dependence of the ensemble generalization error ϵ on $c = 1-\alpha_k/\alpha$ for an ensemble of $K = 10$ identical, under-regularized students. For small learning noise T, the minimum of ϵ at nonzero c persists, whereas for larger T, ϵ is monotonically increasing with c, implying that further diversification of students beyond that caused by the learning noise is wasteful. The plot also shows the performance of the optimal single student (with λ chosen to minimize the generalization error at the given T), demonstrating that the ensemble can perform significantly better by effectively averaging out learning noise.

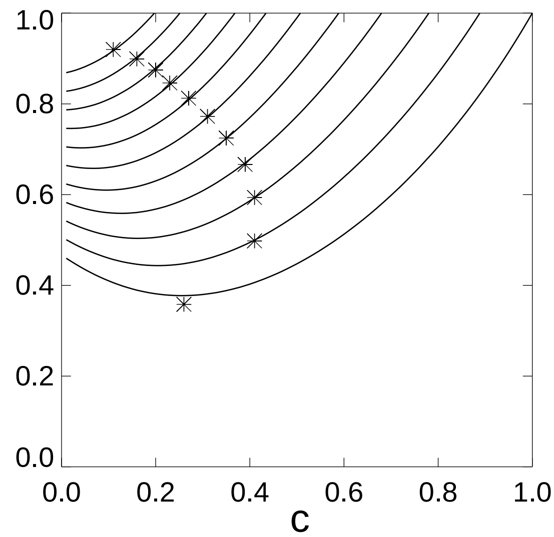

FIG. 2. Generalization error of an ensemble with 10 identical students as a function of the test set fraction c, for various values of the learning noise T. From bottom to top the curves correspond to $T = 0, 0.1, 0.2 \ldots 1.0$. The stars show the error $\epsilon_0(T)$ of the optimal (w.r.t. the choice of weight decay) single perceptron trained on all the examples, which is independent of c. They are placed where the ensemble error is identical to $\epsilon_0(T)$. For $T = 0$, $\epsilon_0(T)$ is always lower than the ensemble error, as shown by the lowest star. The parameters for this example are: $\alpha = 1$, $\lambda = 0.05$, $\sigma^2 = 0.2$.

B. Weight optimization

For realistic ensemble sizes, we have just seen that the presence of learning noise generally reduces the potential for performance improvement by choosing optimal *training set sizes*—the inherently noisy, diverse students should each be trained on a large part of the total data set, the size of the test set being just sufficient to estimate the generalization error reliably. In such cases, however, one can still adapt the *ensemble weights* ω_k to

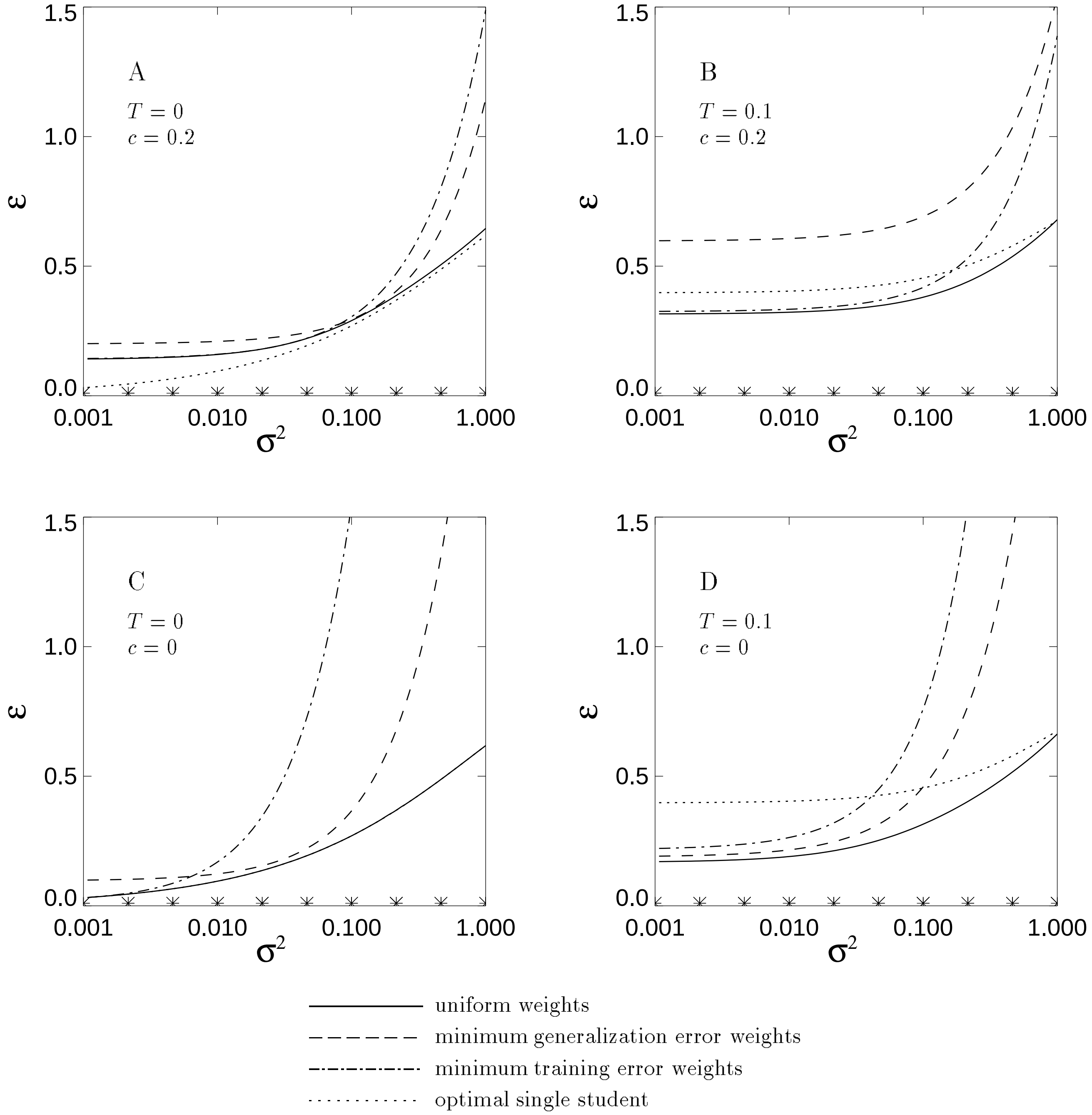

FIG. 3. Generalization error ϵ of an ensemble with 10 students with different weight decays, shown as a function of the noise σ^2. The weight decays of the students are marked by stars on the (logarithmic) x-axis. The dashed lines are for the uniformly weighted ensemble ($\omega_k = 1/K$), and the solid line is for ensemble weights chosen to minimize the ensemble *generalization* error. The dot-dashed lines show the generalization error obtained when the ensemble weights are found instead by minimizing the ensemble *training* error. The dotted lines, finally, are for the optimal single student trained on all data. All the plots are for $\alpha = 1$; the values of the learning noise T and the test set fraction c are shown in the individual plots. Note that in plot C ($T = 0$, $c = 0$), the error that the optimally weighted ensemble achieves is indistinguishable from the error of the single optimal network.

