Appears in
Proceedings of the Fourteenth International Conference on Machine Learning (ICML-97),
Nashville, July 1997.

Characterizing the generalization performance of model selection strategies

Dale Schuurmans[*]
Inst. for Research in Cognitive Science
University of Pennsylvania
Philadelphia, PA 19104-6228
daes@linc.cis.upenn.edu

Lyle H. Ungar
Computer & Infor. Science
University of Pennsylvania
Philadelphia, PA 19104-6389
ungar@central.cis.upenn.edu

Dean P. Foster
Department of Statistics
University of Pennsylvania
Philadelphia, PA 19104-6302
foster@hellspark.wharton.upenn.edu

Abstract

We investigate the structure of model selection problems via the bias/variance decomposition. In particular, we characterize the essential aspects of a model selection task by the bias and variance profiles it generates over the sequence of hypothesis classes. With this view, we develop a new understanding of complexity-penalization methods: First, the penalty terms can be interpreted as postulating a particular profile for the variances as a function of model complexity—if the postulated and true profiles do not match, then systematic under-fitting or over-fitting results, depending on whether the penalty terms are too large or too small. Second, we observe that it is generally best to penalize according to the true variances of the task, and therefore no fixed penalization strategy is optimal across all problems. We then use this characterization to introduce the notion of easy versus hard model selection problems. Here we show that if the variance profile grows too rapidly in relation to the biases, then standard model selection techniques become prone to significant errors. This can happen, for example, in regression problems where the independent variables are drawn from wide-tailed distributions. To counter this, we discuss a new model selection strategy that dramatically outperforms standard complexity-penalization and hold-out methods on these hard tasks.

1 Introduction

When learning a function $h : X \to Y$ from random training examples $\langle x_1, y_1\rangle$, ..., $\langle x_t, y_t\rangle$, there is a well-known tradeoff between the size of the training sample and the complexity of the function class being considered: If the class is too complex for the sample size, there is a risk of ”overfitting” the training data and guessing a function that performs poorly on future test examples. On the other hand, an overly restricted class can prevent us from considering any good candidate functions. The most common strategy for coping with this dilemma in practice is to use some form of automatic *model selection*, such as complexity-penalization or repeated hold-out testing, to balance the tradeoff between complexity and data-fit.

Under the simplest formulation of model selection, the idea is to first stratify the hypothesis class H into a sequence of nested subclasses $H_0 \subset H_1 \subset ... = H$ and then (somehow) choose a class which has the appropriate complexity for the given training data. To understand how we might make this choice, note that for a given training sample $S = \langle x_1, y_1\rangle, ..., \langle x_t, y_t\rangle$ we obtain a corresponding sequence of empirically optimal functions, $h_0^* \in H_0, h_1^* \in H_1, ...,$ *etc*., that achieve minimum average error $\widehat{err}(h_i^*) \overset{\triangle}{=} \sum_{j=1}^{t} loss(h_i^*(x_j), y_j)/t$ on the training set S. The essence of the model selection problem is to choose one of these functions based on their observed empirical errors $\widehat{err}(h_1^*), \widehat{err}(h_2^*), ...$ Note, however, that these errors are monotonically decreasing, and therefore choosing the function with minimum training error simply leads to choosing a function from the largest class. Therefore, the trick is to invoke some other criteria beyond empirical error minimization to make this choice.

Currently, two basic model selection strategies predominate. The most common strategy is *complexity-penalization*. Here one assigns increasing complexity values $c_0, c_1, ...$ to the successive function classes, and then chooses the hypothesis from $h_1^*, h_2^*, ...$ that minimizes some combination of complexity and empirical error (*e.g.*, the additive combination $c_i + \lambda\widehat{err}(h_i^*)$). There are many variants of this basic approach, including the minimum description length principle (Rissanen 1986), “Bayesian” maximum a posteriori selection, structural risk minimization (Vapnik 1982; 1996), “generalized” cross validation (Craven & Wahba 1979), and even regularization (Moody 1992). These strategies differ in the specific complexity values they assign and the particular tradeoff function they

[*] **Also:** NEC Research Institute, Princeton, NJ

optimize, but the basic idea is still the same.

The other most common strategy is *hold-out testing*. Here one asks: for the given set of training data, which hypothesis class H_i generalizes best? We answer this by partitioning the training set, $1, ..., t$, into a pseudo-training set, $1, ..., k$, and a hold-out test set, $k+1, ..., t$, and then using the pseudo-training set to obtain a sequence of pseudo-hypotheses $\hat{h}_0, \hat{h}_1, ...,$ *etc*. We then use the hold-out test set to obtain an *unbiased* estimate of the true errors of these pseudo-hypotheses. (Note that the training set errors tend to be gross underestimates in general.) From these unbiased estimates, we can simply choose the hypothesis class H_i that yields the pseudo-hypothesis $\hat{h}_i$ with the smallest estimated error. Once H_i has been selected, we return the function $h_i^* \in H_i$ that obtains minimum empirical error on the *entire* training sequence. Again, there are many variants to this basic strategy—having to do with repeating the pseudo-train pseudo-test split many times and averaging the results to choose the final hypothesis class; *e.g.*, 10-fold cross validation, leave-one-out testing, bootstrapping, *etc*. (Efron 1979; Weiss & Kulikowski 1991).

