A Bound on the Error of Cross Validation Using the Approximation and Estimation Rates, with Consequences for the Training-Test Split

Michael Kearns
AT&T Bell Laboratories
Murray Hill, NJ 07974
mkearns@research.att.com

Abstract: We give an analysis of the generalization error of cross validation in terms of two natural measures of the difficulty of the problem under consideration: the *approximation rate* (the accuracy to which the target function can be ideally approximated as a function of the number of hypothesis parameters), and the *estimation rate* (the deviation between the training and generalization errors as a function of the number of hypothesis parameters). The approximation rate captures the complexity of the target function with respect to the hypothesis model, and the estimation rate captures the extent to which the hypothesis model suffers from overfitting. Using these two measures, we give a rigorous and general bound on the error of cross validation. The bound clearly shows the tradeoffs involved with making γ — the fraction of data saved for testing — too large or too small. By optimizing the bound with respect to γ, we then argue (through a combination of formal analysis, plotting, and controlled experimentation) that the following qualitative properties of cross validation behavior should be quite robust to significant changes in the underlying model selection problem:

- When the target function complexity is small compared to the sample size, the performance of cross validation is relatively insensitive to the choice of γ.
- The importance of choosing γ optimally increases, and the optimal value for γ decreases, as the target function becomes more complex relative to the sample size.
- There is nevertheless a single *fixed* value for γ that works *nearly* optimally for a wide range of target function complexity.

Category: Learning Theory.
Prefer oral presentation.

1 INTRODUCTION

In this paper we analyze the performance of cross validation in the context of model selection and complexity regularization. We work in a setting in which we must choose the right number of parameters for a hypothesis function in response to a finite training sample, with the goal of minimizing the resulting generalization error. There is a large and interesting literature on cross validation methods, which often emphasizes asymptotic statistical properties, or the exact calculation of the generalization error for simple models. (The literature is too large to survey here; foundational papers include those of Stone [7, 8].) Our approach here is somewhat different, and is primarily inspired by two sources: the work of Barron and Cover [2], who introduced the idea of bounding the error of a model selection method (MDL in their case) in terms of a quantity known as the *index of resolvability*; and the work of Vapnik [9], who provides extremely powerful and general tools for uniformly bounding the deviations between training and generalization errors.

We combine these methods to give a new and general analysis of cross validation performance. In the first and more formal part of the paper, we give a rigorous bound on the error of cross validation in terms of two parameters of the underlying model selection problem (which is defined by a target function, an input distribution, and a nested sequence of increasingly complex classes of hypothesis functions): the *approximation rate* and the *estimation rate*, mentioned in the abstract and defined formally shortly. Taken together, these two problem parameters determine (our analogue of) the index of resolvability, and Vapnik's work yields estimation rates applicable to many natural problems. In the second part of the paper, we investigate the implications of our bound for choosing γ, the fraction of data withheld for testing in cross validation. The most interesting aspect of this analysis is the identification of several qualitative properties (identified in the abstract, and in greater detail in the paper) of the optimal γ that appear to be invariant over a wide class of model selection problems.

2 THE FORMALISM

In this paper we consider model selection as a two-part problem: choosing the appropriate number of parameters for the hypothesis function, and tuning these parameters. The training sample is used in both steps of this process. In many settings, the tuning of the parameters is determined by a fixed learning algorithm such as backpropagation, and then model selection reduces to the problem of choosing the architecture (which determines the number of weights, the connectivity pattern, and so on). For concreteness, here we adopt an idealized version of this division of labor. We assume a nested sequence of function classes $H_1 \subset \cdots \subset H_d \cdots$, called the *structure* [9], where H_d is a class of boolean functions of d parameters, each function being a mapping from some input space X into $\{0, 1\}$. *For simplicity, in this paper we assume that the Vapnik-Chervonenkis (VC) dimension [10, 9] of the class H_d is O(d).* To remove this assumption, one simply replaces all occurrences of d in our bounds by the VC dimension of H_d. We assume that we have in our possession a learning algorithm L that on input any training sample S and any value d will output a hypothesis function $h_d \in H_d$ that minimizes the training error over H_d — that is, $\epsilon_t(h_d) = \min_{h \in H_d}\{\epsilon_t(h)\}$, where $\epsilon_t(h)$ is the fraction of the examples in S on which h disagrees with the given label. In many situations, training error minimization is known to be computationally intractable, leading researchers to investigate heuristics such as backpropagation. The extent to which the theory presented here applies to such heuristics will depend in part on the extent to which they approximate training error minimization for the problem under consideration; however, many of our results have rigorous generalizations for the case in which no assumptions are made on the heuristic L (with the quality of the bounds of course decaying with the quality of the hypotheses output by L).

