No Free Lunch for Early Stopping

Zehra Cataltepe
Bell Laboratories, Lucent Technologies
600 Mountain Ave, Rm 2C-265, Murray Hill, NJ 07974, U.S.A.
zehra@lucent.com

Yaser S. Abu-Mostafa
Malik Magdon-Ismail
Learning Systems Group, California Institute of Technology
MC 136-93, Pasadena, CA 91125, U.S.A
{yaser, magdon}@cs.caltech.edu

Abstract

We show that, with a uniform prior on models having the same training error, early stopping at some fixed training error above the training error minimum results in an increase in the expected generalization error.

1 Introduction

Early stopping of training is one of the methods that aim to prevent overtraining due to too powerful a model class, noisy training examples or a small training set. We study early stopping at a predetermined training error level. If there is no prior information, other than the training examples, all models with the same training error should be equally likely to be chosen as the early stopping solution. When this is the case, we show that, for general linear models, early stopping at any training error level above the training error minimum increases the expected generalization error. Moreover, we also show that the generalization error is an increasing function of the training error. Our results are nonasymptotic and independent of the presence or nature of the training data noise, and they hold when instead of generalization error, test error or off-training-set error [1] [Wolpert, 1996b] are used as the performance criterion. For general nonlinear models, around a small enough neighborhood of a training error minimum, the mean generalization error again increases, when all models with the same training

[1] Off-training-set error does not assume that the training and test inputs come from the same distribution.

error are equally likely. Regularization methods such as weight decay, early stopping using a validation set, or early stopping of training using a hint error are equivalent to early stopping at a fixed training error level but with a nonuniform probability of selection over models with the same training error. If this nonuniform probability agrees with the target function, early stopping may help. One should be aware of what nonuniform probability of selection is implied by the learning procedure.

When they studied early stopping, Wang et. al. [Wang et al., 1994] analyzed the average optimal stopping time for general linear models (one hidden layer neural networks with a linear output and fixed input weights) and introduced and examined the effective size of the learning machine as training proceeds. Sjoberg and Ljung [Sjoberg and Ljung, 1995] linked early stopping using a validation set to regularization, and showed that emphasizing the validation set too much may result in an unregularized solution. Amari et. al. [Amari et al., 1997] determined the best validation set size in the asymptotic limit and showed that even when this validation set size is used, early stopping using a validation set hurts for very large training sets. Dodier [Dodier, 1996] and Baldi and Chauvin [Baldi and Chauvin, 1991] investigated the behavior of validation error curves for linear problems, and the linear auto-association problem respectively.

The term *no free lunch* was introduced in [Wolpert, 1996a, Wolpert, 1996b]. Wolpert shows that when the prior distribution over the target functions is uniform, and the off-training-set error is taken to be the performance criterion, there is no difference between learning algorithms. In other words, if a learning algorithm results in good off-training-set error for one target function, it results in equally worse off-training-set error for another target function. Like [Zhu and Rohwer, 1996] and [Goutte, 1997] who put no-free-lunch theorems into the framework of cross validation, our work puts the no-free-lunch into the framework of early stopping.

Our method of early stopping –choosing a model uniformly among the models with the same training error– is similar to the Gibbs algorithm [Wolpert, 1995]. Although the uniform probability of selection around the training error minimum is equivalent to the isotropic distributions of [Amari et al., 1997], their work concentrates on very large number of training examples. Moreover, for general linear models we need the probability of selection of models to be only symmetric around the training error minimum, and symmetry is a weaker requirement than uniformity.

We are given a fixed training set $\{(\mathbf{x}_1, f_1), \ldots, (\mathbf{x}_N, f_N)\}$ with inputs $\mathbf{x}_n \in \mathcal{R}^{d'}$ and outputs $f_n \in \mathcal{R}$. The model to fit the training data will be denoted by $g_{\mathbf{v}}(\mathbf{x})$, with adjustable parameters $\mathbf{v}$. We will refer to models by their adjustable parameters $\mathbf{v}$, unless indicated otherwise. We assume that the training outputs were generated from the training inputs according to some unknown and fixed distribution $f(\mathbf{x}_n)$, hence $f_n = f(\mathbf{x}_n)$. For example, if the outputs were generated by a teacher model with parameters $\mathbf{v}^*$ and additive zero mean normal noise, we would have $f(\mathbf{x}_n) = g_{\mathbf{v}^*}(\mathbf{x}_n) + e_n$ where $e_n \sim \mathcal{N}(0, \sigma_e^2)$ for $\sigma_e^2 \geq 0$.

We define the quadratic training error E_T and the generalization error E at $\mathbf{v}$ as:

$$E_T(\mathbf{v}) = \frac{1}{N}\sum_{n=1}^{N}(g_{\mathbf{v}}(\mathbf{x}_n) - f_n)^2 \qquad\qquad E(\mathbf{v}) = \left\langle (g_{\mathbf{v}}(\mathbf{x}) - f(\mathbf{x}))^2 \right\rangle_{\mathbf{x}}$$

Let $\mathbf{v}_T$ be a local minimum of the training error E_T. Let $\delta \geq 0$ and $E_\delta = E_T(\mathbf{v}_T) + \delta$. Let $\mathbf{W}_\delta = \{\mathbf{\Delta v} : E_T(\mathbf{v}_T + \mathbf{\Delta v}) = E_\delta\}$. The set of models $\mathbf{v}_T + \mathbf{W}_\delta$ form the early stopping set. We define *early stopping at training error* E_δ as choosing a model from the early stopping set according to a probability distribution on the models in the early stopping set. We denote the probability of selecting $\mathbf{v}_T + \mathbf{\Delta v}$ as the early stopping solution by $P_{\mathbf{W}_\delta}(\mathbf{\Delta v})$. This probability is zero if $\mathbf{\Delta v} \notin \mathbf{W}_\delta$. The mean generalization error at training error level E_δ is:

$$E_{mean}(E_\delta) \quad = \quad \int\limits_{\mathbf{\Delta v} \in \mathbf{W}_\delta} P_{\mathbf{W}_\delta}(\mathbf{\Delta v}) E(\mathbf{v}_T + \mathbf{\Delta v}) d\mathbf{\Delta v}$$