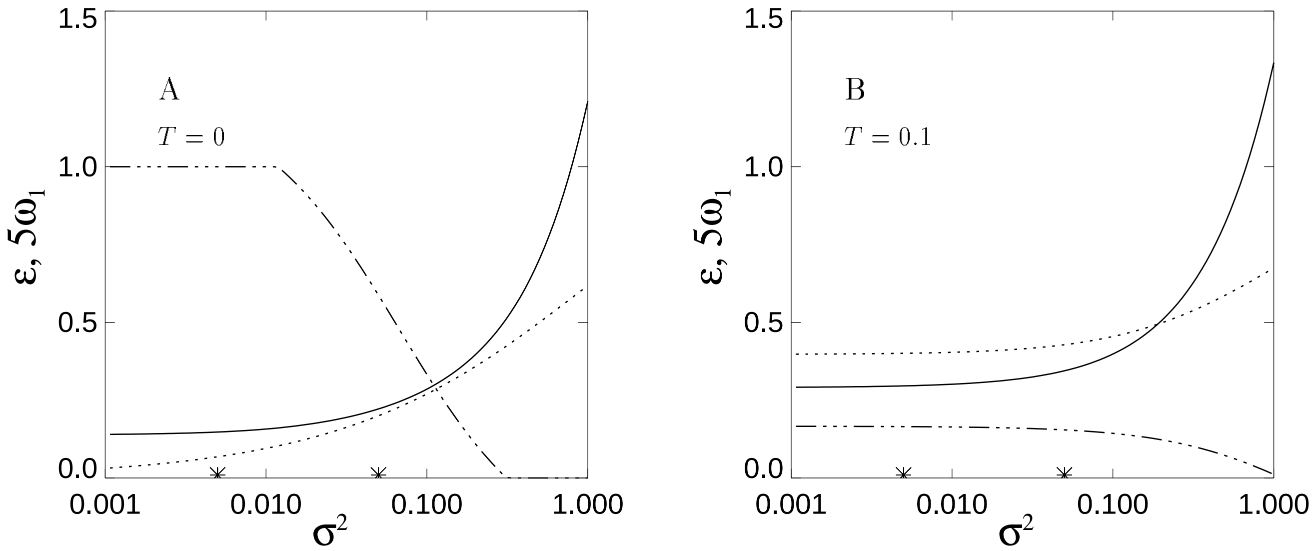

FIG. 4. Generalization error ϵ of an ensemble with 10 students made up of two groups of five identical students (with weight decays $\lambda_1 = \ldots = \lambda_5 = 0.005$, $\lambda_6 = \ldots = \lambda_{10} = 0.05$ as shown by stars on the x-axis), plotted vs. the noise level σ^2. The solid lines show the error for ensemble weights chosen to minimize the ensemble generalization error. The dot-dashed line is the total weight $5\omega_1$ assigned to the group of students with the smaller weight decay; as the noise level increases, the students with larger weight decay are favoured. For comparison, the generalization error of the optimal single student trained on all data (dotted line) is also plotted. As in Figures 3A and 3B, the plots are for $\alpha = 1$ and $c = 0.2$, with learning noise $T = 0$ and $T = 0.1$. Note how the range of noise levels σ^2 for which the ensemble performs better than the optimal single student has now become smaller.

optimize performance, again on the basis of the estimate of the ensemble generalization error (8). Examples of the resulting decrease in generalization error are shown in Figure 3A and 3B for an ensemble of size $K = 10$ with the weight decays λ_k equally spaced on a logarithmic axis between 10^{-3} and 1. For both of the temperatures T shown, the ensemble with uniform weights performs worse than the optimal single student. With weight optimization, the generalization performance approaches that of the optimal single student for $T = 0$, and is actually better at $T = 0.1$ over the whole range of noise levels σ^2 shown. Since even the best single student from the ensemble can never perform better than the optimal single student (which in general will not be contained in the ensemble), this implies that combining the student outputs in a weighted ensemble average is superior to simply choosing the best member of the ensemble by cross-validation, *i.e.*, on the basis of its estimated generalization error. The reason for this is that the ensemble average suppresses the learning noise on the individual students.

In Figure 3 we have also plotted the ensemble generalization error for the case when the ensemble weights are found by minimizing the ensemble training error (18). For small noise level σ^2 and $c = 0.2$ (Figure 3A and B) the result is essentially as good as for the generalization error minimization, but for larger noise levels the system starts to over-fit. Figures 3C and 3D show the case $c = 0$, where all the students are trained on the full data set. The absence of test error contributions from the ensemble training error is seen to lead to substantial over-fitting and can therefore not in general be recommended as a robust method of choosing the ensemble weights. When c is exactly zero, it is of course impossible to choose the ensemble weights by optimizing the estimated generalization error as there are no examples for testing. The corresponding lines in Figures 3C and 3D should therefore be understood as showing the limiting behaviour for $c \to 0$.

We have also studied the effect of weight optimization for ensembles of students whose weight decays cover only a fairly small range. As an example, Figure 4 shows the behaviour of an ensemble of $K = 10$ students consisting of two groups of five identical students each, with the weight decays of the two groups being fairly similar. Contrasting this with the case of an ensemble with a wide spread of different weight decays (see Figures 3A and 3B), we see that the range of noise levels σ^2 for which the generalization error of the ensemble with optimized weights is lower than that of the optimal single student has become smaller. In general, we thus expect it to be advantageous to have an ensemble of students with different degrees and/or kinds of regularization in order to make the performance improvement obtained from an ensemble with optimized weights robust against changes of the (unknown) noise level σ^2.

In Figure 4 we have also plotted the (total) weight which is assigned to the group of five students with the smaller weight decay when the ensemble generalization error is optimized. For low noise levels σ^2 and zero learning noise ($T = 0$, Figure 4A), this group of students carries all the weight, while the students with the higher weight decay are effectively switched off. This means that it is actually better to reduce the effective ensemble size to $K = 5$ than to retain highly over-regularized students in the ensemble. For finite learning noise (Figure 4B), on the other hand, the students with higher weight decay are never switched off completely; being able to average out learning noise by using the whole ensemble is obviously better than removing over-regularized students from the ensemble.

VI. CONCLUSIONS

We have studied ensemble learning for the simple, analytically solvable scenario of an ensemble of linear students. Our main findings, which correlate with experimental results presented in Ref. [9], are: In large ensembles, one should use under-regularized students in order to maximize the benefits of the variance-reducing effects of ensemble learning. In this way, the globally optimal generalization error achievable on the basis of *all* the available data can be reached when the training set sizes of the individual students are chosen optimally, and at the same time an unbiased estimate of the generalization error can be obtained. The ensemble performance is optimized when the more strongly regularized students are trained on a larger part of the available data, making the number of parameters which are well-determined by the training data equal for all students. For ensembles of more realistic size, we found that for students subject to a large amount of noise in the training process it is unnecessary to further increase the diversity of students by training them on smaller, less overlapping training sets. In this case, optimizing the ensemble weights is the method of choice for achieving low ensemble generalization error, and can yield better generalization performance than an optimally chosen single student subject to the same amount of learning noise and trained on all data. This improvement is most insensitive to changes in the unknown noise level σ^2 if the weight decays of the individual students cover a wide range. As mentioned in the introduction, we expect most of the above conclusions to carry over, at least qualitatively, to ensemble learning with more complex, nonlinear models.

APPENDIX A: ENSEMBLE ERRORS

In this appendix, we outline the calculation of the average ensemble generalization error (15, 16) and ensemble training error (18, 19, 20). While most of the averages involved can be carried out directly, the calculation of averages over training inputs is more complicated and is therefore described separately in Appendix B. We only detail the calculation for the case of a linear teacher; the generalization to a general nonlinear perceptron teacher can be obtained straightforwardly using the methods described in Ref. [22].