Model selection is thus the problem of choosing the best value of d. More precisely, we assume an arbitrary *target function* f (which may or may not reside in one of the function classes in the structure $H_1 \subset \cdots \subset H_d \cdots$), and an input distribution D; f and D together define the *generalization error* function $\epsilon_g(h) = \Pr_{x \in D}[h(x) \neq f(x)]$. We are given a training sample S of f, consisting of m random examples drawn according to D and labeled by f (with the labels possibly corrupted by noise). In many model selection methods (such as Rissanen's Minimum Description Length Principle [5] and Vapnik's Guaranteed Risk Minimization [9]), for each value of $d = 1, 2, 3, \ldots$ we give the *entire* sample S and d to the learning algorithm L to obtain the function h_d minimizing the training error in H_d. Some function of d and the training errors $\epsilon_t(h_d)$ is then used to choose among the sequence $h_1, h_2, h_3, \ldots$. Whatever the method, we interpret the goal to be that of minimizing the *generalization* error of the hypothesis selected.

In this paper, we will make the rather mild but very useful assumption that the structure has the property that for any sample size m, there is a value $d_{max}(m)$ such that $\epsilon_t(h_{d_{max}(m)}) = 0$ for *any* labeled sample S of m examples; we will refer to the function $d_{max}(m)$ as the *fitting number* of the structure [1]. The fitting number formalizes the simple notion that with enough parameters, we can always fit the training data perfectly, a property held by most sufficiently powerful function classes (including multilayer neural networks). We typically expect the fitting number to be a linear function of m, or at worst a polynomial in m. The significance of the fitting number for us is that no reasonable model selection method should choose h_d for $d \geq d_{max}(m)$, since doing so simply adds complexity without reducing the training error.

In this paper we concentrate on the simplest version of cross validation. Unlike the methods mentioned above, which use the entire sample for training the h_d, in cross validation we choose a parameter $\gamma \in [0, 1]$, which determines the split between training and test data. Given the input sample S of m examples, let S' be the subsample consisting of the first $(1-\gamma)m$ examples in S, and S'' the subsample consisting of the last γm examples. In cross validation, rather than giving the entire sample S to L, we give only the smaller sample S', resulting in the sequence $h_1, \ldots, h_{d_{max}((1-\gamma)m)}$ of increasingly complex hypotheses. Each hypothesis is now obtained by training on only $(1-\gamma)m$ examples, which implies that we will consider only d smaller than the corresponding fitting number $d_{max}((1-\gamma)m)$; let us introduce the shorthand d^{γ}_{max} for $d_{max}((1-\gamma)m)$. Cross validation chooses the h_d satisfying $h_d = \min_{i \in \{1,\ldots,d^{\gamma}_{max}\}}\{\epsilon''_t(h_i)\}$ where $\epsilon''_t(h_i)$ is the error of h_i on the subsample S''. Notice that we are *not* considering multifold cross validation, or other variants that make more efficient use of the sample, because our analyses will require the independence of the test set. However, we believe that many of the themes that emerge here apply to these more sophisticated variants as well.

We use $\epsilon_{cv}(m)$ to denote the generalization error $\epsilon_g(h_d)$ of the hypothesis h_d chosen by cross validation when given as input a sample S of m random examples of the target function; obviously, $\epsilon_{cv}(m)$ depends on the structure, f, D, and the noise rate. *When bounding $\epsilon_{cv}(m)$, we will use the expression "with high probability" to mean with probability*

[1]Weaker definitions for $d_{max}(m)$ also suffice, but this is the simplest.

$1-\delta$ *over the sample S, for some small fixed constant* $\delta > 0$. All of our results can also be stated with δ as a parameter at the cost of a $\log(1/\delta)$ factor in the bounds, or in terms of the expected value of $\epsilon_{cv}(m)$.

3 THE APPROXIMATION RATE

It is apparent that we should not expect to give nontrivial bounds on $\epsilon_{cv}(m)$ that take no account of some measure of the complexity of the unknown target function f; the correct measure of this complexity is less obvious. Following the example of Barron and Cover's analysis of MDL performance in the context of density estimation [2], we propose the *approximation rate* as a natural measure of the complexity of f and D in relationship to the chosen structure $H_1 \subset \cdots \subset H_d \cdots$. Thus we define the approximation rate function $\epsilon_g(d)$ to be $\epsilon_g(d) = \min_{h \in H_d}\{\epsilon_g(h)\}$. The function $\epsilon_g(d)$ tells us the best generalization error that can be achieved in the class H_d, and it is a nonincreasing function of d. If $\epsilon_g(d') = 0$ for some sufficiently large d', this means that the target function f, at least with respect to the input distribution, is realizable in the class $H_{d'}$, and thus d' is a coarse measure of how complex f is. More generally, even if $\epsilon_g(d) > 0$ for all d, the rate of decay of $\epsilon_g(d)$ still gives a nice indication of how much representational power we gain with respect to f and D by increasing the complexity of our models. Still missing, of course, is some means of determining the extent to which this representational power can be realized by training on a finite sample of a given size, but this will be added shortly. First we give examples of the approximation rate that we will examine in some detail following the general bound on $\epsilon_{cv}(m)$.