$P_{\mathbf{W}_\delta}$ is said to be uniform if $\forall \mathbf{\Delta v}, \mathbf{\Delta v}' \in \mathbf{W}_\delta, \quad P_{\mathbf{W}_\delta}(\mathbf{\Delta v}) = P_{\mathbf{W}_\delta}(\mathbf{\Delta v}')$ – that is, if models with the same training error are equally likely to be chosen as the early stopping solution. (See Figure 1)

The rest of the article is organized as follows. In section 2, we prove that early stopping can not decrease the mean generalization error for general linear models when all models with the same training error are equally likely to be the target. Section 3 proves the same result for nonlinear models but around a training error minimum only. In all these cases, we assume that there is no prior information about the target that generated the training data. In section 4 we experimentally verify the early stopping results for general linear and neural network models. We also compare weight decay,

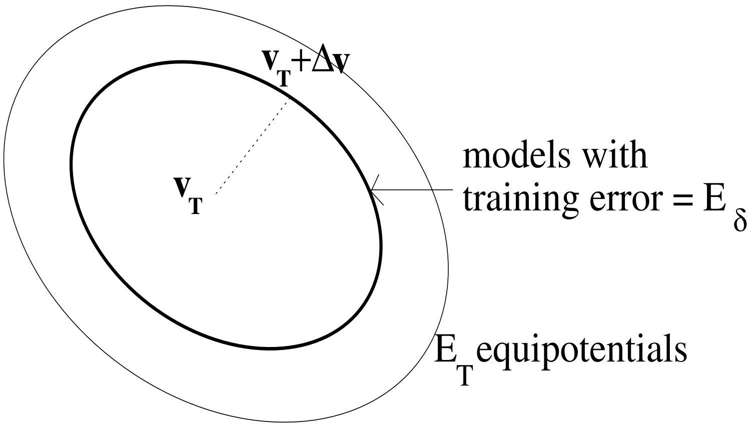
$\mathbf{v}_T+\Delta\mathbf{v}$
$\mathbf{v}_T$
models with training error = E_δ
E_T equipotentials

Figure 1: The models with training error $E_\delta = E_T(\mathbf{v}_T) + \delta$ form the early stopping set at training error level E_δ.

early stopping using a validation set and learning with additional prior information (hints) [Abu-Mostafa, 1994] to our framework and show that early stopping can help when certain additional information is available. Finally, section 5 summarizes the results.

2 Early Stopping for a General Linear Model

In this section we will consider the general linear models. Let $\phi_i(\mathbf{x}) : \mathcal{R}^{d'} \rightarrow \mathcal{R}$, $i = 0, \ldots, d$ be fixed transformation (basis) functions and let $\phi(\mathbf{x}) = [\phi_0(\mathbf{x}), \phi_1(\mathbf{x}), \ldots, \phi_d(\mathbf{x})]^T$. We define a general linear model as $g_{\mathbf{w}}(\mathbf{x}) = \mathbf{w}^T\phi(\mathbf{x})$ with fixed transformation functions $\phi(.)$ and adjustable parameters $\mathbf{w}$ (see Figure 2). If $\phi_0(\mathbf{x}) = 1$ and $\phi_i(\mathbf{x}) = x_i, 1 \leq i \leq d' = d$ we obtain the usual linear model; if $\phi_i(\mathbf{x}) = \prod_{j=1}^{d'} x_j^{k_j}, k_j \geq 0$ we obtain a polynomial model. The output of the general linear model is linear in the model parameters $\mathbf{w}$ and it can be nonlinear in the inputs $\mathbf{x}$. We will denote a general linear model only by the adjustable parameters $\mathbf{w}$.

Let $\mathbf{f}_{N\times 1} = [f_1, \ldots, f_N]^T$ be the training outputs. Let $\Phi_{x(d+1)\times N} = [\phi(\mathbf{x}_1), \ldots, \phi(\mathbf{x}_N)]$ denote the training inputs transformed by the fixed transformation functions. Define $\mathbf{S}_x = \frac{\mathbf{\Phi}_x\mathbf{\Phi}_x^T}{N}$ and $\Sigma_{\phi(x)} = \left\langle \phi(\mathbf{x})\phi(\mathbf{x})^T \right\rangle_{\mathbf{x}}$. When

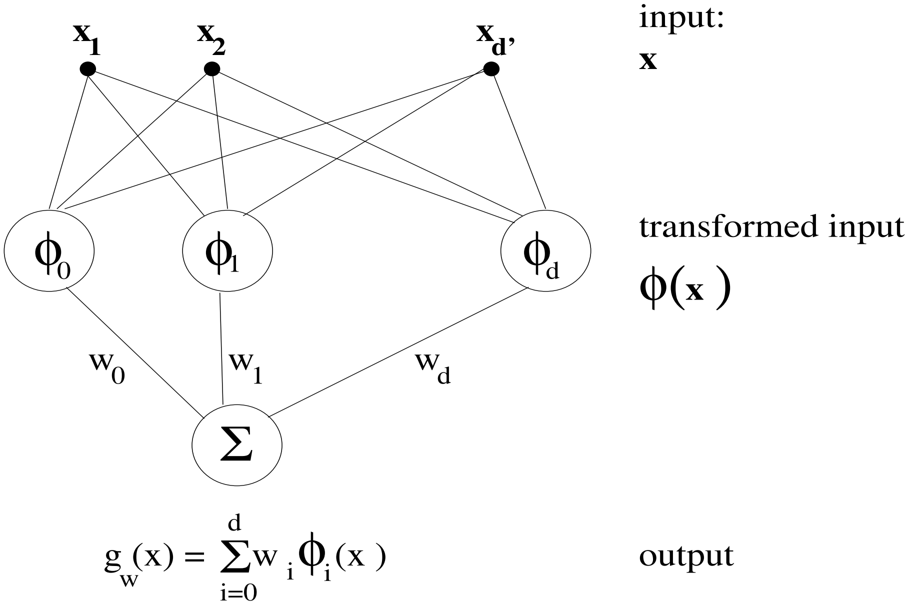

Figure 2: General linear model.

