

No Free Lunch for Early Stopping

Zehra Cataltepe, Yaser S. Abu-Mostafa, Malik Magdon-Ismail
Learning Systems Group
California Institute of Technology
Pasadena, CA 91125
{zehra, yaser, magdon}@cs.caltech.edu

Caltech CS-TR-98-02*

January 19, 1998

Abstract

We show that, with a uniform prior on hypothesis functions 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. We also show that regularization methods are equivalent to early stopping with certain non-uniform prior on the early stopping solutions.

1 Introduction

Early stopping of training is one of the methods that aims to prevent over-training due to powerful hypothesis class, noisy training examples or small training set. Early stopping has been studied by Wang et. al. [9] who analyzed the average optimal stopping time for generalized-linear hypotheses and introduced and examined the effective size of the learning machine as training proceeds. Sjoberg and Ljung [8] 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. [2] determined the best validation set size in the asymptotic limit and showed that early stopping helps little in this limit even when the best stopping point

*Copyright@1998 Zehra Cataltepe, Available at <ftp://ftp.cs.caltech.edu/tr/cs-tr-98-02>

is known. Dodier [6] and Baldi and Chauvin [3] investigated the behavior of validation curves for linear problems, and the linear auto-association problem respectively.

In this paper, we study early stopping at a predetermined training error level. If there is no prior information, other than the training examples, all hypotheses 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 generalized-linear hypotheses, early stopping at any training error level above the training error minimum increases the expected generalization error. For general hypotheses, the same result holds, but only within a small enough neighborhood of the training error minimum. For classification problems and the bin model [1], the expected generalization error increases regardless of the probability of selection of hypotheses. Regularization methods such as weight decay [4] and early stopping using a validation set are equivalent to early stopping at a fixed training error level with a non-uniform probability of selection over hypotheses with the same training error.

We will use the following notation and definitions in this paper. We are given a training set $D = \{\mathbf{x}_i, y_i\}_{i=1}^N$ with inputs $\mathbf{x}_i \in \mathcal{R}^d$ and outputs $y_i \in \mathcal{R}$. The outputs are generated according to $y_i = f(\mathbf{x}_i) + e_i$ where input and noise densities may be unknown. The hypotheses that are used to learn the data D are $g_{\mathbf{w}}(\mathbf{x})$, with adjustable parameters \mathbf{w} . We will refer to hypotheses by their adjustable parameters \mathbf{w} , unless indicated otherwise.

We define the training error E_0 and the generalization error E at \mathbf{w} as:

$$E_0(\mathbf{w}) = \frac{1}{N} \sum_{n=1}^N (g_{\mathbf{w}}(\mathbf{x}_n) - y_n)^2 \quad E(\mathbf{w}) = \left\langle (g_{\mathbf{w}}(\mathbf{x}) - f(\mathbf{x}))^2 \right\rangle_{\mathbf{x}, \mathbf{e}}$$

Let \mathbf{w}_0 be a local minimum of the training error E_0 . Let $\delta \geq 0$ and $E_\delta = E_0(\mathbf{w}_0) + \delta$. Let $\mathbf{W}_\delta = \{\Delta\mathbf{w} : E_0(\mathbf{w}_0 + \Delta\mathbf{w}) = E_\delta\}$. The set of hypotheses $\mathbf{w}_0 + \mathbf{W}_\delta$ form the **early stopping set**. We define **early stopping at training error E_δ** as choosing a hypothesis from the early stopping set. We denote the probability of selecting $\mathbf{w}_0 + \Delta\mathbf{w}$ as the early stopping solution by $P_{\mathbf{W}_\delta}(\Delta\mathbf{w})$. This probability is zero if $\Delta\mathbf{w} \notin \mathbf{W}_\delta$. The **mean generalization error at training error level E_δ** is:

$$E_{mean}(E_\delta) = \int_{\Delta\mathbf{w} \in \mathbf{W}_\delta} P_{\mathbf{W}_\delta}(\Delta\mathbf{w}) E(\mathbf{w}_0 + \Delta\mathbf{w}) d\Delta\mathbf{w}$$

$P_{\mathbf{W}_\delta}$ is said to be **uniform** if $\forall \Delta \mathbf{w}, \Delta \mathbf{w}' \in \mathbf{W}_\delta, P_{\mathbf{W}_\delta}(\Delta \mathbf{w}) = P_{\mathbf{W}_\delta}(\Delta \mathbf{w}')$, i.e. if hypotheses with the same training error are equally likely to be chosen as the early stopping solution.