1. Ensemble generalization error

The ensemble generalization error can be measured with respect to the target output values either before or after noise is added. As mentioned in the text, we have chosen to use the noise free target values in our calculations; inclusion of the noise contribution would simply increase the value of the generalization error by σ^2. By definition, the generalization error of the ensemble with respect to the noise free target values is

$$\epsilon = \left\langle \left(\frac{1}{\sqrt{N}} \mathbf{x}^{\mathrm{T}} \mathbf{w}_0 - \sum_k \omega_k \frac{1}{\sqrt{N}} \mathbf{x}^{\mathrm{T}} \mathbf{w}_k \right)^2 \right\rangle_{\mathbf{x}} = \frac{1}{N} \left\langle \left(\sum_k \omega_k \mathbf{x}^{\mathrm{T}} \mathbf{v}_k \right)^2 \right\rangle_{\mathbf{x}} ,$$

where $\langle \cdot \rangle_{\mathbf{x}}$ is an average over the test input $\mathbf{x}$, $\mathbf{w}_k$ is the parameter vector of the k-th student, and we have introduced

$$\mathbf{v}_k = \mathbf{w}_0 - \mathbf{w}_k . \tag{A1}$$

The average over the assumed Gaussian distribution $P(\mathbf{x}) \propto \exp(-\frac{1}{2}\mathbf{x}^2)$ of test inputs yields $\langle \mathbf{x} \rangle_{\mathbf{x}} = 0$ and $\langle \mathbf{x}\mathbf{x}^{\mathrm{T}} \rangle_{\mathbf{x}} = \mathbf{1}$ and hence

$$\epsilon = \frac{1}{N} \left(\sum_k \omega_k \mathbf{v}_k \right)^2 . \tag{A2}$$

This expression now needs to be averaged over the student parameter vectors $\mathbf{w}_k$ (*i.e.*, over all realizations of the learning noise), and then over all training sets.

As explained in Sections III A and III B, the $\mathbf{w}_k$ are, for a given training set, distributed as $P(\mathbf{w}_k) \propto \exp(-E_k/2T_k)$ with

$$E_k = \sum_{\mu \in \mathcal{S}_k} \left(y^\mu - \frac{1}{\sqrt{N}} \mathbf{w}_k^{\mathrm{T}} \mathbf{x}^\mu \right)^2 + \lambda_k \mathbf{w}_k^2$$

where $\mu \in \mathcal{S}_k$ means that example $(\mathbf{x}^\mu, y^\mu)$ is contained in the training set of student k. The distributions of the different $\mathbf{w}_k$ are (for a fixed training set) independent of each other since each student is assumed to be subject to independent learning noise. Because the energy functions E_k are quadratic in $\mathbf{w}_k$, the joint distribution of the $\mathbf{w}_k$ is Gaussian with means and covariances

$$\begin{aligned} \langle \mathbf{w}_k \rangle &= \mathbf{g}_k \frac{1}{\sqrt{N}} \sum_{\mu \in \mathcal{S}_k} y^\mu \mathbf{x}^\mu \\ \left\langle \Delta\mathbf{w}_k \Delta\mathbf{w}_l^{\mathrm{T}} \right\rangle &= \left\langle \mathbf{w}_k \mathbf{w}_l^{\mathrm{T}} \right\rangle - \langle \mathbf{w}_k \rangle \langle \mathbf{w}_l \rangle^{\mathrm{T}} = \delta_{kl} T_k \mathbf{g}_k \end{aligned} \tag{A3}$$

where, by analogy with (11, 12),

$$\mathbf{g}_k^{-1} = \lambda_k \mathbf{1} + \mathbf{A}_k \qquad \mathbf{A}_k = \frac{1}{N} \sum_{\mu \in \mathcal{S}_k} \mathbf{x}^\mu (\mathbf{x}^\mu)^{\mathrm{T}}. \tag{A4}$$

Since the $\mathbf{v}_k$ differ from the $\mathbf{w}_k$ only by a constant vector, their covariances are identical to those of the $\mathbf{w}_k$, while their average values are

$$\langle \mathbf{v}_k \rangle = \mathbf{w}_0 - \langle \mathbf{w}_k \rangle = \mathbf{g}_k \left(\lambda_k \mathbf{w}_0 - \frac{1}{\sqrt{N}} \sum_{\mu \in \mathcal{S}_k} \eta^\mu \mathbf{x}^\mu \right). \tag{A5}$$

Here we have used the decomposition of the training outputs into noise free target values and additive noise,

$$y^\mu = \frac{1}{\sqrt{N}} \mathbf{w}_0^{\mathrm{T}} \mathbf{x}^\mu + \eta^\mu. \tag{A6}$$

Inserting (A3, A5) into (A2) and averaging over the η^μ yields ($\mathrm{tr}' \ldots = \frac{1}{N} \mathrm{tr} \ldots$)

$$\epsilon = \sum_{kl} \omega_k \omega_l \left[\lambda_k \lambda_l \frac{1}{N} \mathbf{w}_0^{\mathrm{T}} \langle \mathbf{g}_k \mathbf{g}_l \rangle \mathbf{w}_0 + \sigma^2 \, \mathrm{tr}' \langle \mathbf{g} \mathbf{A}_{kl} \mathbf{g} \rangle + \delta_{kl} T_k \, \mathrm{tr}' \langle \mathbf{g}_k \rangle \right], \tag{A7}$$

where

$$\mathbf{A}_{kl} = \frac{1}{N} \sum_{\mu \in \mathcal{S}_k \cap \mathcal{S}_l} \mathbf{x}^\mu (\mathbf{x}^\mu)^{\mathrm{T}}$$

is the covariance matrix of the inputs of the examples on which both student k and student l are trained. Only averages over training inputs now remain. The last term in (A7) is, by definition,

$$\mathrm{tr}' \langle \mathbf{g}_k \rangle = G(\alpha_k, \lambda_k) = G_k.$$

The first term can be simplified using the isotropy of the distribution of training inputs:

$$\frac{1}{N} \mathbf{w}_0^{\mathrm{T}} \langle \mathbf{g}_k \mathbf{g}_l \rangle \mathbf{w}_0 = \frac{1}{N} \mathbf{w}_0^{\mathrm{T}} \mathbf{w}_0 \, \mathrm{tr}' \langle \mathbf{g}_k \mathbf{g}_l \rangle = \mathrm{tr}' \langle \mathbf{g}_k \mathbf{g}_l \rangle$$

(remember that we assumed $\mathbf{w}_0^2 = N$). We are therefore left with two training input averages, which are evaluated in Appendix B:

$$\mathrm{tr}' \langle \mathbf{g}_k \mathbf{g}_l \rangle = \frac{G_k G_l (1 + G_k)(1 + G_l)}{(1 + G_k)(1 + G_l) - \alpha_{kl} G_k G_l} \tag{A8}$$

$$\mathrm{tr}' \langle \mathbf{g}_k \mathbf{A}_{kl} \mathbf{g}_l \rangle = \frac{\alpha_{kl} G_k G_l}{(1 + G_k)(1 + G_l) - \alpha_{kl} G_k G_l}. \tag{A9}$$

Inserting these results into (A7) and making use of (13) to simplify the expressions, one obtains the result (15, 16) given in the text.