The abundance of model selection strategies and different approaches to the problem raises the question of which techniques are best and when. We attempt to answer this in the context of regression by appealing to the standard bias/variance decomposition of generalization error (Geman, Bienenstock, & Doursat 1992). In particular, we characterize model selection problems by the bias and variance profiles they generate over the sequence of hypothesis classes. Given this characterization, we address a number of topics regarding the behavior of model selection strategies and the structure of model selection tasks: First, we investigate complexity-penalization methods, which attempt to directly adjust the empirical error estimates to account for the unseen variances. Here we observe that no single penalization strategy dominates in every situation—all penalization methods have conditions where they perform well and conditions where they fail. Next, we investigate the structure of model selection *problems*, and identify the notion of easy versus hard model selection tasks. Here we show that some problems are inherently more difficult than others for standard complexity-penalization and hold-out methods. For example, regression problems where the independent variables come from wide-tailed distributions cause difficulties for standard selection strategies, because the example vectors encountered in testing tend to be quite distant from the examples seen in training. Given the inadequacy of standard techniques in these cases, we discuss a new model selection procedure that outperforms standard approaches on these hard tasks.

2 Bias/variance decomposition

This paper focuses on *least squares regression problems*. Here the goal is to learn a prediction function $h : X \to \mathbb{R}$ that minimizes the squared difference between predicted $\hat{y}$ and true y-values, as specified by the loss function $loss(\hat{y}, y) = (\hat{y} - y)^2$. Perhaps the simplest prototypical approach to this problem is to first conjecture a suitable class of hypothesis functions H (*e.g.*, by specifying a neural net architecture, or some other representation class), and then choose the hypothesis $h^* \in H$ that minimizes the empirical error $\widehat{err}(h^*) \;\triangleq\; \sum_{j=1}^{t}(h^*(x_j) - y_j)^2/t$ on the training set $S = \langle x_1, y_1\rangle, ..., \langle x_t, y_t\rangle$. This approach is known as *empirical error minimization*; *i.e.*, we implicitly consider a learning function L_H that maps training sets S into hypotheses $h^* = L_H(S)$ from H that obtain minimum error on S. Of course, the key to making such an approach work is to choose the right hypothesis class H.

One way to asses the suitability of a particular hypothesis class H is to consider the *expected* (true) error that results from minimizing empirical error on H. To formalize this, consider a fixed distribution P_{XY} on the space of examples $X \times Y$ and a training sample size t. Notice that this yields a particular distribution, P_{XY}^t, on training sequences of length t. Observe that each such training sequence S determines a particular hypothesis $h^* = L_H(S)$ that obtains minimum error on S. Thus, from the distribution over length t training sequences, we obtain an induced distribution P_H over hypotheses in H. Now notice that each of these hypotheses h^* has a true expected error with respect to the distribution of examples P_{XY}, given by $err(h^*) \triangleq \int_X \int_Y (h^*(x) - y)^2 \, d\mathrm{P}_{Y|x} d\mathrm{P}_X$. Therefore, the induced distribution over hypotheses generates a corresponding distribution over true error values. It is the *expected* value of this distribution,

$$\mathrm{E}_{h^*} err(h^*) \quad \triangleq \quad \int_H err(h^*) \, d\mathrm{P}_H \qquad (1)$$

that is our primary interest. That is, we are interested in the average (true) error of the hypotheses one obtains by minimizing empirical error on H, given that we train on t random examples drawn according to P_{XY} (where we average over hypotheses generated by possible training sets).

Clearly, our goal is to make this expectation as small as possible. That is, we wish to choose a class of hypotheses H that ensures small expected error relative to the unknown target distribution P_{XY}. However, there are two opposing forces to contend with here. If we make H too complex, we obtain a large expected error because similar training sets yield significantly different hypotheses (and not all of these can be simultaneously accurate). On the other hand, if we restrict H too severely, there might not be any good hypotheses left.

This tradeoff can be formalized in terms of the bias/variance decomposition of expected hypothesis error. It is well known that (1) can be decomposed into "bias" and "variance" terms by expanding around the *mean* hypothesis $\bar{h}^*$ of the distribution P_H (Geman, Bienenstock, & Doursat 1992). That is, if we define $\bar{h}^*$ to be the function such that $\bar{h}^*(x) = \int_H h^*(x) d\mathrm{P}_H$ (*i.e.*, on an input x, $\bar{h}^*$ outputs the average prediction of the h^*'s chosen according to P_H), we then obtain the decomposition

$$\mathrm{E}_{h^*} \mathit{err}(h^*) \;=\; \mathit{err}(\bar{h}^*) \;+\; \mathrm{E}_{h^*} \mathit{error}(\bar{h}^*, h^*),$$

where $\mathit{error}(\bar{h}^*, h^*) \triangleq \int_X (\bar{h}^*(x) - h^*(x))^2 \, d\mathrm{P}_X$ is the average discrepancy between the empirical hypotheses h^* and the mean hypothesis $\bar{h}^*$. Here the first and second terms are often referred to as the "bias" and "variance" respectively (of H with respect to P^t_{XY}).

Rather than using this particular decomposition, however, we will find it instructive to consider an alternative decomposition which expands (1) around the *optimal* hypothesis h^{opt} in H, rather than the mean hypothesis $\bar{h}^*$ defined by the distribution P_H. That is, if we define h^{opt} to be the function that obtains minimum *true* error relative to P_{XY} among all hypotheses in H, then an alternative decomposition of (1) can be shown to be

$$\mathrm{E}_{h^*} \mathit{err}(h^*) \;=\; \overset{\textit{“bias”}}{\mathit{err}(h^{opt})} \;+\; \overset{\textit{“variance”}}{\mathrm{E}_{h^*} \mathit{error}(h^{opt}, h^*)}, \tag{2}$$

where $\mathit{error}(h^{opt}, h^*) \triangleq \int_X (h^{opt}(x) - h^*(x))^2 \, d\mathrm{P}_X$ is the average discrepancy between the empirical hypotheses h^* and the class optimal hypothesis h^{opt}.[1] Thus, we can decompose the expected hypothesis error into two slightly different components: the true error of the *optimal* hypothesis in H (irreducible bias), and the average discrepancy between a random data generated hypothesis and this optimal hypothesis (variance).[2]