In sections 2 and 3 we study early stopping at a predetermined training error level with uniform $P_{\mathbf{W}_\delta}$, for generalized-linear and general hypotheses respectively. Section 4 analyzes early stopping for classification problems and the bin model. Section 5 relates early stopping using a validation set and weight decay to our framework and points to possible research directions.

2 Generalized-Linear Hypotheses

Let $\phi(\mathbf{x}) = [\phi_0(\mathbf{x}), \phi_1(\mathbf{x}), \dots, \phi_h(\mathbf{x})]^T$ where $\phi_i(\mathbf{x}) : \mathcal{R}^d \rightarrow \mathcal{R}$ are basis functions and $\langle \phi(\mathbf{x})\phi(\mathbf{x})^T \rangle_{\mathbf{x}}$ exists. We define generalized-linear hypotheses as $g_{\mathbf{w}}(\mathbf{x}) = \mathbf{w}^T \phi(\mathbf{x})$ with fixed $\phi(\cdot)$ and adjustable parameters \mathbf{w} . If $\phi_0(\mathbf{x}) = 1$ and $\phi_i(\mathbf{x}) = x_i, 1 \leq i \leq d$ we obtain the usual linear hypothesis; if $\phi_i(\mathbf{x}) = \prod_{j=1}^d x_j^{k_j}, k_j \geq 0$ we obtain a polynomial hypothesis.

Let $\Phi_{h \times N} = [\phi(\mathbf{x}_1), \dots, \phi(\mathbf{x}_N)]$ be the training inputs transformed by the basis functions $\phi(\cdot)$. Let $\mathbf{y}_{N \times 1} = [y_1, \dots, y_N]^T$ be the training outputs. When $\Phi\Phi^T$ is full rank¹, the unique training error minimum is given by:

$$\mathbf{w}_0 = (\Phi\Phi^T)^{-1}\Phi\mathbf{y}$$

The Hessians of training and generalization errors are constant positive semi-definite² matrices at all \mathbf{w} :

$$H_{E_0}(\mathbf{w}) = 2\frac{\Phi\Phi^T}{N} \quad H_E(\mathbf{w}) = 2\langle \phi(\mathbf{x})\phi(\mathbf{x})^T \rangle_{\mathbf{x}}$$

Any higher derivatives of E and E_0 are 0 everywhere. Hence, for any $\Delta \mathbf{w}$, the generalization and training errors of $\mathbf{w}_0 \pm \Delta \mathbf{w}$ can be written as:

$$E(\mathbf{w}_0 \pm \Delta \mathbf{w}) = E(\mathbf{w}_0) \pm \Delta \mathbf{w}^T \partial E(\mathbf{w}_0) + \Delta \mathbf{w}^T \langle \phi(\mathbf{x})\phi(\mathbf{x})^T \rangle_{\mathbf{x}} \Delta \mathbf{w} \quad (1)$$

$$E_0(\mathbf{w}_0 \pm \Delta \mathbf{w}) = E_0(\mathbf{w}_0) + \Delta \mathbf{w}^T \frac{\Phi\Phi^T}{N} \Delta \mathbf{w} \quad (2)$$

¹We will assume that all matrix inverses needed exist for the rest of the paper.