2. Ensemble training error

The same techniques as above can be used to calculate the ensemble error on the training set, although the resulting expressions are slightly more cumbersome. The (normalized) ensemble training error is defined as

$$\epsilon^{\mathrm{t}} = \left\langle \frac{1}{p}\sum_{\mu}\left(y^{\mu} - \sum_{k}\omega_k f_k(\mathbf{x}^{\mu})\right)^2\right\rangle = \left\langle \frac{1}{p}\sum_{\mu}\left(\frac{1}{\sqrt{N}}\sum_{k}\omega_k \mathbf{v}_k^{\mathrm{T}}\mathbf{x}^{\mu} + \eta^{\mu}\right)^2\right\rangle \tag{A10}$$

where we have made use of (A1) and the decomposition (A6). The average over the distribution of the $\mathbf{v}_k$, *i.e.*, over the learning noise, can be carried out as in the previous section and yields

$$\epsilon^{\mathrm{t}} = \frac{1}{p}\sum_{kl}\omega_k\omega_l\left\langle \bar{\mathbf{v}}_k^{\mathrm{T}}\mathbf{A}\bar{\mathbf{v}}_l\right\rangle + \frac{2}{p\sqrt{N}}\sum_{k}\omega_k\sum_{\mu}\left\langle \eta^{\mu}\bar{\mathbf{v}}_k^{\mathrm{T}}\mathbf{x}^{\mu}\right\rangle + \frac{1}{p}\sum_{\mu}\left\langle(\eta^{\mu})^2\right\rangle + \delta_{kl}T_k\frac{1}{p}\mathrm{tr}\,\langle \mathbf{g}_k\mathbf{A}\rangle \tag{A11}$$

where we have denoted by $\bar{\mathbf{v}}_k$ the averages of the $\mathbf{v}_k$ over the learning noise. Inserting the explicit form (A5) of the $\bar{\mathbf{v}}_k$ and averaging over the η^{μ}, the first term of (A11) becomes

$$\frac{1}{N}\left\langle(\bar{\mathbf{v}}_k)^{\mathrm{T}}\mathbf{A}\bar{\mathbf{v}}^l\right\rangle = \lambda_k\lambda_l\,\mathrm{tr}'\langle\mathbf{g}_k\mathbf{A}\mathbf{g}_l\rangle + \sigma^2\,\mathrm{tr}'\langle\mathbf{A}\mathbf{g}_k\mathbf{A}_{kl}\mathbf{g}_l\rangle\,.$$

For the second term one finds

$$\begin{aligned}\frac{2}{p\sqrt{N}}\left\langle\sum_{\mu}\eta^{\mu}\bar{\mathbf{v}}_k^{\mathrm{T}}\mathbf{x}^{\mu}\right\rangle &= \frac{2}{p\sqrt{N}}\left\langle\sum_{\mu}\eta^{\mu}\left(\lambda\mathbf{w}_0 - \frac{1}{\sqrt{N}}\sum_{\nu\in\mathcal{S}_k}\eta^{\nu}\mathbf{x}^{\nu}\right)^{\mathrm{T}}\mathbf{g}_k\mathbf{x}^{\mu}\right\rangle \\ &= -\frac{2\sigma^2}{p}\mathrm{tr}\,\langle\mathbf{g}_k\mathbf{A}_k\rangle = -\frac{2\sigma^2}{\alpha}(1-\lambda_k G_k)\end{aligned}$$

where (A4) was used. Including the sum over k, this can be written as $(-\sigma^2/\alpha)\ \sum_{kl}\omega_k\omega_l[(1-\lambda_k G_k)+\ (1-\lambda_l G_l)]$. Together with the trivial average $\left\langle(\eta^{\mu})^2\right\rangle = \sigma^2 = \sum_{kl}\omega_k\omega_l\sigma^2$, one thus has

$$\begin{aligned}\epsilon^{\mathrm{t}} &= \sum_{kl}\omega_k\omega_l\epsilon_{kl}^{\mathrm{t}} \\ \epsilon_{kl}^{\mathrm{t}} &= \sigma^2 + \frac{1}{\alpha}\left[\lambda_k\lambda_l\,\mathrm{tr}'\langle\mathbf{g}_k\mathbf{A}\mathbf{g}_l\rangle + \sigma^2\,\mathrm{tr}'\langle\mathbf{A}\mathbf{g}_k\mathbf{A}_{kl}\mathbf{g}_l\rangle - \sigma^2(2-\lambda_k G_k - \lambda_l G_l) + \delta_{kl}T_k\,\mathrm{tr}'\langle\mathbf{g}_k\mathbf{A}\rangle\right].\end{aligned} \tag{A12}$$

There are now three terms that need to be averaged over training inputs. In the last one, the average over the inputs which are not part of the training set of student k can be done directly, yielding

$$\mathrm{tr}'\langle\mathbf{g}_k\mathbf{A}\rangle = (\alpha-\alpha_k)\,\mathrm{tr}'\langle\mathbf{g}_k\rangle +\ \mathrm{tr}'\langle\mathbf{g}_k\mathbf{A}_k\rangle = (\alpha-\alpha_k)G_k + 1 - \lambda_k G_k. \tag{A13}$$

Similarly, the first average in (A12) can be reduced by splitting off the examples on which neither student k nor student l are trained:

$$\begin{aligned}\mathrm{tr}'\langle\mathbf{g}_k\mathbf{A}\mathbf{g}_l\rangle &= (\alpha+\alpha_{kl}-\alpha_k-\alpha_l)\,\mathrm{tr}'\langle\mathbf{g}_k\mathbf{g}_l\rangle +\ \mathrm{tr}'\langle\mathbf{g}_k(\mathbf{A}_k+\mathbf{A}_l-\mathbf{A}_{kl})\mathbf{g}_l\rangle \\ &= (\alpha+\alpha_{kl}-\alpha_k-\alpha_l)\,\mathrm{tr}'\langle\mathbf{g}_k\mathbf{g}_l\rangle +\ \mathrm{tr}'\langle\mathbf{g}_k(1-\lambda_l\mathbf{g}_l)+\mathbf{g}_l(1-\lambda_k\mathbf{g}_k)-\mathbf{g}_k\mathbf{A}_{kl}\mathbf{g}_l\rangle \\ &= (\alpha+\alpha_{kl}-\alpha_k-\alpha_l-\lambda_k-\lambda_l)\,\mathrm{tr}'\langle\mathbf{g}_k\mathbf{g}_l\rangle + G_k + G_l -\ \mathrm{tr}'\langle\mathbf{g}_k\mathbf{A}_{kl}\mathbf{g}_l\rangle\,.\end{aligned} \tag{A14}$$

The two averages in the last expression also occur in the generalization error, see (A8, A9). The only remaining new average in (A12) is shown in Appendix B to be

$$\mathrm{tr}'\langle\mathbf{A}\mathbf{g}_k\mathbf{A}_{kl}\mathbf{g}_l\rangle = \left(\alpha - \frac{\alpha_k G_k}{1+G_k} - \frac{\alpha_l G_l}{1+G_l} + 1\right)\ \mathrm{tr}'\langle\mathbf{g}_k\mathbf{A}_{kl}\mathbf{g}_l\rangle\,. \tag{A15}$$

The final result (18, 19, 20) for the ensemble training error is obtained by inserting (A13, A14, A15) into (A12) and simplifying by making extensive use of (13).

APPENDIX B: AVERAGES OVER TRAINING INPUTS

We now show how the averages over training inputs appearing in the expressions for the ensemble generalization and training error can be calculated. Two methods are described. The diagrammatic technique in Section B 1 may be easier to follow for readers familiar with field theoretic methods, while the differential equation method explained in Section B 2 is somewhat more basic, being based only on simple matrix identities.

1. Diagrammatic technique

The diagrammatic technique we use here was introduced in Refs. [24,31], to which we refer the reader for a detailed exposition. The relevant notation is explained in Figure 5, while Figure 6 gives a summary of the method, using the average $\mathrm{tr}'\langle \mathbf{g}_k \mathbf{g}_l \rangle$ as an example. From the diagrammatic expansion in Figure 6c one sees that

$$\mathrm{tr}'\langle \mathbf{g}_k \mathbf{g}_l \rangle = G_k Z^1_{kl} G_l, \tag{B1}$$

where Z^1_{kl} is the sum of the irreducible diagrams shown in Figure 6d. This sum can be evaluated as

$$Z^1_{kl} = 1 + \alpha_{kl}\, \mathrm{tr}'\langle \mathbf{g}_k \mathbf{g}_l \rangle\, Q_{kl}, \tag{B2}$$

where

$$\begin{aligned} Q_{kl} &= 1 - (G_k + G_l) + (G_k^2 + G_k G_l + G_l^2) + \ldots \\ &= (1 - G_k + G_k^2 - G_k^3 + \ldots)(1 - G_l + G_l^2 - G_l^3 + \ldots) = \frac{1}{(1+G_k)(1+G_l)}\ , \end{aligned}$$

a series that will occur several times below. Combining (B1) and (B2), we deduce the result (A8) stated above.