Now given this decomposition, consider the model selection task: For a given instance of a model selection problem we are given a nested *sequence* of hypothesis classes $H_0 \subset H_1 \subset ...$, and are faced with a particular example distribution P_{XY} and training sample size t. Note that for fixed P_{XY} and t we obtain specific bias and variance values, b_i and v_i, for each hypothesis class H_i. Thus, each instance of a model selection problem yields a particular *profile* of biases and variances over the sequence of hypothesis classes $H_1, H_2, ...$. Intuitively, we expect the variance terms to increase for larger hypothesis classes, as there are a wider variety of functions that give similar fits to the data. On the other hand, we expect the bias terms to decrease as we are better able to approximate the optimal regression for the given distribution. A model selection strategy needs to infer how the combination of bias + variance behaves, based on the structure of $H_1 \subset H_2 \subset \cdots$ and the training set errors $\widehat{err}(h_1^*), \widehat{err}(h_2^*)$...

By adopting the perspective that these bias and variance profiles capture the essential aspects of the task, we are able to make several useful predictions about the behavior of model selection strategies, as well as characterize the difficulty of model selection problems—based solely on the shapes of these bias and variance profiles, and disregarding other aspects of the problem.

3 Performance of penalization strategies

We begin by investigating the behavior of complexity-penalization strategies. Recall that for a training sample S and corresponding hypothesis sequence $h_1^*, h_2^*, ...$, a penalization strategy will choose the hypothesis h_i^* that minimizes some combination of class complexity c_i and empirical error $\widehat{err}(h_i^*)$. The point is that the empirical errors $\widehat{err}(h_i^*)$ tend to be gross underestimates of $\mathit{err}(h_i^*)$ in general (since the h_i^* are explicitly chosen to minimize the error on S), and the degree of underestimation tends to become worse at higher complexity levels. Complexity-penalization, therefore, seeks to adjust the empirical error estimates to compensate for this fact. This results in a generic model selection strategy where one first penalizes the empirical errors to obtain better estimates

$$\widehat{err}_{pen}(h_i^*) \;=\; \widehat{err}(h_i^*) \;+\; \mathit{penalty}_i \tag{3}$$

and then chooses the hypothesis h_i^* with the smallest adjusted estimate $\widehat{err}_{pen}(h_i^*)$.

As mentioned, there are many variants of this strategy, but to illustrate our main points it will suffice to consider two strategies that embody distinct penalization policies. To describe these strategies, let $r = i/t$ be the number of complexity levels being considered per training example.[3] The first penalization strategy we consider is Generalized Cross Validation GCV (Craven & Wahba 1979). Following (Moody & Utans 1992) we can write the adjusted error estimate of this strategy as

$$\widehat{err}_{\mathsf{GCV}}(h_i^*) \;=\; \widehat{err}(h_i^*) \;+\; \frac{2r - r^2}{(1 - r)^2} \widehat{err}(h_i^*).$$

The other penalization strategy we consider is Vapnik's Structural Risk Minimization procedure SRM (Vapnik 1996), which following (Cherkassky, Mulier, & Vapnik

[1] We require some technical conditions to yield this decomposition. In particular, we require that H be closed under linear combinations of functions (as well as Cauchy sequences). This is sufficient to ensure that H is a closed linear subspace of a Hilbert space defined by the inner product $\langle f, g \rangle \triangleq \int_X (f(x) - g(x))^2 d\mathrm{P}_X$; see, *e.g.,* (Ash 1972, Chapter 3). Given these conditions, we can apply the relevant projection theorem to obtain h^{opt}, and the subsequent analysis becomes a simple consequence of generalized Pythagorean relations. Fortunately, this technical condition holds for most hypothesis classes normally considered in practice; including (obviously) linear regression functions, as well as any neural network regressor that uses linear output units.

[2] Note that h^{opt} and $\bar{h}^*$ actually coincide for linear regression when there is a linear generating model and zero mean noise.

[3] For most natural orderings $H_1 \subset H_2 \cdots$, the complexity level i corresponds to the number of free parameters used in the definition of function class H_i. Therefore, intuitively r gives the number of distinct parameters being estimated per training example (Cherkassky, Mulier, & Vapnik 1996; Vapnik 1996).

1996) can be formulated

$$\widehat{err}_{\mathsf{SRM}}(h_i^*) \;=\; \widehat{err}(h_i^*) \;+\; \frac{\sqrt{\tilde{r}}}{\left(1-\sqrt{\tilde{r}}\right)_+}\widehat{err}(h_i^*),$$

where $\tilde{r} = r(1 + \ln 1/r) + (\ln t)/2t$, and $(\cdot)_+$ denotes the positive threshold function; *i.e.*, $(r)_+ = r$ if $r \geq 0$; $(r)_+ = 0$ if $r < 0$. For our purposes, the key difference between these two policies is that SRM uses a much steeper penalization profile than GCV. (See Figures 1–4 below.)