The Intervals Problem. In this problem, the input space X is the real interval [0, 1], and the class H_d of the structure consists of all boolean step functions over $[0, 1]$ of at most d steps; thus, each function partitions the interval [0, 1] into at most d disjoint segments (not necessarily of equal width), and assigns alternating positive and negative labels to these segments. Thus, the input space is one-dimensional, but the structure contains arbitrarily complex functions over $[0, 1]$. It is easily verified that our assumption that the VC dimension of H_d is $O(d)$ holds here, and that the fitting number obeys $d_{max}(m) \leq m$. Now suppose that the input density D is uniform, and suppose that the target function f is the function of s alternating segments of equal width $1/s$, for some s (thus, f lies in the class H_s). We will refer to these settings as the *intervals problem.* Then the approximation rate $\epsilon_g(d)$ is $\epsilon_g(d) = (1/2)(1 - d/s)$ for $1 \leq d < s$ and $\epsilon_g(d) = 0$ for $d \geq s$. Thus, as long as $d < s$, increasing the complexity d gives linear payoff in terms of decreasing the optimal generalization error. For $d \geq s$, there is no payoff for increasing d, since we can already realize the target function. The reader can easily verify that if f lies in H_s, but does not have equal width intervals, $\epsilon_g(d)$ is still piecewise linear, but for $d < s$ is "concave up": the gain in approximation obtained by incrementally increasing d diminishes as d becomes larger. Although the intervals problem is rather simple and artificial, a precise analysis of cross validation behavior can be given for it, and we will argue that this behavior is representative of much broader and more realistic settings.

The Perceptron Problem. In this problem, the input space X is $\Re^N$ for some large natural number N. The class H_d consists of all perceptrons over the N inputs in which at most d weights are nonzero. If the input density is spherically symmetric (for instance, the uniform density on the unit ball in $\Re^N$), and the target function is a function in H_s with all s nonzero weights equal to 1, then it can be shown that the approximation rate function $\epsilon_g(d)$ is $\epsilon_g(d) = (1/\pi)\cos^{-1}(\sqrt{d/N})$ for $d < s$ [6], and of course $\epsilon_g(d) = 0$ for $d \geq s$. This problem provides a nice contrast to the intervals problem, since here the behavior of the approximation rate for small d is concave down: as long as $d < s$, an incremental increase in d yields more approximative power for large d than it does for small d (except for very small values of d).

Power Law Decay. In addition to the specific examples just given, we would also like to study reasonably natural parametric forms of $\epsilon_g(d)$, to determine the sensitivity of our theory to a plausible range of behaviors for the approximation rate. This is important, since in practice we do not expect to have precise knowledge of $\epsilon_g(d)$, since it depends on the target function and input distribution. Following the work of Barron [1], who shows a c/d bound on $\epsilon_g(d)$ for the case of neural networks with one hidden layer under a squared error generalization measure (where c is a measure of target function complexity in terms of a Fourier transform integrability condition) [2], in the later part of the paper we investigate $\epsilon_g(d)$ of the form $(c/d)^\alpha + \epsilon_{min}$, where $\epsilon_{min} \geq 0$ is a parameter representing the "degree of unrealizability" of f with respect to the structure, and $c, \alpha > 0$ are parameters capturing the rate of decay to ϵ_{min}. Our analysis concludes that the qualitative phenomena we identify are invariant to wide ranges of choices for these parameters.

Note that for all three cases, there is a natural measure of target function complexity captured by $\epsilon_g(d)$: in the intervals problem it is the number of target intervals, in the perceptron problem it is the number of nonzero weights in the target,

[2]Since the bounds we will give have straightforward generalizations to real-valued function learning under squared error (details in the full paper), examining behavior for $\epsilon_g(d)$ in this setting seems reasonable.

and in the more general power law case it is captured by the parameters of the power law. In the later part of the paper, we will study cross validation performance as a function of these complexity measures, and obtain remarkably similar predictions.

4 THE ESTIMATION RATE

For a fixed f, D and $H_1 \subset \cdots \subset H_d \cdots$, we say that a function $\rho(d,m)$ is an *estimation rate bound* if for all d and m, with high probability over the sample S we have $|\epsilon_t(h_d) - \epsilon_g(h_d)| \leq \rho(d,m)$, where as usual h_d is the result of training error minimization on S within H_d. Thus $\rho(d,m)$ simply bounds the deviation between the training error and the generalization error of h_d [3]. Note that the best such bound may depend in a complicated way on all of the elements of the problem: f, D and the structure. Indeed, much of the recent work on the statistical physics theory of learning curves has documented the wide variety of behaviors that such deviations may assume [6, 3]. However, for many natural problems it is both convenient and accurate to rely on a *universal* estimation rate bound provided by the powerful theory of uniform convergence: Namely, for any f, D and any structure the function $\rho(d,m) = \sqrt{(d/m)\log(m/d)}$ is an estimation rate bound [9] [4]. Depending upon the details of the problem, it is sometimes appropriate to omit the $\log(m/d)$ factor, and often appropriate to refine the $\sqrt{d/m}$ behavior to a function that interpolates smoothly between d/m behavior for small ϵ_t to $\sqrt{d/m}$ for large ϵ_t. Although such refinements are both interesting and important, many of the qualitative claims and predictions we will make are invariant to them as long as the deviation $|\epsilon_t(h_d) - \epsilon_g(h_d)|$ is well-approximated by a power law $(d/m)^\alpha$ ($\alpha > 0$); it will be more important to recognize and model the cases in which power law behavior is grossly violated.