$\Phi_x \Phi_x^T$ is full rank [2], the unique training error minimum is given by the ordinary least squares solution:

$$\mathbf{w}_T \;\; = \;\; (\Phi_x \Phi_x^T)^{-1} \Phi_x \mathbf{f} = \mathbf{S}_x^{-1} \frac{\Phi_x \mathbf{f}}{N}$$

The Hessians of training and generalization errors are constant positive semidefinite[3] matrices at all $\mathbf{w}$:

$$HE_T(\mathbf{w}) = 2\mathbf{S}_x \qquad\qquad HE(\mathbf{w}) = 2\mathbf{\Sigma}_{\phi(x)}$$

Any higher derivatives of E and E_T are zero everywhere. Hence, for any $\mathbf{\Delta w}$, the generalization and training errors of $\mathbf{w}_T \pm \mathbf{\Delta w}$ can be written as:

$$E(\mathbf{w}_T \pm \mathbf{\Delta w}) \; = \; E(\mathbf{w}_T) \pm \mathbf{\Delta w}^T \nabla E(\mathbf{w}_T) + \mathbf{\Delta w}^T \mathbf{\Sigma}_{\phi(x)} \mathbf{\Delta w} \qquad (1)$$

[2] Hence we restrict ourselves to problems where $N \geq d+1$. When the transformation functions are real valued, for most cases $\mathbf{\Phi}_x \mathbf{\Phi}_x^T$ is likely to be full rank.

[3] Any matrix of the form AA^T is positive semidefinite, because for any $\mathbf{w}$ of proper dimensions, $\mathbf{w}^T AA^T \mathbf{w} \; = \; ||A^T \mathbf{w}||^2 \; \geq \; 0$, hence $\mathbf{S}_x \; = \; \frac{\Phi_x \Phi_x^T}{N}$ is positive semidefinite. $\mathbf{\Sigma}_{\phi(x)} = \left\langle \phi(\mathbf{x}) \phi(\mathbf{x})^T \right\rangle_{\mathbf{x}}$ is also positive semidefinite since $\frac{\Phi_x \Phi_x^T}{N} \longrightarrow_{N \to \infty} \left\langle \phi(\mathbf{x}) \phi(\mathbf{x})^T \right\rangle_{\mathbf{x}}$

$$E_T(\mathbf{w}_T \pm \mathbf{\Delta w}) \,=\, E_T(\mathbf{w}_T) + \mathbf{\Delta w}^T \mathbf{S}_x \mathbf{\Delta w} \tag{2}$$

The following lemma proves that when all models with the training error $E_T(\mathbf{w}_T) + \delta, \delta \geq 0$ are equally likely to be chosen as the solution, the mean generalization error at training error level $E_T(\mathbf{w}_T) + \delta$ can not be smaller than the generalization error of the training error minimum $\mathbf{w}_T$.

Lemma 1: *When all models with training error $E_\delta = E_T(\mathbf{w}_T) + \delta \geq E_T(\mathbf{w}_T)$ are equally likely to be chosen as the early stopping solution, the mean generalization error at training error level $E_T(\mathbf{w}_T) + \delta$ is at least as much as the generalization error of the training error minimum. More specifically, for any $\delta \geq 0$, $E_{mean}(E_\delta) = E(\mathbf{w}_T) + \beta(\delta)$, for some $\beta(\delta) \geq 0$.*

The proof is given in appendix A. See Figure 3 for an illustration of the lemma.

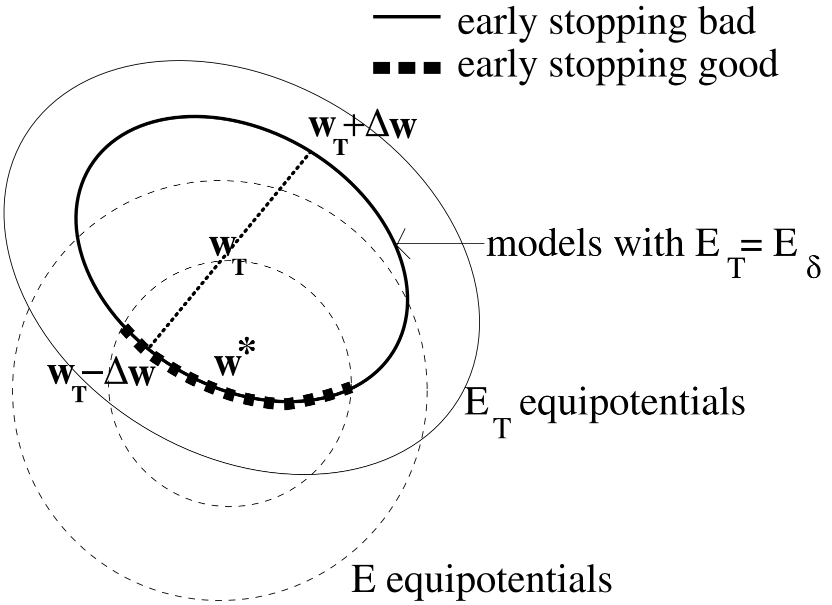

Figure 3: Early stopping at a training error δ above $E_T(\mathbf{w}_T)$ results in higher generalization error when all models having the same training error are equally likely to be chosen as the early stopping solution.

This result does not depend on the noise level, number of training examples, or the target function versus model complexity. Even if the target function is a constant and the model is a 100th degree polynomial, lemma 1 tells us that we should stop only at the training error minimum.

If the error criterion is the test error on independently and identically distributed (i.i.d.) or non-i.i.d. inputs $\{\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_M\}$, the lemma still holds. Because $\mathbf{S}_{\tilde{x}} = \frac{\Phi_{\tilde{x}} \Phi_{\tilde{x}}^T}{M}$ is positive semidefinite.

Furthermore, lemma 1 holds not only for quadratic loss, but for any loss function which has a positive semidefinite test error Hessian and small enough third and higher derivatives at the training error minimum.

The following theorem compares the mean generalization error between any two training error levels:

Theorem 1: *When all models with the same training error are equally likely to be chosen as the early stopping solution, the mean generalization error is an increasing function of the early stopping training error. In other words, for* $0 < \delta_1 < \delta_2$, $E_{mean}(E_{\delta_1}) < E_{mean}(E_{\delta_2})$.

The proof is given in appendix B.

Therefore, when the model is general linear, the best strategy is to minimize the training error as much as possible.

3 Early Stopping for a Nonlinear Model

When the model is general linear we are able to prove lemma 1 without any assumptions about the location of the generalization error minimum with respect to the training error minimum. Also, our results are valid for all models with the same training error, regardless of how far they are from the training error minimum. For the nonlinear model we will assume that the distance between the training error minimum and the generalization error minimum is $\mathcal{O}\left(\frac{1}{N}\right)$, which asymptotically is the case (see e.g., [Amari et al., 1997]). Also we will prove the increase in the mean generalization error only around the training error minimum.