²Any matrix of the form AA^T is positive semi-definite, because for any \mathbf{w} of proper dimensions, $\mathbf{w}^T AA^T \mathbf{w} = \|A^T \mathbf{w}\|^2 \geq 0$, hence $\Phi\Phi^T$ is positive semi-definite. $\langle \phi(\mathbf{x})\phi(\mathbf{x})^T \rangle_{\mathbf{x}}$ is also positive semi-definite since the expectation exists and $\Phi\Phi^T \xrightarrow{N \rightarrow \infty} \langle \phi(\mathbf{x})\phi(\mathbf{x})^T \rangle_{\mathbf{x}}$

Theorem 1: When all hypotheses with the same training error are equally likely to be chosen as the early stopping solution, the mean generalization error at any training error level above the training error minimum is greater than the generalization error of the training error minimum. More specifically, for any $\delta \geq 0$, $E_{mean}(E_\delta) = E(\mathbf{w}_0) + \beta(\delta)$, for some $\beta(\delta) \geq 0$.

Proof: Proofs for all theorems are given in the appendix.

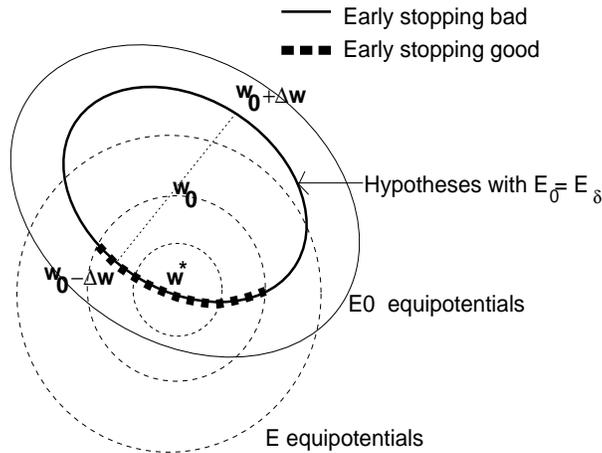


Figure 1: Early stopping at a training error δ above $E_0(\mathbf{w}_0)$ results in higher generalization error when all hypotheses having the same training error are equally likely to be chosen as the early stopping solution.

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

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

Theorem 2: When all hypotheses 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 $0 < \delta_1 < \delta_2$, $E_{mean}(E_{\delta_1}) < E_{mean}(E_{\delta_2})$.

3 General Hypotheses

When the hypotheses are continuous and differentiable functions, we can approximate the hypotheses around a training error minimum by means of a generalized-linear hypothesis function. There can be at most N (number of training examples) basis functions, otherwise $\Phi\Phi^T$ would be singular. For example, if Taylor approximation is used, each basis function would be a derivative of the hypothesis at the training error minimum. Depending on N :

- If N is finite, the approximation by the generalized-linear function would be accurate only around the training error minimum. Hence we can claim that theorems 1 and 2 hold but only for small enough δ depending on N and the derivatives of the general hypothesis function at the training error minimum.
- As $N \rightarrow \infty$, the approximation error goes to 0, and hence we can claim theorems 1 and 2 hold. In other words, if there are infinitely many examples and no other information, the best strategy is to descend on the training error till the minimum.

4 Classification Problems and the Bin Model

For classification problems, bin model [1] can be utilized to prove that mean generalization error increases as the training error increases. Since the proof does not have any assumptions about the probability distribution on the hypotheses with the same training error, it is worth mentioning here.

We will use the following version of the bin model: Let the M hypotheses in the learning model have generalization errors π_1, \dots, π_M . Determine the training errors of hypotheses ν_1, \dots, ν_M by picking N i.i.d. inputs and finding the errors on these samples for each bin. π_m corresponds to the generalization error E of a hypotheses, and ν_m corresponds to the training error E_0 . $P[\cdot]$ will denote the probability of the occurrence of an event.

The mean generalization error for training error level ν is:

$$E_{mean}(\nu) = \mathcal{E}[\pi|\nu] = \sum_{m=1}^M \pi_m P[\pi_m | \nu_m = \nu] \quad (3)$$

Theorem 3: For classification problems and hypotheses that can be formalized using the bin model, the mean generalization error is an increasing function of the training error.