Note that this universal estimation rate bound holds only under the assumption that the training sample is noise-free, but straightforward generalizations exist. For instance, if the training data is corrupted by random label noise at rate $0 \leq \eta < 1/2$, then $\rho(d,m) = \sqrt{(d/(1-2\eta)^2 m) log(m/d)}$ is again a universal estimation rate bound.

After giving a general bound on $\epsilon_{cv}(m)$ in which the approximation and estimation rate functions are parameters, we investigate the behavior of $\epsilon_{cv}(m)$ (and more specifically, of the parameter γ) for specific choices of these parameters.

5 THE BOUND

Theorem 1 *Let $H_1 \subset \cdots \subset H_d \cdots$ be any structure, where the VC dimension of H_d is $O(d)$. Let f and D be any target function and input distribution, let $\epsilon_g(d)$ be the approximation rate function for the structure with respect to f and D, and let $\rho(d,m)$ be an estimation rate bound for the structure with respect to f and D. Then for any m, with high probability*

$$\epsilon_{cv}(m) \leq \min_{1 \leq d \leq d^\gamma_{max}} \{\epsilon_g(d) + \rho(d,(1-\gamma)m)\} + O\left(\sqrt{\frac{\log(d^\gamma_{max})}{\gamma m}}\right) \tag{1}$$

where γ is the fraction of the training sample used for testing, and $d^\gamma_{max} = d_{max}((1-\gamma)m)$. Using the universal estimation bound rate and the rather weak assumption that $d_{max}(m)$ is polynomial in m, we obtain that with high probability

$$\epsilon_{cv}(m) \leq \min_{1 \leq d \leq d^\gamma_{max}} \left\{\epsilon_g(d) + O\left(\sqrt{\frac{d}{(1-\gamma)m}\log\left(\frac{m}{d}\right)}\right)\right\} + O\left(\sqrt{\frac{\log((1-\gamma)m)}{\gamma m}}\right). \tag{2}$$

Straightforward generalizations of these bounds for the case where the data is corrupted by random label noise can be obtained, using the modified estimation rate bound mentioned in Section 4 (details in the full paper).

Proof Sketch: We have space only to highlight the main ideas of the proof. For each d from 1 to d^γ_{max}, fix a function $f_d \in H_d$ satisfying $\epsilon_g(f_d) = \epsilon_g(d)$; thus, f_d is the best possible approximation to the target f within the class H_d. By a standard Chernoff bound argument it can be shown that with high probability we have $|\epsilon_t(f_d) - \epsilon_g(f_d)| \leq \sqrt{\log(d^\gamma_{max})/m}$ for all $1 \leq d \leq d^\gamma_{max}$. This means that within each H_d, the minimum training error $\epsilon_t(h_d)$ is at

[3] In the later and less formal part of the paper, we will often assume that specific forms of $\rho(d,m)$ are not merely upper bounds on this deviation, but accurate approximations to it.

[4] The results of Vapnik actually show the stronger result that $|\epsilon_t(h) - \epsilon_g(h)| \leq \sqrt{(d/m)\log(m/d)}$ for *all* $h \in H_d$, not only for the training error minimizer h_d.

most $\epsilon_g(d) + \sqrt{\log(d^{\gamma}_{max})/m}$. Since $\rho(d,m)$ is an estimation rate bound, we have that with high probability, for all $1 \leq d \leq d^{\gamma}_{max}$,

$$\begin{aligned} \epsilon_g(h_d) &\leq \epsilon_t(h_d) + \rho(d,m) && (3) \\ &\leq \epsilon_g(d) + \sqrt{\log(d^{\gamma}_{max})/m} + \rho(d,m). && (4) \end{aligned}$$

Thus we have bounded the generalization error of the d^{γ}_{max} hypotheses h_d, only one of which will be selected. If we knew the actual values of these generalization errors (equivalent to having an infinite test sample in cross validation), we could bound our error by the minimum over all $1 \leq d \leq d^{\gamma}_{max}$ of the expression (4) above. However, in cross validation we do not know the exact values of these generalization errors but must instead use the γm testing examples to estimate them. Again by standard Chernoff bound arguments, this introduces an additional $\sqrt{\log(d^{\gamma}_{max})/(\gamma m)}$ error term, resulting in our final bound. This concludes the proof sketch.

In the bounds given by (1) and (2), the $\min\{\cdot\}$ expression is analogous to Barron and Cover's index of resolvability [2]; the final term in the bounds represents the error introduced by the testing phase of cross validation. These bounds exhibit tradeoff behavior with respect to the parameter γ: as we let γ approach 0, we are devoting more and more of the sample to training the h_d, and the estimation rate bound term $\rho(d,(1-\gamma)m)$ is decreasing. However, the test error term $O(\sqrt{\log(d^{\gamma}_{max})/(\gamma m)})$ is increasing, since we have less data to accurately estimate the $\epsilon_g(h_d)$. The reverse phenomenon occurs as we let γ approach 1.