Let the model $g_{\mathbf{v}}$ be a nonlinear (continuous and differentiable) model with adjustable parameters $\mathbf{v}$. Let $\mathbf{v}_T$ be a minimum of the training error, and let $\mathbf{v}^*$ be a minimum of the generalization error.

Now we assert the counterpart of lemma 1 for the nonlinear models: **Theorem 2:** *Let* $\mathbf{v}_T - \mathbf{v}^* = \mathcal{O}\left(\frac{1}{N^{0.5}}\right)$, $\mathbf{\Delta v} = \mathcal{O}\left(\frac{1}{N^{0.5}}\right)$, $\delta \geq 0$ *and* $\delta = \mathcal{O}\left(\frac{1}{N}\right)$. *Let* $E_\delta = E_T(\mathbf{v}_T) + \delta + \mathcal{O}\left(\frac{1}{N^{1.5}}\right)$. *When all models with training error* E_δ *are equally likely to be chosen as the early stopping solution, their mean generalization error is* $E_{mean}(E_\delta) = E(\mathbf{v}_T) + \beta(\delta) + \mathcal{O}\left(\frac{1}{N^{1.5}}\right)$, *for some* $\beta(\delta) \geq 0$.

The proof is given in appendix C.

4 Weight Decay, Early Stopping Using a Validation Set and Hints

We see from lemma 1 and theorem 2 that if all models with a given training error are chosen with equal probability (density) then no strategy beats the strategy of choosing the training error minimum. We emphasize that the only assumption required for the proof of the theorem is that the models with the same training error be chosen with equal probability [4]. We make no assumptions on the input probability distribution, target function, or presence or nature of the noise. This is a strikingly general statement especially given the plethora of evidence in favor of methods picking a solution other than the training error minimum [Reed, 1993]. It must therefore be the case that these algorithms are violating the assumptions of our theorem; some models with a given training error are chosen with higher probability than others.

First we establish that the commonly used regularization techniques do not choose uniformly among models with a given training error. This is easy to see for weight-decay-type regularizers. Given two weight vectors with the same training error, the model with the smaller weights is favored. In this way, models with lower complexity are favored. Early stopping works in a similar way [Sjoberg and Ljung, 1995]. From the data set one picks a training set and the remaining data points are used as a validation set. Along the path from the starting point of the training algorithm to the training set minimum, one picks the weights that obtain a minimim for the validation set error. The key observation is that the training algorithm usually starts at small weights. This means that if the validation set minimum happens to

[4] In fact for the proof we actually only need symmetry.

have smaller weights than the training set minimum (roughly half the time [Amari et al., 1997]), then the final solution will have smaller weights. If the validation set minimum happens to have larger weights than the training minimum (roughly half the time) then the final solution will be the training minimum because of the direction of approach. Averaging over possible training sets, the training set minimum will average to the minimum of the entire training set, therefore we see that on average the solution will have smaller weights than the entire training set solution, much like a weight decay type regularizer. Thus once again we see that the algorithm favors smaller weights (less complex functions)[5]. Thus we see that the assumption of the theorem is being violated. What remains is to see that it is being violated in a way that favors the right models. In real data where noise is usually present, the data represents a function that is more complex than the target function. Thu,s given two models with the same training error, the less complex one should be favored.

We have given an intuitive explanation as to why regularizing algorithms tend to work, and how they are violating the no-free-lunch theorem we have proved. We would like to end on a more general note on the use of prior information such as hints and invariances that are known ahead of time about the target function. By starting at small weights or using regularization, we are enforcing a prior about the learning problem: that noise is present and so the data alone represents too complex a function. In general one should incorporate all the prior information into the objective function and then minimize that objective function. This is usually done in a Bayesian framework. If one has no prior information, then all models yielding the same training error should be equally likely and we are in the world of our no-free-lunch theorem. Thus, we see that in order to get better performance than the training error minimum, it is necessary to incorporate some prior information into the learning process. It is in this sense that our theorem is a no-free-lunch theorem.

4.1 Experiments

We experimented with linear and nonlinear models to verify our results.

[5] If one in addition averages over possible starting points for the training algorithm as well, then this would remove the asymetry and the theorem would apply. Thus we see that the key to these early stopping algorithms is in fact the use of small weights for the initial starting point.

4.1.1 Linear Model

We computed the minimum training error (least squares) solution $\mathbf{w}_T$; then we computed the average generalization error of solutions $\mathbf{w}$ with training error $E_T(\mathbf{w}_T)+\delta$. For comparison, we also computed the generalization error of the weight decay solution with training error $E_T(\mathbf{w}_T)+\delta$. In Figure 4[6] we show the behavior of the mean generalization error as the training error increases. When all models with the same training error are chosen with the same probability, in agreement with lemma 1, the mean generalization error increases as the training error increases. On the other hand, the weight decay solution has smaller generalization error for a small enough weight decay parameter. Note that choosing the weight decay solution with probability 1 corresponds to a nonuniform (delta function) probability distribution on models with the same training error, therefore lemma 1 does not apply. Note also that, for this experiment, both the target and the model are linear and the training points have zero mean normal noise, therefore, the weight decay provably results in better generalization error when the weight decay parameter is small enough [Bishop, 1995].

4.1.2 Nonlinear Model

We experimented with a neural network model, and a noisy and even target function, also generated by a (teacher) neural network model. We first found a training error minimum using the gradient descent with adaptive learning rate. Then we chose random weights $\boldsymbol{\Delta}\mathbf{v}$[7] such that $E_T(\mathbf{v}_T+\boldsymbol{\Delta}\mathbf{v}) \approx E_T(\mathbf{v}_T)+\delta$. In Figure 5[8] we show the mean test error

[6] For this experiment, both the target and the model were linear. Input dimensionality was $d = 5$, plus constant bias 1. Training inputs were chosen from a zero mean unit normal. There were $N = 20$ training input-outputs. The target (teacher) model was also linear with weights chosen from zero mean 9 variance normal. Zero mean normal noise was added to the training outputs. Noise variance was determined according to 0.1 signal-to-noise ratio. The mean generalization/test error for the uniform P was computed on 500 different models with the same training error. The generalization/test error was computed as the squared distance between the target and the model.