Now reconsider the bias/variance characterization developed above. This offers an interesting interpretation of complexity-penalization methods. This can be seen by directly comparing equations (2) and (3) and noting that the first terms can be naturally aligned. Notice here that, although $\widehat{err}(h_i^*)$ is normally considered to be a direct (but poor) estimate of h_i^*'s true error, we can alternatively view $\widehat{err}(h_i^*)$ as an estimate of the true error of the *optimal* hypothesis in the class, h_i^{opt}. In fact, $\widehat{err}(h_i^*)$ is typically a much *better* estimate of $err(h_i^{opt})$ than it is of $err(h_i^*)$! To see this, consider Figure 0 which depicts the relationship between the training set estimate $\widehat{err}(h_i^*)$ and the fixed quantities $err(h_i^*)$ and $err(h_i^{opt})$. First notice that $\widehat{err}(h_i^*) < \widehat{err}(h_i^{opt})$, since h_i^* is explicitly chosen to minimize $\widehat{err}$. However, notice also that h_i^{opt} is a *fixed* hypothesis which has not been chosen as a function of S, and therefore we know that after relatively few training examples we will have $\widehat{err}(h_i^{opt}) \approx err(h_i^{opt})$ with high probability. Thus, combining this with the fact that $err(h_i^{opt}) < err(h_i^*)$ by the definition of h^{opt}, we obtain the chain of inequalities

$$\widehat{err}(h_i^*) \;<\; \widehat{err}(h_i^{opt}) \;\overset{\mathrm{P}}{\approx}\; err(h_i^{opt}) \;<\; err(h_i^*).$$

This shows that $\widehat{err}(h_i^*)$ must be closer to $err(h_i^{opt})$ than $err(h_i^*)$ with high probability after relatively few training examples.[4] In fact, the superiority of interpreting $\widehat{err}(h_i^*)$ as an estimate of $err(h_i^{opt})$ rather than $err(h_i^*)$ can be easily demonstrated experimentally, as shown in Table 5 below.

Although not often explicitly made, this elementary observation leads to an interesting interpretation of penalization strategies: if the empirical error term $\widehat{err}(h_i^*)$ accurately estimates the the bias term for class H_i, then the *penalty*$_i$ term must be accounting for the unobserved *variance* of H_i. Thus, we can interpret the sequence of penalty terms, *penalty*$_1$, *penalty*$_2$, ..., as in effect postulating a particular profile of variance terms for the classes $H_1, H_2, \ldots$. So for example, a steep penalization profile encodes the assumption that the variances grow rapidly as a function of complexity level i, whereas a flat profile asserts that the variances grow more slowly. This observation leads to a series of specific predictions about the behavior of penalization strategies: (1) if the penalization profile is much steeper than the true variance profile, we expect systematic *underfitting* since the latter hypotheses will be over-penalized relative to the true variances; (2) on the other hand, if the penalization profile is much flatter than the true variance profile, we expect systematic *overfitting* since the latter hypotheses will be under-penalized; and finally (3) we expect good generalization performance if the penalty profile matches the true variance profile for the task.

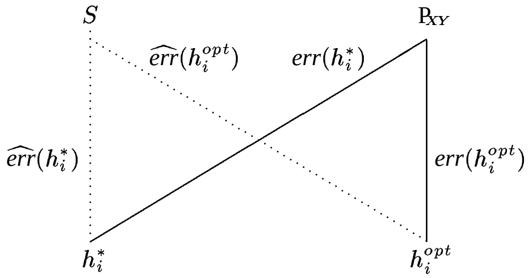

Figure 0: Illustrating the relation between the true error and estimated error of the empirically best function h_i^* and of the true best function h_i^{opt}. Solid lines indicate the true errors determined by P_{XY}, and dashed lines indicate the empirical errors obtained on a training sample S.

Experiment To test these hypotheses we ran a series of experiments to investigate the behavior of GCV and SRM on model selection tasks with different bias and variance profiles. Recall that GCV and SRM propose very different penalization policies and therefore we expect them to behave quite differently as we vary the task structure. To conduct our experiments we considered a traditional linear regression problem where the goal is to learn a linear function $h(x_1, ..., x_n) = a_1x_1 + \cdots + a_nx_n$ that minimizes the mean squared error on an unknown P_{XY}. In this context, a natural model selection task arises by considering the nested sequence of function classes $H_1 \subset H_2 \subset \cdots$ defined by the first $1, 2, \ldots$ variables respectively (which assumes in effect that the variables have been ordered by importance). To design test problems, we set $n = 10$, $t = 20$, and considered a series of distributions P_{XY} that yield different bias and variance profiles for the task. Specifically, we used distributions defined by a simple additive model $Y = \alpha_1X_1 + \cdots + \alpha_nX_n + \varepsilon$, where the X_i's and ε are independent and $\varepsilon \sim N(0, \sigma^2)$. We generated X_i's by a Cauchy$(0, 1)$ distribution, which was then truncated at $(-\beta_i, +\beta_i)$ for different choices of β_i. We also set the linear model coefficients to be $\alpha_i = 1/\beta_i$, to normalize the X_i variances. Thus, our test distributions P_{XY} were determined by σ and the truncation constants $\beta_1, \ldots, \beta_{10}$.

The reason for using these Cauchy-like distributions instead of more conventional Gaussians is that we wished to construct *difficult* model selection problems. That is, wide-tailed distributions like Cauchy create difficult variable selection problems, because small training samples will not

[4] This argument could be formalized into precise quantitative statements, for example, by an elementary application of Hoeffding-Chernoff bounds (Hoeffding 1963), but we do not pursue this here. The intuition is clear in any case.

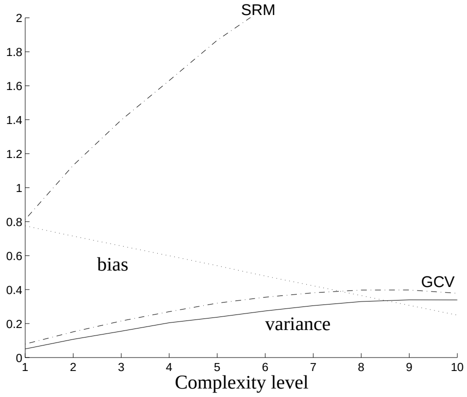

Figure 1: Bias and variance profiles for Problem 1 (*flat* variance), showing the corresponding penalty profiles used by GCV and SRM.