While we believe Theorem 1 to be enlightening and potentially useful in its own right, we would now like to take its interpretation a step further. More precisely, suppose we assume that the bound is an approximation to the actual behavior of $\epsilon_{cv}(m)$. Then in principle we can optimize the bound to obtain the best value for γ. Of course, in addition to the assumptions involved (the main one being that $\rho(d,m)$ is a good approximation to the training-generalization error deviations of the h_d), this analysis can only be carried out given information that we should not expect to have in practice (at least in exact form) — in particular, the approximation rate function $\epsilon_g(d)$, which depends on f and D. *However, we argue in the coming sections that several interesting qualitative phenomena regarding the choice of* γ *are largely invariant to a wide range of natural behaviors for* $\epsilon_g(d)$.

6 A CASE STUDY: THE INTERVALS PROBLEM

We begin by performing the suggested optimization of γ for the intervals problem. Recall that the approximation rate here is $\epsilon_g(d) = (1/2)(1 - d/s)$ for $d < s$ and $\epsilon_g(d) = 0$ for $d \geq s$, where s is the complexity of the target function. Here we analyze the behavior obtained by assuming that the estimation rate $\rho(d,m)$ actually behaves as $\rho(d,m) = \sqrt{d/(1-\gamma)m}$ (so we are omitting the log factor from the universal bound) [5], and to simplify the formal analysis a bit (but without changing the qualitative behavior) we replace the term $\sqrt{\log((1-\gamma)m)/(\gamma m)}$ by the weaker $\sqrt{\log(m)/m}$. Thus, if we define the function

$$F(d,m,\gamma) = \epsilon_g(d) + \sqrt{d/(1-\gamma)m} + \sqrt{\log(m)/(\gamma m)} \qquad (5)$$

then following (1), we are approximating $\epsilon_{cv}(m)$ by $\epsilon_{cv}(m) \approx \min_{1 \leq d \leq d^{\gamma}_{max}}\{F(d,m,\gamma)\}$ [6] (see Figure 1 for a plot of this approximation).

The first step of the analysis is to fix a value for γ and differentiate $F(d,m,\gamma)$ with respect to d to discover the minimizing value of d. This differentiation must have two regimes due to the discontinuity at $d = s$ in $\epsilon_g(d)$. It is easily verified that the derivative is $-(1/2s) + 1/(2\sqrt{d(1-\gamma)m})$ for $d < s$ and $1/(2\sqrt{d(1-\gamma)m})$ for $d \geq s$. It can be shown that provided that $(1-\gamma)m \geq 4s$ then $d = s$ is a global minimum of of $F(d,m,\gamma)$, and if this condition is violated then the value we obtain for $\epsilon_{cv}(m)$ is vacuously large anyway (meaning that this fixed choice of γ can not end up being the optimal one, or, if $m < 4s$, that our analysis claims we simply do not have enough data for nontrivial generalization, regardless of how we split it between training and testing). Plugging in $d = s$ yields $\epsilon_{cv}(m) \approx F(s,m,\gamma) = \sqrt{s/(1-\gamma)m} + \sqrt{\log(m)/(\gamma m)}$ for this fixed choice of γ. Now by differentiating $F(s,m,\gamma)$ with respect to γ, it can be shown that the optimal choice of γ under the assumptions is $\gamma_{opt} = (\log(m)/s)^{1/3}/(1 + (\log(m)/s)^{1/3})$.

[5] It can be argued that this power law estimation rate is actually a rather accurate approximation for the true behavior of the training-generalization deviations of the h_d for this problem.

[6] Although there are hidden constants in the $O(\cdot)$ notation of the bounds, it is the *relative* weights of the estimation and test error terms that is important, and choosing both constants equal to 1 is a reasonable choice (since both terms have the same Chernoff bound origins).

It is important to remember at this point that despite the fact that we have derived a precise expression for γ_{opt}, due to the assumptions and approximations we have made in the various constants, any *quantitative* interpretation of this expression is meaningless. However, we *can* reasonably expect that this expression captures the qualitative way in which the optimal γ changes as the amount of data m changes in relation to the target function complexity s. On this score the situation initially appears rather bleak, as the function $(\log(m)/s)^{1/3}/(1+(\log(m)/s)^{1/3})$ is quite sensitive to the ratio $\log(m)/s$. For example, for $m = 10000$, if $s = 10$ we obtain $\gamma_{opt} = 0.524\cdots$, if $s = 100$ we obtain $\gamma_{opt} = 0.338\cdots$, and if $s = 1000$ we obtain $\gamma_{opt} = 0.191\cdots$. Thus γ_{opt} is becoming smaller as $\log(m)/s$ becomes small, and the analysis suggests vastly different choices for γ_{opt} depending on the target function complexity, which is something we do not expect to have the luxury of knowing in practice.