5 Link to Early Stopping Using a Validation Set and Weight Decay and Future Research

Early stopping using a validation set operates as follows: the whole data is partitioned into a training and a validation set. The training error is minimized, and the validation error is monitored in the mean time. The hypothesis at which the validation error reaches its minimum is taken to be the early stopping solution. Early stopping using a validation set has two components that affect its performance:

- **Validation set size:** If the validation set is too small or large, early stopping may result in worse performance than minimizing the training error on all data. Although [2] suggests a validation set size, it is valid for very large training sets. A possible remedy for the validation set size problem is using all data for training, and early stopping not based on a validation set, but some other criterion. For example, if prior information or hints about the target function, such as, invariances [5], monotonicity [7] etc. are known, the hint error can be monitored while descending on the training error and training can be stopped according to the hint error. Another criteria is the level of training error, which we have examined in this paper.
- **Optimization algorithm:** Which hypotheses are visited by the optimization algorithm [8], or the initial hypothesis at which training starts [2] highly affect the performance of early stopping. Instead of the optimization algorithm determining which solution should be chosen for early stopping, any prior information about the data should determine how the optimization algorithm should behave. In this paper we have used the probability distribution on the hypotheses with the same training error both to incorporate any prior information about the target and to specify the optimization algorithm. If there is no prior information, all solutions with the same training error should be equally likely to be chosen as the solution, which is the case we have analyzed.

Success of weight decay depends on whether certain priors about the data are true or not. Weight decay results in a solution with smaller weights than the training error minimum. The training error minimum has large weights if certain criteria about the data and hypotheses are satisfied (in the case of generalized-linear hypotheses, if target outputs have additive noise and the

target function is also generalized-linear with the same basis functions as the hypotheses). Provided that the training error minimum has weights larger than the target and the weight decay parameter is chosen small enough, the weight decay solution has smaller generalization error than the training error minimum. Since weight decay solution's training error is larger than that of the training error minimum, weight decay can be seen as a method of early stopping. Different from the ones we have analyzed, in this case $P_{\mathbf{W}_\delta}$ is not uniform. It is actually a delta function nonzero only at the weight decay solution. This shows us that based on the prior information we have about the target function, we can have a nonuniform $P_{\mathbf{W}_\delta}$ and a decrease in the mean generalization error. Incorporation of other kinds of prior information, such as invariances, monotonicity etc. into early stopping is a promising research direction.

For generalized-linear hypotheses functions, when training starts from small weights and a small decent rate is used, the hypotheses visited during the descent usually lies close to weight decay solutions [4]. Provided that the validation set is not too large, early stopping using a validation set stops at a hypothesis close to a weight decay solution with a certain weight decay parameter. As long as the corresponding weight decay parameter is small enough, and the assumptions about target and noise are true, the solution is likely to have generalization error less than the training error minimum. This may be the reason why early stopping seems to be resulting in better generalization error in practice.

Appendix

Proof of Theorem 1:

Let the early stopping training error level be $E_\delta = E_0(\mathbf{w}_0) + \delta$ for some $\delta \geq 0$. Then, from equation (2), the early stopping set consists of $\mathbf{w}_0 + \mathbf{W}_\delta = \mathbf{w}_0 + \{\Delta \mathbf{w} : \Delta \mathbf{w}^T \frac{\Phi \Phi^T}{N} \Delta \mathbf{w} = \delta\}$. The mean generalization error is:

$$E_{mean}(E_\delta) = \int_{\Delta \mathbf{w} \in \mathbf{W}_\delta} P_{\mathbf{W}_\delta}(\Delta \mathbf{w}) E(\mathbf{w}_0 + \Delta \mathbf{w}) d\Delta \mathbf{w}$$

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

$$E_{mean}(E_\delta) =$$

$$0.5 \int_{\Delta \mathbf{w} \in \mathbf{W}_\delta} (P_{\mathbf{W}_\delta}(\Delta \mathbf{w})E(\mathbf{w}_0 + \Delta \mathbf{w}) + P_{\mathbf{W}_\delta}(-\Delta \mathbf{w})E(\mathbf{w}_0 - \Delta \mathbf{w})) d\Delta \mathbf{w}$$