[7] Since the gradient at the minimum $\mathbf{v}_T$ is very small but not exactly zero, we scaled $\boldsymbol{\Delta}\mathbf{v}$ as $k\boldsymbol{\Delta}\mathbf{v}$ where k is the best possible solution for $k\boldsymbol{\Delta}\mathbf{v}^T\nabla E_T(\mathbf{v}_T)+k^2\frac{1}{2}\boldsymbol{\Delta}\mathbf{v}^T H E_T(\mathbf{v}_T)\boldsymbol{\Delta}\mathbf{v} = \delta$. Hence $k = \frac{-b\pm\sqrt{b^2+4a\delta}}{2a}$ where $a = \frac{1}{2}\boldsymbol{\Delta}\mathbf{v}^T H E_T(\mathbf{v}_T)\boldsymbol{\Delta}\mathbf{v}$ and $b = \boldsymbol{\Delta}\mathbf{v}^T\nabla E_T(\mathbf{v}_T)$.

[8] The training outputs were generated by (teacher) neural network whose weights were drawn from unit normal. First a neural network with 5 hidden units was generated. Then the function was made even by adding five more hidden units with exactly the

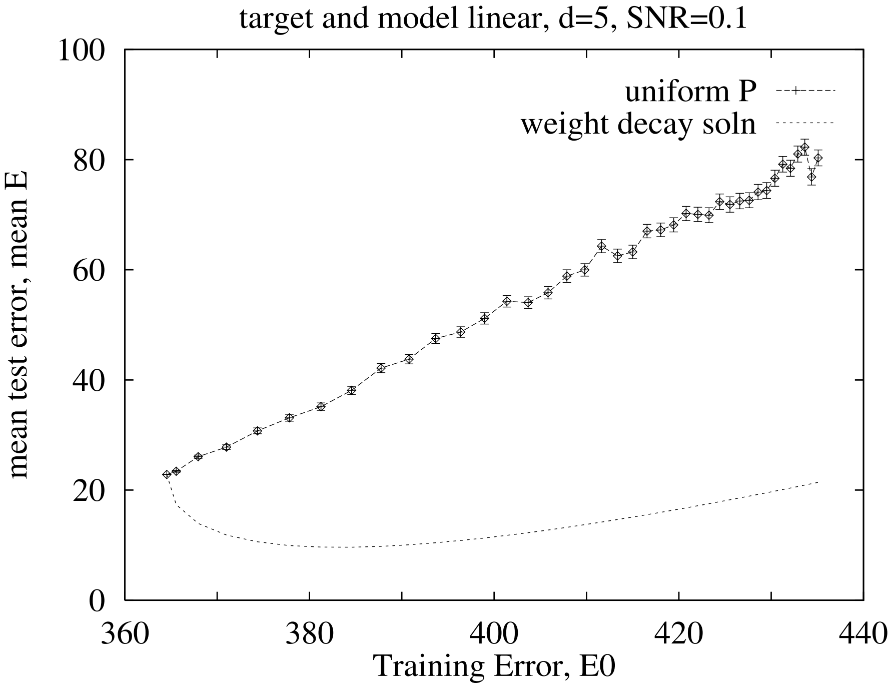

Figure 4: The mean generalization/test error versus training error of a linear model for a given target and training set. The mean generalization error increases as the training error increases when all models with the same training error are given equal probability of selection. When the weight decay parameter is small enough, choosing the weight decay solution with probability 1 and all other models with the same training error with probability 0 improves the generalization error.

versus the training error for a specific target, training set and model $g_{\mathbf{v}_T}$. When the mean test error for a certain training error level is computed by giving each model with the same training error equal probability, the mean test error increases. On the other hand, when the models with smaller evenness hint error $E_1(\mathbf{v}_T + \mathbf{\Delta v})$ are given more weight, the mean test error seems to decrease for sometime and then increase. In other words, early stopping, choosing models with smaller hint errors with higher probability can decrease the mean test error.

Note that, as shown in Figure 6, the decrease in the mean test error using the hint is dependent on not only the number of training examples N, but also the signal-to-noise ratio. For the same N, but now for $SNR = 10$, selecting the models according to the evenness hint error, in the same way we did for the previous experiment that had $SNR = 0.01$, does not decrease the mean test error. It is possible that the probability of selection of a model should depend not only on the hint error E_1, but also the level of training error and the signal-to-noise ratio.

5 Conclusions

We analyzed early stopping at a certain training error minimum, and showed that one should minimize the training error as much as possible when all the information available about the target is the training set. We also demonstrated that when additional information is available, early stopping can help.

same connections, except negative of the input weights of the first five hidden units. The training and test inputs were drawn from a zero mean and variance 10 normal. The training outputs were obtained by adding zero mean noise to the teacher outputs on the training inputs. The noise variance was determined according to the signal-to-noise ratio. The test outputs were not noisy. There were $N = 30$ training and $M = 50$ test examples. The student (model) neural network had 10 hidden units, and its weights were drawn from a zero mean 0.001 variance normal. The training method was gradient descent. The learning rate was initially 0.0001, during training, it was multiplied by 1.1 when the training error decreased and halved otherwise. Training continued for 1000 passes and the model with the smallest training error was taken to be $g_{\mathbf{v}_T}$. When computing the mean test error using the evenness hint [Abu-Mostafa, 1994], we weighed the model $g_{\mathbf{v}_T + \mathbf{\Delta v}^i}$ according to: $\frac{\exp - E_1(\mathbf{v}_T + \mathbf{\Delta v}^i)}{\sum_{i=1}^{1000} \exp - E_1(\mathbf{v}_T + \mathbf{\Delta v}^i)}$ for $i = 1, \ldots, 1000$.