For the second of the averages required, $\mathrm{tr}'\langle \mathbf{g}_k \mathbf{A}_{kl} \mathbf{g}_l \rangle$, the diagrammatic expansion is similar (Figure 7). The irreducible diagrams sum to

$$Z^2_{kl} = \alpha_{kl} Q_{kl} + \alpha_{kl}\, \mathrm{tr}'\langle \mathbf{g}_k \mathbf{A}_{kl} \mathbf{g}_l \rangle\, Q_{kl},$$

and using $\mathrm{tr}'\langle \mathbf{g}_k \mathbf{A}_{kl} \mathbf{g}_l \rangle = G_k Z^2_{kl} G_l$ one derives the result (A9).

Finally, the diagrammatic expansion of the average $\mathrm{tr}'\langle \mathbf{A} \mathbf{g}_k \mathbf{A}_{kl} \mathbf{g}_l \rangle$ required for the calculation of the ensemble training error is shown in Figure 8. The four series into which the diagrams can be sorted sum to

$$\begin{aligned} \mathrm{tr}'\langle \mathbf{A} \mathbf{g}_k \mathbf{A}_{kl} \mathbf{g}_l \rangle \ &= \left(\alpha - \frac{\alpha_k G_k}{1+G_k}\right) \mathrm{tr}'\langle \mathbf{g}_k \mathbf{A}_{kl} \mathbf{g}_l \rangle + \alpha_{kl} G_k G_l Q_{kl} \\ &\quad - \frac{\alpha_l G_l}{1+G_l}\, \mathrm{tr}'\langle \mathbf{g}_k \mathbf{A}_{kl} \mathbf{g}_l \rangle + \alpha_{kl} G_l G_k Q_{kl}\, \mathrm{tr}'\langle \mathbf{g}_k \mathbf{A}_{kl} \mathbf{g}_l \rangle\, . \end{aligned} \tag{B3}$$

From (A9) one sees that

$$\alpha_{kl} G_k G_l Q_{kl} \left(1 + \mathrm{tr}'\langle \mathbf{g}_k \mathbf{A}_{kl} \mathbf{g}_l \rangle\right) = \mathrm{tr}'\langle \mathbf{g}_k \mathbf{A}_{kl} \mathbf{g}_l \rangle\, ,$$

and inserting this into (B3) yields the result (A15) stated in the previous section.

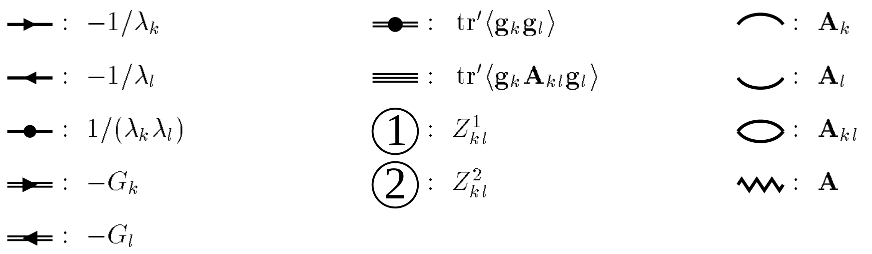

FIG. 5. The correspondence between the diagrams and the mathematical expressions.

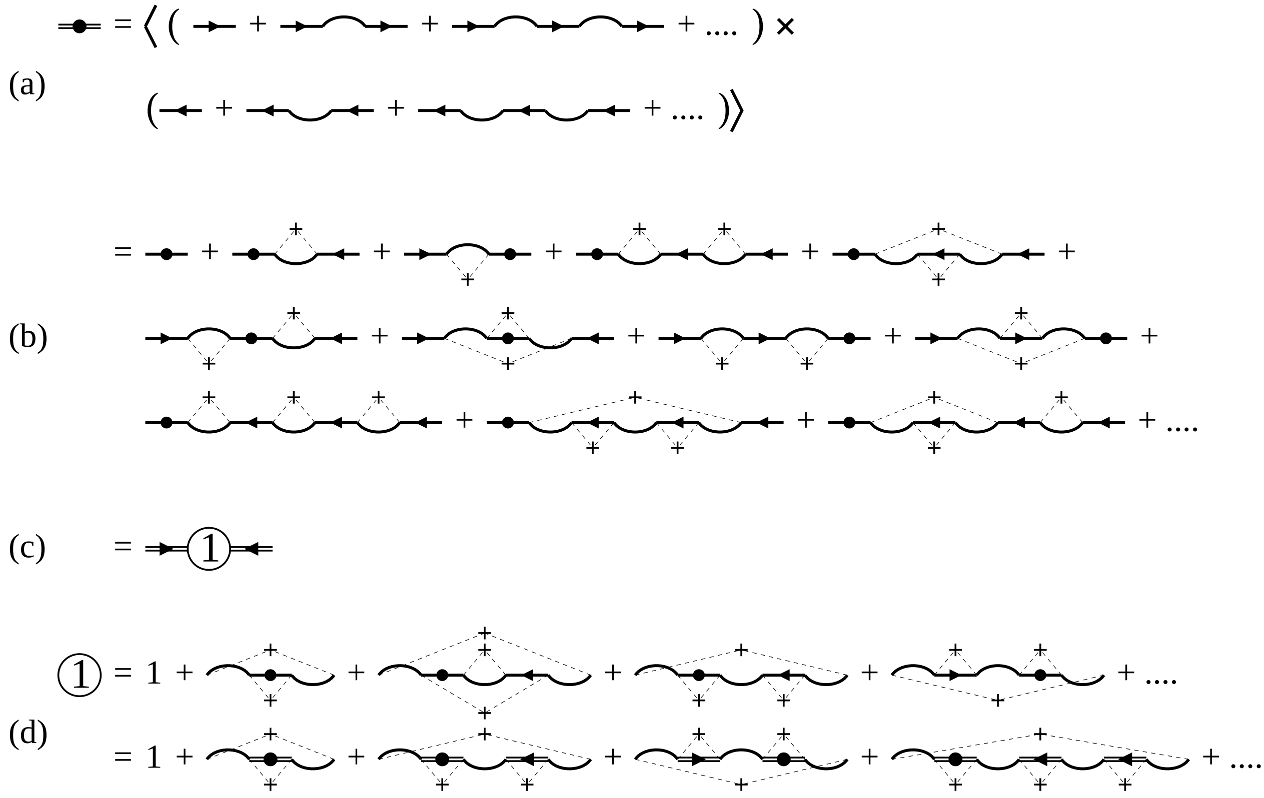

FIG. 6. Diagrams for calculation of $\mathrm{tr}'\langle \mathbf{g}_k \mathbf{g}_l \rangle$. All the symbols are explained in Figure 5. (a) $\mathbf{g}_k \mathbf{g}_l$ is drawn as a product of expansions $\mathbf{g}_k = (\lambda_k \mathbf{1} + \mathbf{A}_k)^{-1} = -(-\lambda_k^{-1}\mathbf{1} + \lambda_k^{-1}\mathbf{A}_k\lambda_k^{-1} - \lambda_k^{-1}\mathbf{A}_k\lambda_k^{-1}\mathbf{A}_k\lambda_k^{-1} + \ldots)$, and similarly for $\mathbf{g}_l$. (b) All the terms arising from the above product must be averaged. The averaging can be done by pairing the training inputs that occur in the products of the matrices $\mathbf{A}_k$ and $\mathbf{A}_l$ (see Ref. [31]), as indicated by dashed lines with a "+". Only diagrams where the dashed lines do not cross survive for $N \to \infty$. (c) When all the irreducible diagrams are collected as shown in (d), the expression becomes simple. The irreducible diagrams are those which cannot be cut in two without cutting a dashed line. In the last line the subdiagrams corresponding to $-G_k$, $-G_l$, and $\mathrm{tr}'\langle \mathbf{g}_k \mathbf{g}_l \rangle$ have been identified ('dressing').