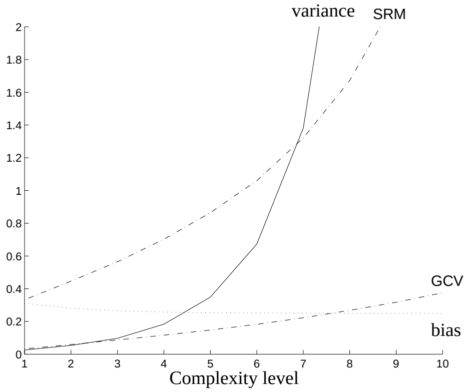

Figure 2: Bias and variance profiles for Problem 2 (*steep* variance), showing the corresponding penalty profiles used by GCV and SRM.

	mean	percentiles of error ratios after 1000 repetitions				mean compl.
	ratio	50	75	95	100	diff.
GCV	*1.90*	*1.421*	2.114	4.357	37.47	-0.962
SRM	*2.86*	*2.287*	3.383	6.567	37.47	-5.834
VAR	*1.60*	*1.057*	1.677	3.584	37.47	0.030
10CV	*2.08*	*1.610*	2.469	4.657	12.73	-1.812
ADJ	*2.00*	*1.544*	2.358	4.492	35.74	-2.223
$\widehat{\text{ADJ}}$	*2.02*	*1.572*	2.328	4.568	35.74	-2.266

Table 1: Results for Problem 1—*flat* variance profile.

	mean	percentiles of error ratios after 1000 repetitions				mean compl.
	ratio	50	75	95	100	diff.
GCV	*73.1*	*1.745*	4.886	248.5	8007	0.605
SRM	*2.3*	*1.613*	2.274	4.41	117	-1.151
VAR	*1.9*	*1.454*	2.077	4.07	33	-0.830
10CV	*17.8*	*1.643*	3.021	26.97	2745	0.009
ADJ	*1.5*	*1.229*	1.724	3.10	8	-0.550
$\widehat{\text{ADJ}}$	*1.8*	*1.252*	1.798	3.66	34	-0.286

Table 2: Results for Problem 2—*steep* variance profile.

accurately capture the significant range of X_i values that will be observed in testing. Therefore small errors in $\widehat{\alpha}_i$ result in hypotheses with huge test set errors, since we evaluate these functions on large unobserved X_i values. In this way we achieve a large *variance* between hypotheses.

For these tasks, we evaluated model selection strategies by measuring the *ratio* of the true error of the hypotheses h_i^* they chose to the true error of the best hypothesis in the sample-dependent sequence $h_1^*, h_2^*, ...$. (The rationale for this is that we wish to measure the selection strategy's ability to approximate the best hypothesis in the given sequence—not find a better function from outside the sequence.) We ran our experiments by fixing a distribution P_{XY}, repeatedly generated training samples of size $t = 20$, and recording the ratio of chosen to best-in-sequence errors achieved by each strategy. This was repeated 1000 times to estimate the performance of the model selection strategies, as well as to estimate the bias/variance characteristics of the given problem.

The first problem we considered, shown in Figure 1, was designed to have a *flat* variance profile comparable in size to the bias profile (defined by setting $\sigma = 0.5$, $\beta_i = 10$). Here we expect GCV to outperform SRM, since its penalization profile more closely matches the true variance profile of this task (Figure 1). In fact our results show exactly this. Table 1 shows that GCV significantly outperforms SRM at this task, obtaining a mean approximation ratio of 1.9 over 1000 trials, compared to 2.9 obtained by SRM. (The other strategies mentioned in Table 1 are explained below.) That is, GCV chose a function from the sample-determined sequence $h_1^*, h_2^*, ...$ that had a true error 1.9 times larger than the best true error of any function in the sequence, on average. Moreover, GCV chose functions at complexity levels that were close to the optimum complexity levels for the given training sets—the last column of Table 1 shows that GCV underestimated the best complexity level by only 1.0 on average. For this problem SRM significantly underfit the data, choosing function complexities that were 5.8 levels smaller than optimum complexity on average. These results support our predictions based on the variance and penalization profiles involved.

We next considered a problem that had a much steeper variance profile, more closely resembling the penalization profile of SRM (defined by setting $\sigma = 0.5$, $\beta_i = 10 \times 2^{i-1}$); see Figure 2. In sharp contrast to the previous results, Table 2 shows that SRM significantly outperforms GCV in this case, achieving a mean approximation ratio of 2.3 versus GCV's mean ratio of 73.1. The last column in Table 2 also shows that GCV now overshoots the best complexity by an average of 0.6 levels, which leads to devastating consequences given the sharply increasing variances in

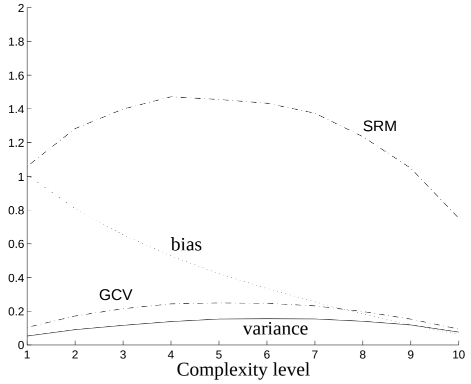

Figure 3: Bias and variance profiles for Problem 3 (*low* variance), showing the corresponding penalty profiles used by GCV and SRM.

	mean	percentiles of error ratios				mean
		after 1000 repetitions				compl.
	ratio	50	75	95	100	diff.
GCV	*1.22*	*1.0*	1.000	2.246	12.97	-0.111
SRM	*2.46*	*1.0*	2.151	8.985	51.12	-0.806
VAR	*1.03*	*1.0*	1.000	1.054	3.17	0.071
10CV	*1.47*	*1.0*	1.000	3.686	14.25	-0.246
ADJ	*2.10*	*1.0*	2.220	6.325	27.16	-0.630
$\widehat{\text{ADJ}}$	*2.13*	*1.0*	2.175	6.293	43.10	-0.620

Table 3: Results for Problem 3—*low* variance profile; easy problem.