Now, since $P_{\mathbf{W}_\delta}$ is uniform, it is also symmetric, i.e. $P_{\mathbf{W}_\delta}(\Delta \mathbf{w}) = P_{\mathbf{W}_\delta}(-\Delta \mathbf{w})$. For the proof of this theorem 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

$$\int_{\Delta \mathbf{w} \in \mathbf{W}_\delta} P_{\mathbf{W}_\delta}(\Delta \mathbf{w}) d\Delta \mathbf{w} = 1:$$

$$\begin{aligned} E_{mean}(E_\delta) &= E(\mathbf{w}_0) + \int_{\Delta \mathbf{w} \in \mathbf{W}_\delta} P_{\mathbf{W}_\delta}(\Delta \mathbf{w}) \Delta \mathbf{w}^T \left\langle \phi(\mathbf{x}) \phi(\mathbf{x})^T \right\rangle_{\mathbf{x}} \Delta \mathbf{w} d\Delta \mathbf{w} \\ &= E(\mathbf{w}_0) + \beta(\delta) \end{aligned}$$

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

$$\beta(\delta) = \int_{\Delta \mathbf{w} \in \mathbf{W}_\delta} P_{\mathbf{W}_\delta}(\Delta \mathbf{w}) \Delta \mathbf{w}^T \left\langle \phi(\mathbf{x}) \phi(\mathbf{x})^T \right\rangle_{\mathbf{x}} \Delta \mathbf{w} d\Delta \mathbf{w} \geq 0 \quad (4)$$

□

Proof of Theorem 2:

By theorem 1, $E_{mean}(E_{\delta_1}) = E(\mathbf{w}_0) + \beta(\delta_1)$ and $E_{mean}(E_{\delta_2}) = E(\mathbf{w}_0) + \beta(\delta_2)$ 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 $V(\delta) = \int_{\Delta \mathbf{w} \in \mathbf{W}_\delta} \Delta \mathbf{w}^T \left\langle \phi(\mathbf{x}) \phi(\mathbf{x})^T \right\rangle_{\mathbf{x}} \Delta \mathbf{w} d\Delta \mathbf{w}$, and let $\frac{1}{P_\delta}$ be the surface area of the h dimensional ellipsoid $\Delta \mathbf{w}^T \frac{\Phi \Phi^T}{N} \Delta \mathbf{w} = \delta$. Since $P_{\mathbf{W}_\delta}$ is uniform, from equation 4:

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

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

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

$$\frac{1}{P_\delta} = \lim_{l \rightarrow 0} \frac{\int_{\Delta \mathbf{w}^T \frac{\Phi \Phi^T}{N} \Delta \mathbf{w} \leq \delta+l} d\Delta \mathbf{w} - \int_{\Delta \mathbf{w}^T \frac{\Phi \Phi^T}{N} \Delta \mathbf{w} \leq \delta} d\Delta \mathbf{w}}{l}$$

$$\begin{aligned}
&= \lim_{l \rightarrow 0} \frac{\left(\frac{\delta+l}{\delta}\right)^{\frac{h+1}{2}} - 1}{l} \int_{\Delta \mathbf{w}^T \frac{\Phi \Phi^T}{N} \Delta \mathbf{w} \leq \delta} d\Delta \mathbf{w} \\
&= \frac{h+1}{2\delta} \int_{\Delta \mathbf{w}^T \frac{\Phi \Phi^T}{N} \Delta \mathbf{w} \leq \delta} d\Delta \mathbf{w}
\end{aligned}$$

Hence $\frac{1}{P_{\delta_1}} = \frac{h+1}{2\delta_1} \int_{\Delta \mathbf{w}^T \frac{\Phi \Phi^T}{N} \Delta \mathbf{w} \leq \delta_1} d\Delta \mathbf{w}$. By means of change of variables

$\Delta \mathbf{u} = \frac{\Delta \mathbf{w}}{k}$ we have $\frac{1}{P_{\delta_2}} = k^{h-1} \frac{1}{P_{\delta_1}}$. Therefore, $\frac{P_{\delta_2}}{P_{\delta_1}} = k^{-h+1}$.