2. Differential equation method

An alternative method for calculating averages over training inputs, which we describe in the present section, was introduced in Ref. [22]. It is based on considering the effect of incremental changes in the size of the students' training sets, which in the thermodynamic limit result in partial differential equations for the required averages. The basic building block is the matrix identity

$$\left(\mathbf{M} + \tfrac{1}{N}\mathbf{x}\mathbf{x}^{\mathrm{T}}\right)^{-1} = \mathbf{M}^{-1} - \frac{1}{N}\frac{\mathbf{M}^{-1}\mathbf{x}\mathbf{x}^{\mathrm{T}}\mathbf{M}^{-1}}{1 + \frac{1}{N}\mathbf{x}^{\mathrm{T}}\mathbf{M}^{-1}\mathbf{x}} \tag{B4}$$

which, as can easily be verified, holds for any vector $\mathbf{x}$ and any positive definite symmetric matrix $\mathbf{M}$.

Consider now the average $G_{kl} = \mathrm{tr}'\langle \mathbf{g}_k \mathbf{g}_l \rangle$, which is a function of the size of the training sets of students k and l, α_k and α_l, their overlap α_{kl}, and the weight decay parameters λ_k and λ_l. Writing $\alpha_k = \alpha_{kl} + \Delta_k$ and $\alpha_l = \alpha_{kl} + \Delta_l$, we calculate the variation of G_{kl} with α_{kl} for fixed Δ_k and Δ_l. Varying α_{kl} by $1/N$ means adding one new training example (whose input vector we simply write as $\mathbf{x}$) to the training sets of students k and l. Denoting the resulting new 'response matrices' by $\mathbf{g}_k^+$ and $\mathbf{g}_l^+$, we have from (B4)

$$\frac{1}{N}\,\mathrm{tr}\,\mathbf{g}_k^+\mathbf{g}_l^+ = \frac{1}{N}\,\mathrm{tr}\,\mathbf{g}_k\mathbf{g}_l + \frac{1}{N}\left[-\frac{\frac{1}{N}\mathbf{x}^{\mathrm{T}}\mathbf{g}_k\mathbf{g}_l\mathbf{g}_k\mathbf{x}}{1 + \frac{1}{N}\mathbf{x}^{\mathrm{T}}\mathbf{g}_k\mathbf{x}} - \frac{\frac{1}{N}\mathbf{x}^{\mathrm{T}}\mathbf{g}_l\mathbf{g}_k\mathbf{g}_l\mathbf{x}}{1 + \frac{1}{N}\mathbf{x}^{\mathrm{T}}\mathbf{g}_l\mathbf{x}} + \frac{(\frac{1}{N}\mathbf{x}^{\mathrm{T}}\mathbf{g}_k\mathbf{g}_l\mathbf{x})^2}{(1 + \frac{1}{N}\mathbf{x}^{\mathrm{T}}\mathbf{g}_k\mathbf{x})(1 + \frac{1}{N}\mathbf{x}^{\mathrm{T}}\mathbf{g}_l\mathbf{x})}\right] \tag{B5}$$

To get an equation for G_{kl}, this has to be averaged over both the new and the existing training inputs. The average over the new input can be done by noting that for the assumed Gaussian input distribution $P(\mathbf{x}) \propto \exp(-\mathbf{x}^2/2)$ one has

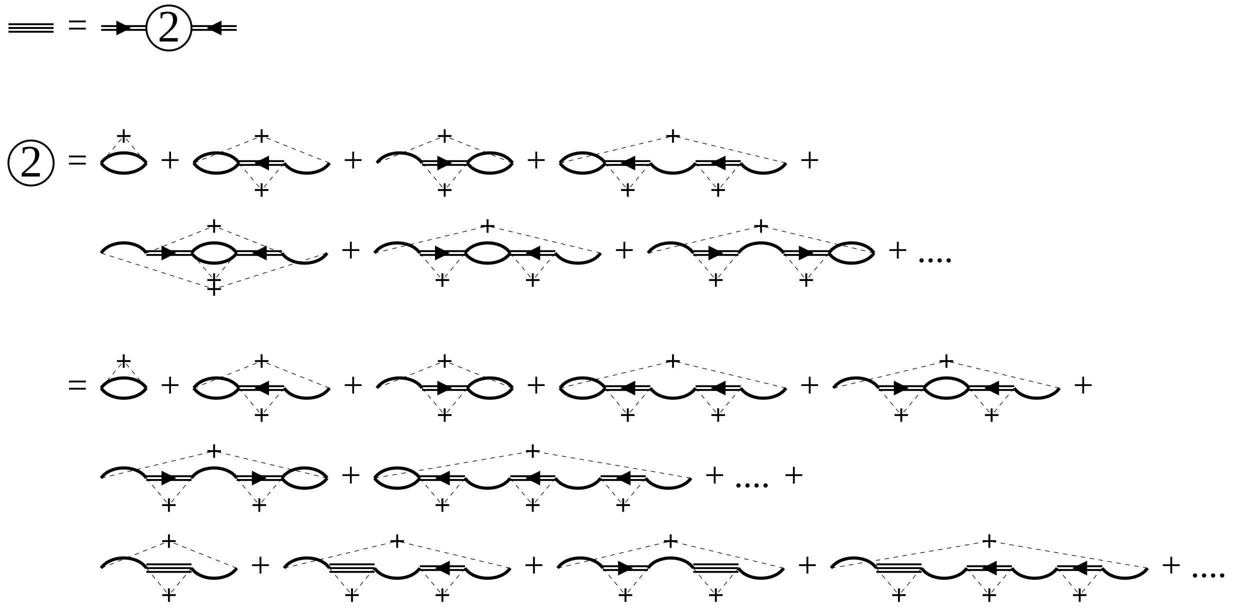

FIG. 7. Diagrams for calculation of $\mathrm{tr}'\langle \mathbf{g}_k \mathbf{A}_{kl} \mathbf{g}_l \rangle$. In the last line those of the irreducible diagrams containing $\mathrm{tr}'\langle \mathbf{g}_k \mathbf{A}_{kl} \mathbf{g}_l \rangle$ itself are singled out.