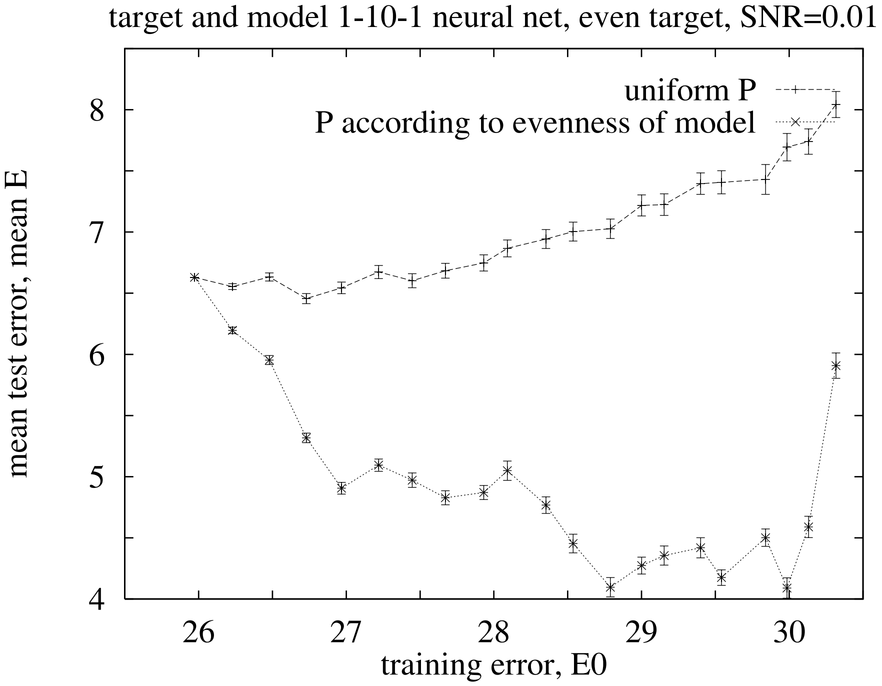

Figure 5: The mean test error versus training error of a nonlinear model for a given even target and training set. The mean test error increases as the training error increases when all models with the same training error are given equal probability of selection. Choosing the models with the smaller evenness error with higher probability reduces the mean test error.

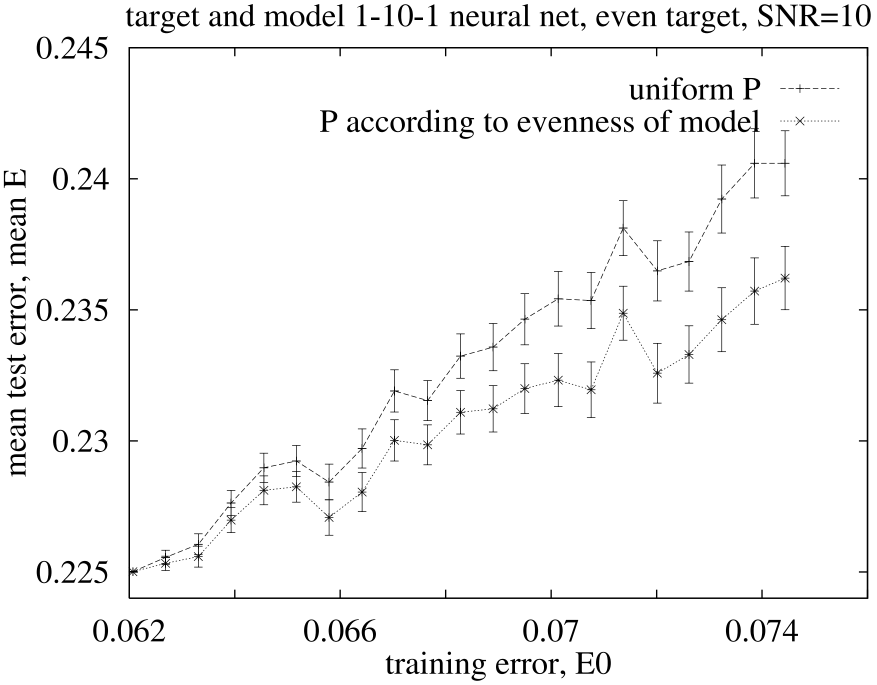

Figure 6: When the signal-to-noise ratio is high and the target is even, even if the models with the same training error are weighed according to their hint error, the mean test error around the training error minimum may increase.

A Proof of Lemma 1:

Let the early stopping training error level be $E_\delta = E_T(\mathbf{w}_T) + \delta$ for some $\delta \geq 0$. Then, from equation 2, the early stopping set consists of $\mathbf{w}_T + \mathbf{W}_\delta = \mathbf{w}_T + \{\boldsymbol{\Delta}\mathbf{w} : \boldsymbol{\Delta}\mathbf{w}^T \mathbf{S}_x \boldsymbol{\Delta}\mathbf{w} = \delta\}$. The mean generalization error is:

$$E_{mean}(E_\delta) = \int\limits_{\boldsymbol{\Delta}\mathbf{w} \in \mathbf{W}_\delta} P_{\mathbf{W}_\delta}(\boldsymbol{\Delta}\mathbf{w}) E(\mathbf{w}_T + \boldsymbol{\Delta}\mathbf{w}) d\boldsymbol{\Delta}\mathbf{w}$$

For any $\boldsymbol{\Delta}\mathbf{w} \in \mathbf{W}_\delta$, hence satisfying $\boldsymbol{\Delta}\mathbf{w}^T \mathbf{S}_x \boldsymbol{\Delta}\mathbf{w} = \delta$, there exists a $-\boldsymbol{\Delta}\mathbf{w} \in \mathbf{W}_\delta$, therefore we can rewrite the mean generalization error as:

$$\begin{aligned} E_{mean}(E_\delta) &= 0.5 \int\limits_{\boldsymbol{\Delta}\mathbf{w} \in \mathbf{W}_\delta} (P_{\mathbf{W}_\delta}(\boldsymbol{\Delta}\mathbf{w}) E(\mathbf{w}_T + \boldsymbol{\Delta}\mathbf{w}) \\ &\qquad + P_{\mathbf{W}_\delta}(-\boldsymbol{\Delta}\mathbf{w}) E(\mathbf{w}_T - \boldsymbol{\Delta}\mathbf{w}))\, d\boldsymbol{\Delta}\mathbf{w} \end{aligned}$$

Now, since $P_{\mathbf{W}_\delta}$ is uniform, it is also symmetric, i.e. $P_{\mathbf{W}_\delta}(\boldsymbol{\Delta}\mathbf{w}) = P_{\mathbf{W}_\delta}(-\boldsymbol{\Delta}\mathbf{w})$. For the proof of this lemma symmetry is the only restriction we need on $P_{\mathbf{W}_\delta}$. Using symmetry of $P_{\mathbf{W}_\delta}$, equation 1, and the fact that $\int\limits_{\boldsymbol{\Delta}\mathbf{w} \in \mathbf{W}_\delta} P_{\mathbf{W}_\delta}(\boldsymbol{\Delta}\mathbf{w}) d\boldsymbol{\Delta}\mathbf{w} = 1$:

$$\begin{aligned} E_{mean}(E_\delta) &= E(\mathbf{w}_T) + \int\limits_{\boldsymbol{\Delta}\mathbf{w} \in \mathbf{W}_\delta} P_{\mathbf{W}_\delta}(\boldsymbol{\Delta}\mathbf{w}) \boldsymbol{\Delta}\mathbf{w}^T \boldsymbol{\Sigma}_{\phi(x)} \boldsymbol{\Delta}\mathbf{w} d\boldsymbol{\Delta}\mathbf{w} \\ &= E(\mathbf{w}_T) + \beta(\delta) \end{aligned}$$