However, it is both fortunate and interesting that γ_{opt} does not tell the entire story. In Figure 2, we plot the function $F(s, m, \gamma)$ [7] as a function of γ for $m = 10000$ and for several different values of s (note that for consistency with the later experimental plots, the x axis of the plot is actually the training fraction $1 - \gamma$). Here we can observe four important qualitative phenomena:

(A) When s is small compared to m, the predicted error is relatively insensitive to the choice of γ: as a function of γ, $F(s, m, \gamma)$ has a wide, flat bowl, indicating a wide range of γ yielding essentially the same near-optimal error.

(B) As s becomes larger in comparison to the fixed sample size m, the relative superiority of γ_{opt} over other values for γ becomes more pronounced. In particular, large values for γ become progressively worse as s increases. For example, the plots indicate that for $s = 10$ (again, $m = 10000$), even though $\gamma_{opt} = 0.524\cdots$ the choice $\gamma = 0.75$ will result in error quite near that achieved using γ_{opt}. However, for $s = 500$, $\gamma = 0.75$ is predicted to yield greatly suboptimal error.

(C) Because of the insensitivity to γ for s small compared to m, there is a *fixed* value of γ which seems to yield reasonably good performance for a wide range of values for s; this value is essentially the value of γ_{opt} for the case where s is large (but nontrivial generalization is still possible), since choosing the best value for γ is more important there than for the small s case. Note that for very large s, the bound predicts vacuously large error for *all* values of γ, so that the choice of γ again becomes irrelevant.

(D) The value of γ_{opt} is *decreasing* as s increases. This is slightly difficult to confirm from the plot, but can be seen clearly from the precise expression for γ_{opt}.

Despite the fact that our analysis so far has been rather specialized (addressing the behavior for a fixed structure, target function and input distribution), it is our belief that (A), (B), (C) and (D) above are rather universal phenomena that hold for many other model selection problems. For instance, we shall shortly demonstrate that our theory again predicts that (A), (B), (C) and (D) hold for the perceptron problem, and for case of power law decay of $\epsilon_g(d)$ described earlier. First we give an experimental demonstration that at least the predicted properties (A), (B) and (C) truly do hold for the intervals problem.

In Figures 3, 4 and 5, we plot the results of experiments in which labeled random samples of size $m = 5000$ were generated for a target function of s equal width intervals, for $s = 10, 100$ and 500. The samples were corrupted by random label noise at rate $\eta = 0.3$. For each value of γ and each value of d, $(1-\gamma)m$ of the sample was given to a program implementing training error minimization for the class H_d [8]; the remaining γm examples were used to select the best h_d according to cross validation. The plots show the true generalization error of the h_d selected by cross validation as a function of γ; this generalization error can be computed exactly for this problem. Each point in the plots represents an average over 10 trials.

While there are obvious and significant quantitative differences between these experimental plots and Figure 2, the properties (A), (B) and (C) are rather clearly borne out by the data: (A) In Figure 3, where s is small compared to m, there is a wide range of acceptable γ; it appears that any choice of γ between 0.10 and 0.50 yields nearly optimal generalization error. (B) By the time $s = 100$ (Figure 4), the sensitivity to γ is considerably more pronounced. For example, the choice $\gamma = 0.50$ now results in clearly suboptimal performance, and it is more important to have γ close to 0.10. (C) Despite these complexities, there does indeed appear to be single value of γ — approximately 0.10 — that performs nearly optimally for the entire range of s examined.

The property (D) — namely, that the optimal γ decreases as the target function complexity is increased relative to a fixed m — is certainly not refuted by the experimental results, but any such effect is simply too small to be verified. It

[7] In the plots, we now use the more accurate test penalty term $\sqrt{\log((1-\gamma)m)/(\gamma m)}$ since we are no longer concerned with simplifying the calculation.

[8] A nice feature of the intervals problem is the fact that training error minimization can be performed in almost linear time using a dynamic programming approach [4].

would be interesting to verify this prediction experimentally, perhaps on a different problem where the predicted effect is more pronounced.

7 POWER LAW DECAY AND THE PERCEPTRON PROBLEM

For the cases where the approximation rate $\epsilon_g(d)$ obeys either power law decay or is that derived for the perceptron problem discussed in Section 3, the behavior of $\epsilon_{cv}(m)$ as a function of γ predicted by our theory is largely the same. For example, if $\epsilon_g(d) = (c/d)$ and we use the standard estimation rate $\rho(d,m) = \sqrt{d/(1-\gamma)m}$, then an analysis similar to that performed for the intervals problem reveals that for fixed γ the minimizing choice of d is $d = (4c^2(1-\gamma)m)^{1/3}$; plugging this value of d back into the bound on $\epsilon_{cv}(m)$ yields Figure 6, which, like Figure 2 for the intervals problem, shows the predicted behavior of $\epsilon_{cv}(m)$ as a function of γ for a fixed m and several different choices for the complexity parameter c. We again see that properties (A), (B), (C) and (D) hold strongly despite the change in $\epsilon_g(d)$ from the intervals problem (although quantitative aspects of the prediction, which already must be taken lightly for reasons previously stated, have obviously changed, such as the "interesting" values of the ratio of sample size to target function complexity).