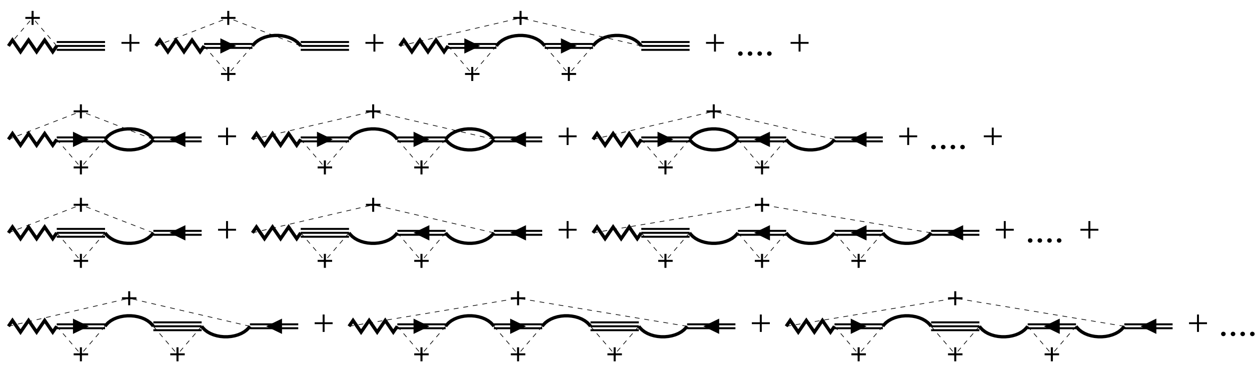

FIG. 8. Diagrams for calculation of $\mathrm{tr}'\langle \mathbf{A} \mathbf{g}_k \mathbf{A}_{kl} \mathbf{g}_l \rangle$. They are naturally separated into four series (from the top): Diagrams which contain $\mathrm{tr}'\langle \mathbf{g}_k \mathbf{A}_{kl} \mathbf{g}_l \rangle$ as a factor; those in which $\mathbf{A}_{kl}$ cannot be incorporated in a $\mathrm{tr}'\langle \mathbf{g}_k \mathbf{A}_{kl} \mathbf{g}_l \rangle$ average; and those containing an average of $\mathbf{A}$ and $\mathrm{tr}'\langle \mathbf{g}_k \mathbf{A}_{kl} \mathbf{g}_l \rangle$ in an irreducible combination, grouped according to whether $\mathbf{A}$ appears next to $\mathrm{tr}'\langle \mathbf{g}_k \mathbf{A}_{kl} \mathbf{g}_l \rangle$ or not.

$$\frac{1}{N} \mathbf{x}^{\mathrm{T}} \mathbf{M} \mathbf{x} = \mathrm{tr}' \, \mathbf{M} + O\left(N^{-1/2}\right)$$

where $\mathbf{M}$ can be any product of powers of $\mathbf{g}_k$ and $\mathbf{g}_l$ [32]. This yields

$$\frac{\partial G_{kl}}{\partial \alpha_{kl}} = \left\langle \frac{(\partial/\partial\lambda_k) \frac{1}{N} \, \mathrm{tr} \, \mathbf{g}_k \mathbf{g}_l}{1 + \frac{1}{N} \, \mathrm{tr} \, \mathbf{g}_k} + \frac{(\partial/\partial\lambda_l) \frac{1}{N} \, \mathrm{tr} \, \mathbf{g}_k \mathbf{g}_l}{1 + \frac{1}{N} \, \mathrm{tr} \, \mathbf{g}_l} + \frac{(\frac{1}{N} \, \mathrm{tr} \, \mathbf{g}_k \mathbf{g}_l)^2}{(1 + \frac{1}{N} \, \mathrm{tr} \, \mathbf{g}_k)(1 + \frac{1}{N} \, \mathrm{tr} \, \mathbf{g}_l)} \right\rangle$$

up to terms of $O(N^{-1/2})$; the remaining average is over the existing training inputs. Using the self-averaging property of the response functions, $\frac{1}{N} \, \mathrm{tr} \, \mathbf{g}_{k/l} = G_{k/l} + O(N^{-1/2})$ and $\frac{1}{N} \, \mathrm{tr} \, \mathbf{g}_k \mathbf{g}_l = G_{kl} + O(N^{-1/2})$ (which can derived from the recursion relation (B5), compare the discussion in Ref. [22]), this average becomes trivial in the thermodynamic limit and one obtains the partial differential equation

$$\frac{\partial G_{kl}}{\partial \alpha_{kl}} - \frac{1}{1+G_k} \frac{\partial G_{kl}}{\partial \lambda_k} - \frac{1}{1+G_l} \frac{\partial G_{kl}}{\partial \lambda_l} = \frac{G_{kl}^2}{(1+G_k)(1+G_l)}. \tag{B6}$$

This can now be solved using the method of characteristic curves (see, *e.g.*, Ref. [33], or Ref. [22] for a brief review). The characteristic curves of (B6) are defined by

$$\frac{d\alpha_{kl}}{dt} = 1 \,, \quad \frac{d\lambda_k}{dt} = -\frac{1}{1+G_k} \,, \quad \frac{d\lambda_l}{dt} = -\frac{1}{1+G_l} \,, \quad \frac{dG_{kl}}{dt} = \frac{G_{kl}^2}{(1+G_k)(1+G_l)} \tag{B7}$$

(t being the curve parameter), and the solution 'surface' $G_{kl} = G_{kl}(\alpha_{kl}, \lambda_k, \lambda_l)$ is the union of those characteristic curves which satisfy the required initial condition $G_{kl}|_{\alpha_{kl}=0} = G_k G_l$. Using (14)—which is, in fact, the solution of the differential equation $\partial G/\partial\alpha - (1+G)^{-1}\partial G/\partial\lambda = 0$, derived analogously to (B6) as described in Ref. [22]—one verifies that G_k and G_l are constant along the characteristic curves. This makes the integration of (B7) trivial: Selecting the arbitrary origin of the t-scale such that $\alpha_{kl} = 0$ at $t = 0$, the first and last equations of (B7) yield directly

$$-\frac{1}{G_{kl}} = -\frac{1}{G_{kl}|_{\alpha_{kl}=0}} + \frac{\alpha_{kl}}{(1+G_k)(1+G_l)} = -\frac{1}{G_k G_l} + \frac{\alpha_{kl}}{(1+G_k)(1+G_l)}$$

which gives the desired result (A8).

The remaining averages can be deduced from G_{kl} by applying the identity (B4). Considering $\mathrm{tr}'\langle \mathbf{g}_k \mathbf{A}_{kl} \mathbf{g}_l \rangle$, we first write explicitly

$$\mathrm{tr}'\, \mathbf{g}_k \mathbf{A}_{kl} \mathbf{g}_l = \frac{1}{N^2} \sum_{\mu \in \mathcal{S}_k \cap \mathcal{S}_l} (\mathbf{x}^\mu)^{\mathrm{T}} \mathbf{g}_k \mathbf{g}_l \mathbf{x}^\mu .$$

Because both $\mathbf{g}_k$ and $\mathbf{g}_l$ depend on $\mathbf{x}^\mu$, one *cannot* replace $\frac{1}{N}(\mathbf{x}^\mu)^{\mathrm{T}} \mathbf{g}_k \mathbf{g}_l \mathbf{x}^\mu \to \frac{1}{N}\,\mathrm{tr}\; \mathbf{g}_k \mathbf{g}_l + O(N^{-1/2})$. Instead, one needs to 'pull $\mathbf{x}^\mu$ out' of $\mathbf{g}_k$ and $\mathbf{g}_l$ by using (B4) in reverse: Writing $(\mathbf{g}_k)^{-1} = (\mathbf{g}_k^\mu)^{-1} + \frac{1}{N}\mathbf{x}^\mu(\mathbf{x}^\mu)^{\mathrm{T}}$, one has

$$\mathbf{g}_k \mathbf{x}^\mu = \left(\mathbf{g}_k^\mu - \frac{1}{N} \frac{\mathbf{g}_k^\mu \mathbf{x}^\mu (\mathbf{x}^\mu)^{\mathrm{T}} \mathbf{g}_k^\mu}{1 + \frac{1}{N}(\mathbf{x}^\mu)^{\mathrm{T}} \mathbf{g}_k^\mu \mathbf{x}^\mu} \right) \mathbf{x}^\mu = \frac{\mathbf{g}_k^\mu \mathbf{x}^\mu}{1 + \frac{1}{N}\mathbf{x}^{\mathrm{T}} \mathbf{g}_k^\mu \mathbf{x}}$$

and similarly for $\mathbf{g}_l \mathbf{x}^\mu$. Since $\mathbf{g}_k^\mu$ and $\mathbf{g}_l^\mu$ are independent of $\mathbf{x}^\mu$, one can now invoke self-averaging: $\frac{1}{N}\mathbf{x}^{\mathrm{T}} \mathbf{g}_k^\mu \mathbf{x} = \frac{1}{N}\,\mathrm{tr}\; \mathbf{g}_k^\mu + O(N^{-1/2}) = \mathrm{tr}'\langle \mathbf{g}_k^\mu \rangle + O(N^{-1/2})$, and since removing example μ corresponds to reducing α_k by $1/N$, $\mathrm{tr}'\langle \mathbf{g}_k^\mu \rangle = G_k + O(N^{-1})$. One can thus write

$$\frac{1}{N}(\mathbf{x}^\mu)^{\mathrm{T}} \mathbf{g}_k \mathbf{g}_l \mathbf{x}^\mu = \frac{G_{kl}}{(1+G_k)(1+G_l)} + O\left(N^{-1/2}\right) ;$$

summing this over μ one obtains $\mathrm{tr}'\langle \mathbf{g}_k \mathbf{A}_{kl} \mathbf{g}_l \rangle = \alpha_{kl} G_{kl}/(1+G_k)(1+G_l)$ and hence (A9).