Since $\boldsymbol{\Sigma}_{\phi(x)} = \left\langle \phi(\mathbf{x})\phi(\mathbf{x})^T \right\rangle_{\mathbf{x}}$ is positive semidefinite and $P_{\mathbf{W}_\delta}(\boldsymbol{\Delta}\mathbf{w}) \geq 0$,

$$\beta(\delta) = \int\limits_{\boldsymbol{\Delta}\mathbf{w} \in \mathbf{W}_\delta} P_{\mathbf{W}_\delta}(\boldsymbol{\Delta}\mathbf{w}) \boldsymbol{\Delta}\mathbf{w}^T \boldsymbol{\Sigma}_{\phi(x)} \boldsymbol{\Delta}\mathbf{w} d\boldsymbol{\Delta}\mathbf{w} \geq 0 \qquad (3)$$

□

B Proof of Theorem 1:

By lemma 1, $E_{mean}(E_{\delta_1}) = E(\mathbf{w}_T) + \beta(\delta_1)$ and $E_{mean}(E_{\delta_2}) = E(\mathbf{w}_T) + \beta(\delta_2)$ for $\beta(\delta_1), \beta(\delta_2) > 0$. Let $0 < \delta_1 < \delta_2$. We need to prove $\beta(\delta_1) < \beta(\delta_2)$.

Let $V(\delta) = \int\limits_{\mathbf{\Delta w} \in \mathbf{W}_\delta} \mathbf{\Delta w}^T \mathbf{\Sigma}_{\phi(x)} \mathbf{\Delta w} d\mathbf{\Delta w}$, and let $\frac{1}{P_\delta}$ be the surface area of the d-dimensional ellipsoid $\mathbf{\Delta w}^T \mathbf{S}_x \mathbf{\Delta w} = \delta$. Since $P_{\mathbf{W}_\delta}$ is uniform, from equation 3:

$$\frac{\beta(\delta_2)}{\beta(\delta_1)} = \frac{P_{\delta_2}}{P_{\delta_1}} \frac{V(\delta_2)}{V(\delta_1)}$$

Define $k^2 = \frac{\delta_2}{\delta_1} > 1$. Let $\mathbf{W}_{\delta_1} = \{\mathbf{\Delta w} : \mathbf{\Delta w}^T \mathbf{S}_x \mathbf{\Delta w} = \delta_1\}$. Then $\mathbf{W}_{\delta_2} = \{k\mathbf{\Delta w} : \mathbf{\Delta w} \in \mathbf{W}_{\delta_1}\}$. By means of change of variables $\mathbf{\Delta u} = k\mathbf{\Delta w}$ in $V(\delta_2)$ we have $\frac{V(\delta_2)}{V(\delta_1)} = k^{d+1}$.

We can define the surface area as the derivative of the volume:

$$\begin{aligned}
\frac{1}{P_\delta} &= \lim_{l \to 0} \frac{\int\limits_{\mathbf{\Delta w}^T \mathbf{S}_x \mathbf{\Delta w} \leq \delta + l} d\mathbf{\Delta w} - \int\limits_{\mathbf{\Delta w}^T \mathbf{S}_x \mathbf{\Delta w} \leq \delta} d\mathbf{\Delta w}}{l} \\
&= \lim_{l \to 0} \frac{\left(\frac{\delta + l}{\delta}\right)^{\frac{h+1}{2}} - 1}{l} \int\limits_{\mathbf{\Delta w}^T \mathbf{S}_x \mathbf{\Delta w} \leq \delta} d\mathbf{\Delta w} \\
&= \frac{h+1}{2\delta} \int\limits_{\mathbf{\Delta w}^T \mathbf{S}_x \mathbf{\Delta w} \leq \delta} d\mathbf{\Delta w}
\end{aligned}$$

Hence $\frac{1}{P_{\delta_1}} = \frac{h+1}{2\delta_1} \int\limits_{\mathbf{\Delta w}^T \mathbf{S}_x \mathbf{\Delta w} \leq \delta_1} d\mathbf{\Delta w}$. By means of change of variables $\mathbf{\Delta u} = \frac{\mathbf{\Delta w}}{k}$ we have $\frac{1}{P_{\delta_2}} = k^{d-1} \frac{1}{P_{\delta_1}}$. Therefore, $\frac{P_{\delta_2}}{P_{\delta_1}} = k^{-d+1}$.

Hence, $\frac{\beta(\delta_2)}{\beta(\delta_1)} = k^{-d+1} k^{d+1} = k^2 > 1$. □

C Proof of Theorem 2:

Let $\nabla E(\mathbf{v}_T), \nabla E_T(\mathbf{v}_T), HE(\mathbf{v}_T), HE_T(\mathbf{v}_T)$ denote the gradient and Hessians of the generalization error and the training error at the training error minimum $\mathbf{v}_T$.

Similar to equations 1 and 2, the training and generalization errors at $\mathbf{v}_T + \mathbf{\Delta v}$ are:

$$\begin{aligned}
E(\mathbf{v}_T \pm \mathbf{\Delta v}) = {} & E(\mathbf{v}_T) \pm \mathbf{\Delta v}^T \nabla E(\mathbf{v}_T) \\
& + \frac{1}{2} \mathbf{\Delta v}^T HE(\mathbf{v}_T) \mathbf{\Delta v} + \mathcal{O}\left(\frac{1}{N^{1.5}}\right) \qquad (4)
\end{aligned}$$

$$E_T(\mathbf{v}_T \pm \mathbf{\Delta v}) \; = \; E_T(\mathbf{v}_T) + \frac{1}{2}\mathbf{\Delta v}^T H E_T(\mathbf{v}_T)\mathbf{\Delta v} + \mathcal{O}\left(\frac{1}{N^{1.5}}\right) \qquad (5)$$