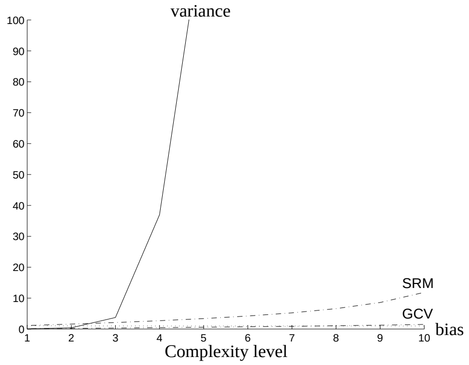

Figure 4: Bias and variance profiles for Problem 4 (*extreme* variance), showing the corresponding penalty profiles used by GCV and SRM.

	mean	percentiles of error ratios				mean
		after 1000 repetitions				compl.
	ratio	50	75	95	100	diff.
GCV	*2118*	*1.49*	11.4	9205	2.3×10^5	0.658
SRM	*198*	*1.07*	1.78	5.37	1.1×10^5	-0.366
VAR	*1.73*	*1.00*	1.70	3.26	75.3	-0.502
10CV	*869*	*1.45*	4.04	1482	2.3×10^5	0.308
ADJ	*1.31*	*1.00*	1.32	2.59	9.48	-0.257
$\widehat{\text{ADJ}}$	*23*	*1.00*	1.53	8.63	13,422	0.069

Table 4: Results for Problem 4—*catastrophic* variance profile; hard problem.

this task. Notice that here SRM avoids these large errors by systematically *underfitting* the data (by an average of 1.2 complexity levels), and thus avoids the more complex hypotheses that cause big problems. Thus, it seems that *under*fitting the data has far less severe consequences than overfitting in this case. In fact, there seems to be an inherent asymmetry here which favors conservative (steep) penalization methods over credulous (flat) penalizers; *i.e.*, the losses associated with overfitting are far more significant than underfitting if the variances are sharply increasing.

These results show that there is, in general, no *best* penalization method. The performance one obtains depends on how closely the penalization profile of the strategy matches the true variances of the task. If the penalty terms are much larger than the variances, systematic underfitting results; whereas if the penalties are much smaller, the data are systematically overfit.[5] Not surprisingly, directly penalizing by the true variances of the task (Procedure VAR) always seems to yield good performance for any slope of variance profile. *E.g.*, Tables 1 and 2 show that VAR performs well on Problems 1 and 2, even though the variance profiles behave quite differently in these two cases. Of course, VAR avoids systematic over or under-fitting by using different penalization profiles for each problem, and this is certainly not achievable in practice. However, this suggests that one should try to directly penalize according to the true variances of the task as much as possible.

Overall, these results suggest that one can interpret penalizers as asserting a particular structure for the problem: The postulated penalization profile makes a specific assumption about the behavior of the variances in the given task. The obvious conclusion is that one should set the penalty terms according to whatever prior knowledge one has about the variance profile for the task at hand. Accurate assumptions tend to yield excellent generalization performance, whereas inaccurate assumptions lead to poor performance. However, we will see that there are situations where one might not want to use penalty methods regardless.

4 Difficulty of model selection problems

The bias/variance decomposition can also be used to characterize the notion of *hard* versus *easy* model selection problems. Specifically, in terms of our previous definitions, we find that if the variance profile is flat (grows slowly

[5] Note that this is similar to an observation made by Kearns *et al.* (Kearns *et al.* 1995) in the context of learning classifications. However, they do not explicitly invoke a bias/variance characterization of model selection problems to explain their results.

	hypothesis class (# of variables)			
interpretation	$i = 1$	5	7	10
$\widehat{err}(h_i^*)$ ests. $err(h_i^*)$	0.042	119	782	1632
$\widehat{err}(h_i^*)$ ests. $err(h_i^{opt})$	0.036	0.048	0.069	0.12
	variance on 1000 trials			

Table 5: Empirical comparison of the two interpretations of $\widehat{err}(h_i^*)$: *(A)* as a direct estimate of its true error $err(h_i^*)$, versus *(B)* as an estimate of the class bias $err(h_i^{opt})$. Shows variances obtained for hypothesis classes in Problem 4.

and is not large in comparison to the bias profile) then almost any sensible penalization strategy will do reasonably well. On the other hand, if the variance profile grows explosively relative to the bias profile, then disaster results for any penalization strategy that does not use the exact variance profile for the task (or at least a sufficiently steep profile). These difficult problems occur naturally in regression whenever the x values have wide-tailed distributions—for example, as occurs with Cauchy distributions or in polynomial regression problems.