Through a combination of formal analysis and plotting, it is possible to demonstrate that the properties (A), (B), (C) and (D) are robust to wide variations in the parameters α and ϵ_{min} in the parametric form $\epsilon_g(d) = (c/d)^\alpha + \epsilon_{min}$, as well as wide variations in the form of the estimation rate $\rho(d,m)$. For example, if $\epsilon_g(d) = (c/d)^2$ (faster than the approximation rate examined above) and $\rho = d/m$ (faster than the estimation rated examined above), then for the interesting ratios of m to c (that is, where the generalization error predicted is bounded away from 0 and the trivial value of $1/2$), a figure quite similar to Figure 6 is obtained. Similar predictions can be derived for the perceptron problem using the universal estimation rate bound or any similar power law form.

In summary, our theory predicts that although significant quantitative differences in the behavior of cross validation may arise for different model selection problems, the properties (A), (B), (C) and (D) should be present in a wide range of problems. At the very least, the behavior of our *bounds* exhibits these properties for a wide range of problems. It would be interesting to try to identify natural problems for which one or more of these properties is strongly violated; a potential source for such problems may be those for which the underlying learning curve deviates from classical power law behavior [6, 3].

8 ACKNOWLEDGEMENTS

I give warm thanks to Yishay Mansour, Dana Ron and Yoav Freund for their contributions to this work and for many interesting discussions. The code used in the experiments was written by Andrew Ng.

References

[1] A. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, 19:930–944, 1991.

[2] A. R. Barron and T. M. Cover. Minimum complexity density estimation. *IEEE Transactions on Information Theory*, 37:1034–1054, 1991.

[3] D. Haussler, M. Kearns, H.S. Seung, and N. Tishby. Rigourous learning curve bounds from statistical mechanics. In *Proceedings of the Seventh Annual ACM Confernce on Computational Learning Theory*, pages 76–87, 1994.

[4] M. Kearns, Y. Mansour, A. Ng, and D. Ron. An experimental and theoretical comparison of model selection methods. In *Proceedings of the Eigth Annual ACM Conference on Computational Learning Theory*, 1995.

[5] J. Rissanen. *Stochastic Complexity in Statistical Inquiry*, volume 15 of *Series in Computer Science*. World Scientific, 1989.

[6] H. S. Seung, H. Sompolinsky, and N. Tishby. Statistical mechanics of learning from examples. *Physical Review*, A45:6056–6091, 1992.

[7] M. Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society B*, 36:111–147, 1974.

[8] M. Stone. Asymptotics for and against cross-validation. *Biometrika*, 64(1):29–35, 1977.

[9] V. N. Vapnik. *Estimation of Dependences Based on Empirical Data*. Springer-Verlag, New York, 1982.

[10] V. N. Vapnik and A. Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264–280, 1971.

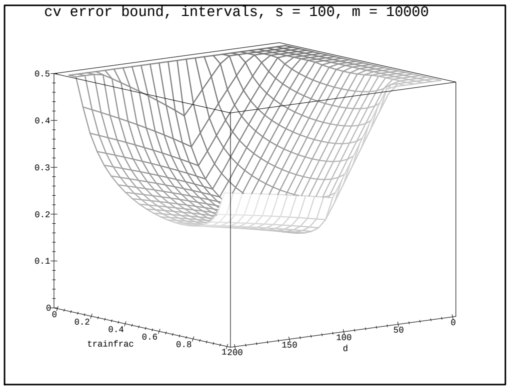

Figure 1: Plot of the bound/approximation $\epsilon_{cv}(m) \approx \epsilon_g(d) + \sqrt{d/(1-\gamma)m} + \sqrt{\log((1-\gamma)m)/(\gamma m)}$ as a function of the training fraction $1-\gamma$ and d for $m = 10000$, and $\epsilon_g(d)$ for the intervals problem with a target function of complexity $s = 100$. Several interesting features are evident for these values, including the predicted choice of $d = s = 100$, and the curvature as a function of γ, indicating a unique optimal choice for γ bounded away from 0 and 1. (Note that we have plotted the minimum of the bound and 0.5, since this generalization error can be trivially achieved by flipping a fair coin.)