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

Proof of Theorem 3:

Expanding the mean generalization error from equation (3):

$$\begin{aligned}
E_{mean}(\nu) &= \mathcal{E}[\pi|\nu] = \sum_{m=1}^M \pi_m P[\pi_m | \nu_m = \nu] \\
&= \frac{\sum_{m=1}^M \pi_m P[\nu_m = \nu | \pi_m] P[\pi_m]}{\sum_{m=1}^M P[\nu_m = \nu | \pi_m] P[\pi_m]} \\
&= \frac{\sum_{m=1}^M \pi_m P[\pi_m] \pi_m^{N\nu} (1 - \pi_m)^{N(1-\nu)}}{\sum_{m=1}^M P[\pi_m] \pi_m^{N\nu} (1 - \pi_m)^{N(1-\nu)}}
\end{aligned}$$

Taking the derivative of $\mathcal{E}[\pi|\nu]$ w.r.to ν :

$$\frac{d\mathcal{E}[\pi|\nu]}{d\nu} = Q_0 \sum_{\pi_m < \pi_k} Q_{m,k} (\pi_m - \pi_k) \ln \left(\frac{\pi_m}{1 - \pi_m} \frac{1 - \pi_k}{\pi_k} \right)$$

where $Q_0 = \frac{1}{N \left(\sum_{m=1}^M P[\pi_m] \pi_m^{N\nu} (1 - \pi_m)^{N(1-\nu)} \right)^2} > 0$ and

$Q_{m,k} = \pi_m^{N\nu} (1 - \pi_m)^{N(1-\nu)} \pi_k^{N\nu} (1 - \pi_k)^{N(1-\nu)} > 0$. When $\pi_m < \pi_k$ both $(\pi_m - \pi_k)$ and $\ln \left(\frac{\pi_m}{1 - \pi_m} \frac{1 - \pi_k}{\pi_k} \right)$ are negative hence the derivative is positive. Therefore the mean generalization error is an increasing function of the training error. \square

Acknowledgements

We would like to thank members of the Caltech Learning Systems Group: Dr. Amir Atiya, Alexander Nicholson, Joseph Sill and Xubo Song for many

useful discussions.

References

- [1] Y. S. Abu-Mostafa and X. Song (1996), "Bin Model for Neural Networks" in Proceedings of the International Conference on Neural Information Processing, Hong Kong, 1996, pp. 169–173.
- [2] S. Amari, N. Murata, K. Muller, M. Finke, H. H. Yang (1997), "Asymptotic Statistical Theory of Overtraining and Cross-Validation", *IEEE Trans. on Neural Networks*, vol. 8, no. 5, pp. 985–996.
- [3] P. Baldi and Y. Chauvin (1991), "Temporal Evolution of Generalization during Learning in Linear Networks", *Neural Computation*, **3**, 589–603.
- [4] C. Bishop (1995), *Neural Networks for Pattern Recognition*, Clarendon Press, Oxford, 1995.
- [5] Z. Cataltepe and Y. S. Abu-Mostafa (1994), "Estimating Learning Performance Using Hints", Proceedings of the 1993 Connectionist Models Summer School, M. Mozer et. al. (Eds.), Lawrence Erlbaum Associates, Publishers, Hillsdale, NJ. pp.380-386.
- [6] R. Dodier (1996), "Geometry of Early Stopping in Linear Networks" In G. Tesauro, D. S. Touretzky and T.K. Leen (eds.), *Advances in Neural Information Processing Systems 8*. Cambridge, MA: MIT Press.
- [7] J. Sill and Y. S. Abu-Mostafa (1997), "Monotonicity Hints", in M. Mozer, M. Jordan and T. Petsche (eds.), *Advances in Neural Information Processing Systems 9*. Cambridge, MA: MIT Press, pp.634-640.
- [8] J. Sjöberg and L. Ljung (1995), "Overtraining, regularization, and searching for a minimum, with application to neural networks", *Int. J. Control*, vol. **62**, no. 6, pp. 1391–1407.
- [9] C. Wang, S. S. Venkatesh, J. S. Judd (1994), "Optimal Stopping and Effective Machine Complexity in Learning", In G. Tesauro, D. S. Touretzky and T.K. Leen (eds.), *Advances in Neural Information Processing Systems 6*. Cambridge, MA: MIT Press.