Experiment To demonstrate the distinction between easy and hard problems, we conducted a series of experiments in the same setup as before. The first case we considered was a model selection problem which had a *low* variance profile in relation to the bias terms (defined by setting $\sigma = 0.1$, $\beta_i = i$; Figure 3). We expect such a problem to be easy for most reasonable selection strategies, since the variances play a minor role and there are no serious consequences to minor over or under-fitting. Table 3 demonstrates the relatively benign behavior of the penalization strategies on this task; although the variance profile distinctly favors GCV in this case and this is reflected in the results.[6]

To contrast with this, we next considered a problem (Figure 4) which had a variance profile that grows explosively in complexity of the hypothesis class (defined by setting $\sigma = 1$, $\beta_i = 10^i$, and $\alpha_i = (1/\beta_i)^{3/4}$). We expect this to give a *hard* model selection problem because of the drastic consequences that would befall even minor overfitting. Table 4 shows that both GCV and SRM fail badly at this task. Both strategies make *catastrophic* mistakes from time to time, choosing hypotheses that are many orders of magnitude worse than the best available. Interestingly, Procedure VAR, which penalizes according to the true variances of the task, still works reasonably well in this case (Table 4). But of course VAR is not a practically realizable strategy. Overall, we found that model selection problems of this type tend to be inherently difficult for penalization strategies. In fact, we tried an entire suite of penalization methods on this task and obtained uniformly poor performance. These included Akaike's AIC, Schwarz's BIC, and Mallow's C_p, among others; see *e.g.*, (Foster & George 1994; Cherkassky, Mulier, & Vapnik 1996) for a discussion of several such methods.

These results lead us to conclude that complexity-penalization can be an inherently risky strategy. There seems to be a potential for disaster whenever the task happens to be hard; *i.e.*, whenever the variance profile grows explosively in an unpredictable manner.

Alternative hold-out methods An obvious idea in these situations is to consider alternative hold-out–based methods, like 10-fold cross-validation (10CV) or some other resampling procedure (Kohavi 1995; Weiss & Kulikowski 1991). The common folklore surrounding these techniques is that they can often be better behaved than penalization methods. However, it turns out that these strategies are prone to the very same drastic mistakes suffered by penalty-based methods, as Table 4 clearly demonstrates for 10CV. The strikingly bad performance obtained by all standard model selection methods on these difficult tasks raises the question of whether it is possible to do better on hard problems, or whether we have to live with the potential of making disastrous mistakes.

5 A new model selection technique

In a recent paper, (Schuurmans 1997), one of the authors introduces a new strategy for model selection that takes a fundamentally different approach to the problem than previous techniques. This new strategy seems to avoid many of the catastrophic overfitting errors that plague standard complexity-penalization and hold-out methods on difficult model selection tasks. This procedure implicitly attempts to estimate the variance of a function class H_j by examining how h_j^* compares to h_i^* for $i < j$.

The basic idea behind this new strategy is to exploit the intrinsic geometry of the function learning task which arises from a simple statistical model of the problem: Assume the training and test examples are independent random observations drawn from a joint distribution P_{XY} on $X \times Y$. Then we can decompose this distribution into the conditional distribution of Y given X, $P_{Y|X}$, and the marginal distribution P_X on X. Note that when learning a function $h : X \to Y$ we are really only interested in approximating the conditional $P_{Y|X}$. However our approach is to exploit knowledge about P_X to help us make better decisions about which hy-

[6] It might seem surprising at first that the penalty and variance terms can actually *decrease* on these problems, as shown in Figure 3. However, this is a consequence of the fact that the penalty terms depend on the training set errors, which can decrease faster than the multiplicative adjustments used by GCV and SRM. For the variances, the easiest way to see how they can decrease is to imagine a case where Y is a deterministic linear function of the variables. Here, any linearly independent set of training examples determines the target function exactly, and thus we would observe *zero* variance if given a linearly independent set (which could happen with probability 1 for $t >= n$). The additive noise component ε has the effect of monotonically increasing these variances, in counterbalance.

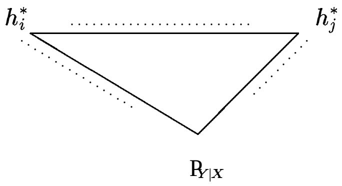

Figure 5: The real and estimated distances between successive hypotheses h_i and h_j and the target $P_{Y|X}$. Solid lines indicate real distances, dashed lines indicate empirical distance estimates.

pothesis h to choose. In fact, for now, assume that we actually *know* P_X and see how far this gets us. (Note that any information we require about P_X can be obtained from *unlabeled* training examples.)

Knowing P_X is important because it allows us to define a natural metric $d(h, g) \triangleq \left(\int_X (h(x) - g(x))^2 \, dP_X\right)^{1/2}$ on the space of hypotheses that measures the "distance" between two hypotheses h and g. Moreover, we can extend this definition to include the target conditional $P_{Y|X}$ via the definition $d(h, P_{Y|X}) \triangleq \left(\int_X \int_Y (h(x) - y)^2 \, dP_{Y|x} dP_X\right)^{1/2}$; which means that we can interpret the true error of a function h as the *distance* between h and the target object $P_{Y|X}$. Importantly, these definitions are compatible in the sense that the defined distance measure d satisfies the standard (pseudo) metric axioms over $H \cup \{P_{Y|X}\}$. This gives us a nice geometric view of the problem: We are given a nested sequence of spaces $H_0 \subset H_1 \subset ...$, each with a closest function $h_0, h_1, ...$ to the target $P_{Y|X}$, where the distances are decreasing. However, we do not observe these real distances. Rather, we are given a training sample $S = \langle x_1, y_1 \rangle, ..., \langle x_t, y_t \rangle$, and have to choose from the sequence of *empirically* closest functions $h_0^*, h_1^*, ...$, which have monotonically decreasing distance estimates $d(\widehat{h, P_{Y|X}}) \triangleq \left(\sum_{j=1}^t (h(x_j) - y_j)^2 / t\right)^{1/2}$ on S. The key point though is that we now have more information at our disposal: not only do we have estimated distances to $P_{Y|X}$, we now know the true distances *between* functions in the sequence!