Since $\mathbf{v}_T = \mathbf{v}^* + \mathcal{O}\left(\frac{1}{N^{0.5}}\right)$:

$$HE\left(\mathbf{v}_T\right) \;\; = \;\; HE\left(\mathbf{v}^* + \mathcal{O}\left(\frac{1}{N^{0.5}}\right)\right) = HE\left(\mathbf{v}^*\right) + \mathcal{O}\left(\frac{1}{N^{0.5}}\right)$$

Using the fact that $\mathbf{\Delta v} = \mathcal{O}\left(\frac{1}{N^{0.5}}\right)$, and equation 4, we can write the average generalization error among $\mathbf{v}_T + \mathbf{\Delta v}$ and $\mathbf{v}_T - \mathbf{\Delta v}$ as:

$$\frac{E(\mathbf{v}_T + \mathbf{\Delta v}) + E(\mathbf{v}_T - \mathbf{\Delta v})}{2} \;\; = \;\; E(\mathbf{v}_T) + \frac{1}{2}\mathbf{\Delta v}^T HE(\mathbf{v}^*)\mathbf{\Delta v} + \mathcal{O}\left(\frac{1}{N^{1.5}}\right)$$

Define $\mathbf{W}_\delta = \{\mathbf{\Delta v} : E_T(\mathbf{v}_T + \mathbf{\Delta v}) = E_T(\mathbf{v}_T) + \delta + \mathcal{O}\left(\frac{1}{N^{1.5}}\right)\}$. (Hence $\delta = \mathcal{O}\left(\frac{1}{N}\right)$.) For each $\mathbf{\Delta v} \in \mathbf{W}_\delta$, there is a $-\mathbf{\Delta v} \in \mathbf{W}_\delta$. As we did for the proof of lemma 1, using the uniform probability of selection $P_{\mathbf{W}_\delta}$, we can compute the mean generalization error as:

$$\begin{aligned}
E_{mean}(E_\delta) &= \int\limits_{\mathbf{\Delta v} \in \mathbf{W}_\delta} P_{\mathbf{W}_\delta}(\mathbf{\Delta v}) E(\mathbf{v}_T + \mathbf{\Delta v}) d\mathbf{\Delta v} \\
&= 0.5 \int\limits_{\mathbf{\Delta v} \in \mathbf{W}_\delta} \left(P_{\mathbf{W}_\delta}(\mathbf{\Delta v}) \; E(\mathbf{v}_T + \mathbf{\Delta v})\right. \\
&\qquad\qquad \left. + P_{\mathbf{W}_\delta}(-\mathbf{\Delta v}) E(\mathbf{v}_T - \mathbf{\Delta v})\right) d\mathbf{\Delta v} \\
&= E(\mathbf{v}_T) + 0.5 \int\limits_{\mathbf{\Delta v} \in \mathbf{W}_\delta} P_{\mathbf{W}_\delta}(\mathbf{\Delta v}) \mathbf{\Delta v}^T HE(\mathbf{v}^*) \mathbf{\Delta v} d\mathbf{\Delta v} \\
&\quad + \mathcal{O}\left(\frac{1}{N^{1.5}}\right) \\
&= E(\mathbf{v}_T) + \beta(\delta) + \mathcal{O}\left(\frac{1}{N^{1.5}}\right)
\end{aligned}$$

Since $\mathbf{v}^*$ is the generalization error minimum, $HE(\mathbf{v}^*)$ is positive semidefinite. Hence, $\beta(\delta) = 0.5 \int\limits_{\mathbf{\Delta v} \in \mathbf{W}_\delta} P_{\mathbf{W}_\delta}(\mathbf{\Delta v}) \mathbf{\Delta v}^T HE(\mathbf{v}^*) \mathbf{\Delta v} d\mathbf{\Delta v} \geq 0$. □

Acknowledgements

We thank members of the Caltech Learning Systems Group: Amir Atiya, Alexander Nicholson, Joseph Sill and Xubo Song for many useful discussions, and to two anonymous referees for their comments that improved the presentation of this paper.

References

[Abu-Mostafa, 1994] Abu-Mostafa, Y. (1994). Learning from hints. *Journal of Complexity*, 10:165–178.

[Amari et al., 1997] Amari, S., Murata, N., Muller, K., Finke, M., and Yang, H. H. (1997). Asymptotic statistical theory of overtraining and cross-validation. *IEEE Transactions on Neural Networks*, 8(5):985–996.

[Baldi and Chauvin, 1991] Baldi, P. and Chauvin, Y. (1991). Temporal evolution of generalization during learning in linear networks. *Neural Computation*, 3:589–603.

[Bishop, 1995] Bishop, C. (1995). *Neural Networks for Pattern Recognition.* Clarendon Press, Oxford.

[Dodier, 1996] Dodier, R. (1996). Geometry of early stopping in linear networks. In Touretzky, D. S., Mozer, M. C., and Hasselmo, M. E., editors, *Advances in Neural Information Processing Systems*, volume 8, pages 365–371. The MIT Press, Cambridge.

[Goutte, 1997] Goutte, C. (1997). Note on free lunches and cross-validation. *Neural Computation*, 9(6):1053–1059.

[Reed, 1993] Reed, R. (1993). Pruning algorithms – a survey. *IEEE Transactions on Neural Networks*, 4(5):740–747.

[Sjoberg and Ljung, 1995] Sjoberg, J. and Ljung, L. (1995). Overtraining, regularization, and searching for a minimum, with application to neural networks. *International Journal of Control*, 62(6):1391–1407.

[Wang et al., 1994] Wang, C., Venkatesh, S. S., and Judd, J. S. (1994). Optimal stopping and effective machine complexity in learning. In Cowan, J. D., Tesauro, G., and Alspector, J., editors, *Advances in Neural Information Processing Systems*, volume 6, pages 303–310. Morgan Kaufmann, San Francisco.

[Wolpert, 1995] Wolpert, D. H. (1995). *The Mathematics of Generalization, the Proceedings of the SFI/CNLS Workshop on Formal Approaches to Supervised Learning.* Addison Wesley, Reading, MA.

[Wolpert, 1996a] Wolpert, D. H. (1996a). The existence of A priori distinctions between learning algorithms. *Neural Computation*, 8(7):1391–1420.

[Wolpert, 1996b] Wolpert, D. H. (1996b). The lack of A priori distinctions between learning algorithms. *Neural Computation*, 8(7):1341–1390.

[Zhu and Rohwer, 1996] Zhu, H. and Rohwer, R. (1996). No free lunch for cross-validation. *Neural Computation*, 8(7):1421–1426.