The final average can be obtained by the same technique:

$$\mathrm{tr}'\langle \mathbf{A} \mathbf{g}_k \mathbf{A}_{kl} \mathbf{g}_l \rangle = \mathrm{tr}'\langle \mathbf{g}_k \mathbf{A}_{kl} \mathbf{g}_l \rangle \left[(\alpha - \alpha_k - \alpha_l + \alpha_{kl}) + \frac{\alpha_k - \alpha_{kl}}{1+G_k} + \frac{\alpha_l - \alpha_{kl}}{1+G_l} \right] + \mathrm{tr}'\langle \mathbf{A}_{kl} \mathbf{g}_k \mathbf{A}_{kl} \mathbf{g}_l \rangle . \tag{B8}$$

The terms on the r.h.s. correspond, from left to right, to training examples not contained in either $\mathcal{S}_k$ or $\mathcal{S}_l$, contained in $\mathcal{S}_k$ but not in $\mathcal{S}_l$ and vice versa, and contained in $\mathcal{S}_k \cap \mathcal{S}_l$. The last term can be written as

$$\frac{1}{N} \sum_{\mu \in \mathcal{S}_k \cap \mathcal{S}_l} \left[\frac{1}{N}(\mathbf{x}^\mu)^{\mathrm{T}} \mathbf{g}_k \left(\mathbf{A}_{kl} - \frac{1}{N}\mathbf{x}^\mu (\mathbf{x}^\mu)^{\mathrm{T}} \right) \mathbf{g}_l \mathbf{x}^\mu + \frac{1}{N}(\mathbf{x}^\mu)^{\mathrm{T}} \mathbf{g}_k \mathbf{x}^\mu \frac{1}{N}(\mathbf{x}^\mu)^{\mathrm{T}} \mathbf{g}_l \mathbf{x}^\mu \right] =$$

$$\frac{\alpha_{kl}}{(1+G_k)(1+G_l)}\, \mathrm{tr}'\langle \mathbf{g}_k \mathbf{A}_{kl} \mathbf{g}_l \rangle + \alpha_{kl} \frac{G_k}{1+G_k} \frac{G_l}{1+G_l} + O\left(N^{-1/2}\right) .$$

Inserting this into (B8), one is led back to (B3), from which the result (A15) follows.

---

* Present address: Centre for Biological Sequence Analysis, University of Denmark, Building 206, DK-2800 Lyngby, Denmark. email: `krogh@cbs.dtu.dk`

** email: P.Sollich@ed.ac.uk

[1] J. Hertz, A. Krogh, and R. G. Palmer, *Introduction to the Theory of Neural Computation* (Addison-Wesley, Redwood City, California, 1991).
[2] T. L. H. Watkin, A. Rau, and M. Biehl, Rev. Modern Phys. **65**, 499 (1993).
[3] C. W. J. Granger, Journal of Forecasting **8**, 167 (1989).
[4] D. H. Wolpert, Neural Networks **5**, 241 (1992).
[5] L. Breimann, Technical Report No. 367, Department of Statistics, UC Berkeley (unpublished).
[6] L. K. Hansen and P. Salamon, IEEE Transactions on Pattern Analysis and Machine Intelligence **12**, 993 (1990).
[7] M. P. Perrone and L. N. Cooper, in *Neural Networks for Speech and Image processing*, edited by R. J. Mammone (Chapman-Hall, 1993).
[8] S. Hashem, IEEE Transactions on Neural Networks **6**, 792 (1995).
[9] A. Krogh and J. Vedelsby, in *Advances in Neural Information Processing Systems 7*, edited by G. Tesauro, D. S. Touretzky, and T. K. Leen (MIT Press, Cambridge, MA, 1995), pp. 231–238.
[10] R. Jacobs, M. Jordan, S. Nowlan, and G. Hinton, Neural Computation **3**, 79 (1991).
[11] R. E. Schapire, Machine Learning **5**, 197 (1990).
[12] Y. Freund, Information and Computation **121**, 256 (1995).
[13] R. Meir, in *Advances in Neural Information Processing Systems 7*, edited by G. Tesauro, D. S. Touretzky, and T. K. Leen (MIT Press, Cambridge, MA, 1995), pp. 295–302.
[14] P. Sollich and A. Krogh, in *Advances in Neural Information Processing Systems 8*, edited by D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo (MIT Press, Cambridge, MA, 1996), pp. 190–196.
[15] S. Geman, E. Bienenstock, and R. Doursat, Neural Comp. **4**, 1 (1992).
[16] In order for the estimate (8) to be unbiased, the ambiguity must be estimated from a sample of inputs representative of the distribution $P(\mathbf{x})$ as a whole, rather than from the training inputs alone. Otherwise, the ambiguity will normally be underestimated, since all students trained on a particular example will tend to reproduce the corresponding training output and therefore differ less than typical in their predictions for this input.
[17] D. J. C. MacKay, Neural Comp. **4**, 448 (1992).
[18] W. L. Buntine and A. S. Weigend, Complex Syst. **5**, 603 (1991).
[19] A. Bruce and D. Saad, J. Phys. A **27**, 3355 (1994).
[20] H. S. Seung, H. Sompolinsky, and N. Tishby, Phys. Rev. A **45**, 6056 (1992).
[21] The factor of two in the Gibbs distribution is introduced merely in order to eliminate cumbersome numerical factors in later results.
[22] P. Sollich, J. Phys. A **27**, 7771 (1994).
[23] A. Krogh and J. A. Hertz, J. Phys. A **25**, 1135 (1992).
[24] J. A. Hertz, A. Krogh, and G. I. Thorbergsson, J. Phys. A **22**, 2133 (1989).
[25] P. Sollich, J. Phys. A **28**, 6125 (1995).
[26] S. Bös, W. Kinzel, and M. Opper, Phys. Rev. E **47**, 1384 (1993).
[27] T. L. H. Watkin, Europhys. Lett. **21**, 871 (1993).
[28] J. Berger, *Statistical Decision Theory and Bayesian Analysis* (Springer, NewYork, 1985).
[29] J. Pilz, *Bayesian Estimation and Experimental Design in Linear Regression Models*, 2nd ed. (John Wiley, Chichester, 1991), (1st edition Teubner, Leipzig, 1983).
[30] D. J. C. MacKay, Neural Comp. **4**, 415 (1992).
[31] A. Krogh, J. Phys. A **25**, 1119 (1992).
[32] In fact, it suffices that the norm of $\mathbf{M}$ (*i.e.*, its maximal eigenvalue) be of $O(1)$.
[33] M. L. Eaton, *Multivariate Statistics - A Vector Space Approach* (John Wiley, New York, 1983).