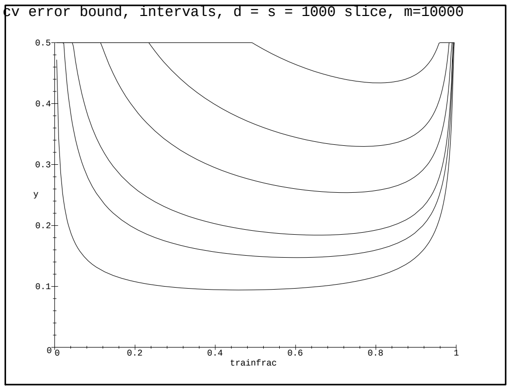

Figure 2: Plot of the predicted generalization error of cross validation for the intervals model selection problem, as a function of the fraction $1-\gamma$ of data used for training. (In the plot, the fraction of training data is 0 on the left ($\gamma = 1$) and 1 on the right ($\gamma = 0$)). The fixed sample size $m = 10,000$ was used, and the 6 plots show the error predicted by the theory for target function complexity values $s = 10$ (bottom plot), 50, 100, 250, 500, and 1000 (top plot). The movement and relative superiority of the optimal choice for γ identified by properties (A), (B), (C) and (D) can be seen clearly. (We have again plotted the minimum of the the predicted error and 0.5.)

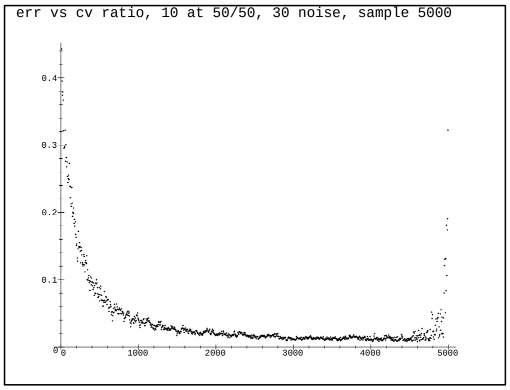

Figure 3: Experimental plot of cross validation generalization error as a function of training set size $(1-\gamma)m$, for $s = 10$ and $m = 5000$, with 30% label noise added. As in Figure 2, the x axis indicates the amount of data used for training. For this small value of the ratio of s to m, property (A) predicted by the theory is confirmed: there is a wide range of γ values yielding essentially the same generalization error.

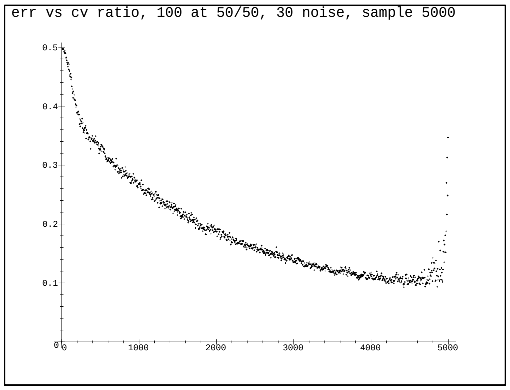

Figure 4: Experimental plot of cross validation generalization error as a function of training set size $(1-\gamma)m$ as in Figure 3, but now for the target function complexity $s = 100$. Property (B) predicted by the theory can be seen: compared to Figure 3, the relative superiority of the optimal γ over other values is increasing, and in particular larger values of γ (such as 0.5) are now clearly suboptimal choices.

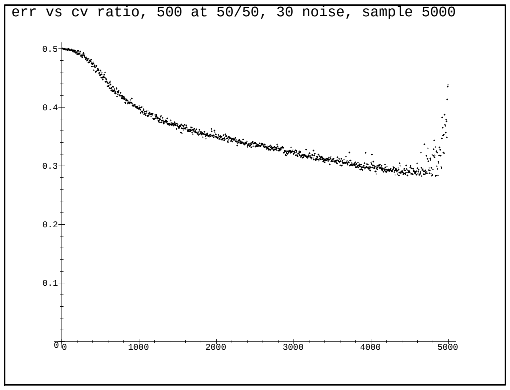

Figure 5: Experimental plot of cross validation generalization error as a function of training set size $(1-\gamma)m$, now for $s = 500$. Figures 3, 4 and 5 together confirm the predicted property (C): as in the theoretical plot of Figure 2, despite the differing shapes of the three plots there is a fixed value (about $\gamma = 0.1$) yielding near-optimal performance for the entire range of s.

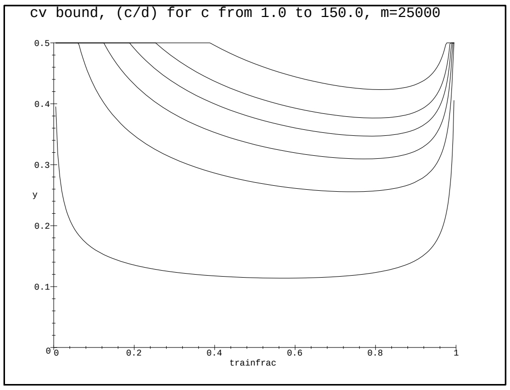

Figure 6: Plot of the predicted generalization error of cross validation for the case $\epsilon_g(d) = (c/d)$, as a function of the fraction $1-\gamma$ of data used for training. The fixed sample size $m = 25,000$ was used, and the 6 plots show the error predicted by the theory for target function complexity values $c = 1$ (bottom plot), 25, 50, 75, 100, and 150 (top plot). Properties (A), (B), (C) and (D) are again evident. (We have again plotted the minimum of the the predicted error and 0.5.)