Our idea is to use this additional information to choose a better hypothesis. In fact, notice that we are dealing with two metrics here: the true metric d defined by the joint distribution P_{XY} and an empirical metric $\hat{d}$ determined by the labeled training sequence. Given these two metrics, consider the triangle formed by two hypotheses h_i^* and h_j^* and the target conditional $P_{Y|X}$ (Figure 5). Note that there are six distances involved, three real and three estimated—of which the true distances to $P_{Y|X}$ are the only two we care about, and yet these are the only two we don't have! The key observation though is that the real and estimated distances between hypotheses $d(h_i^*, h_j^*)$ and $d(\widehat{h_i^*, h_j^*})$ give us an *observable* relationship between d and $\hat{d}$ in the local vicinity. In fact, we can adopt the naive assumption that observed relationship between h_i^* and h_j^* also holds between h_j^* and $P_{Y|X}$. Note that if this were the case, we would obtain a better estimate of $d(h_j^*, P_{Y|X})$ simply by adjusting the training set distance $d(\widehat{h_j^*, P_{Y|X}})$ according to the observed ratio $d(h_i^*, h_j^*)/d(\widehat{h_i^*, h_j^*})$.[7] In fact, adopting this as a simple heuristic leads to a surprisingly effective model selection procedure (ADJ): given the hypothesis sequence $h_1^*, h_2^*, ...$, first multiply each estimated distance $d(\widehat{h_j^*, P_{Y|X}})$ by the largest observed ratio $d(h_i^*, h_j^*)/d(\widehat{h_i^*, h_j^*})$, $i < j$, and then choose the function in the sequence with the smallest *adjusted* distance estimate to $P_{Y|X}$. (Note that this adjustment to h_j^*'s distance can be interpreted as an estimate for the variance of H_j, indirectly achieved by referring to $H_i \subset H_j$.)

Experiment Tables 1–4 show that this technique does indeed work effectively on the model selection problems considered here. In particular, Table 4 shows that ADJ completely avoids the catastrophic mistakes made by the standard model selection strategies, and even outperforms the ideal variance penalizer VAR. This is somewhat surprising since VAR exploits exact knowledge of the true variances for the task. However, the reason for VAR's failure is that it does not pay explicit attention to the inter-hypothesis distances, and can therefore sometimes be fooled. Of course, we do not expect a free lunch in general (Schaffer 1993), and there are certainly model selection problems where ADJ does not dominate, *e.g.*, Table 3. However, one should be able to exploit additional information about the task (here knowledge of P_X) to obtain significant improvements across a wide range of problem types and conditions. Our empirical results support this view for the case of hard model selection tasks. (Further support to this claim is provided in (Schuurmans 1997) which considers a different class of polynomial curve-fitting problems.)

To summarize, the new metric-based technique ADJ appears to effectively avoid dangerous under and over-fitting, and provides a safe and responsive model selection strategy, at least for the regression problems considered here. Interestingly, the performance of ADJ does not seem to degrade too severely when we move to consider hard model selection problems, even when these hard problems cause tremendous difficulty for standard techniques.

Of course, one can always argue that these results are not terribly useful since the metric-based strategy ADJ requires knowledge of the true domain distribution P_X. This is clearly an unreasonable assumption in practice. However, one can obtain information about P_X from *unlabeled* training instances. In fact, many important function learning applications have large corpora of unlabeled training data

[7] Note that since we expect $\hat{d}$ to be an underestimate in general, we expect this ratio to be typically larger than 1.

available (*e.g.*, image, speech and text databases), so these metric-based techniques could still apply to a wide range of practical situations—provided they are robust to using only *estimated* distances. In fact, ADJ turns out to be reasonably robust to using approximate distances. Tables 1–3 show that as few as 100 reference examples were sufficient for the approximate $\widehat{\mathsf{ADJ}}$ procedure to perform nearly as well as ADJ (except for the difficult problem in Table 4). Finally, note that this still yields a reasonably efficient model selection procedure since computing inter-hypothesis distances involves making only a single pass down the reference list of unlabeled examples. This is a strong advantage over standard hold-out techniques like 10CV which repeatedly call the hypothesis generating mechanism to generate pseudo-hypotheses.

6 Conclusions

We considered a simple characterization of model selection problems based on the standard bias/variance decomposition of expected hypothesis error. This analysis allows us to make predictions about and distinguish the performance of different model selection strategies based on two simple but essential aspects of the task: the shapes of the bias and variance profiles generated across the sequence of hypothesis classes. With this characterization, we distinguished between easy and hard model selection problems. This distinction is important because difficult model selection problems arise in fairly natural conditions. For example, we demonstrated this for regression where the independent variables are drawn from wide-tailed distributions. In such cases, prediction variance increases sharply as additional terms are added to the model, since the x's in the out-of-sample testing data sets may be far from those in the training set. Another example where steep variance profiles occur is polynomial curve-fitting (Cherkassky, Mulier, & Vapnik 1996; Vapnik 1996).

These observations lead to specific recommendations: First, one should use as much prior knowledge as possible about the shape of the variance profile to choose a model selection policy that works effectively while avoiding disastrous mistakes. For example, in the case of steep variance profiles, standard complexity penalization methods do not penalize sufficiently, which leads to disastrous results. Second, the new metric-based model selection strategies seem to be much more robust against catastrophic overfitting errors than standard techniques, and apparently can be usefully applied in difficult cases.

Among the many avenues for future work, we are currently extending the same style of bias/variance analysis to *classification* (as opposed to regression) problems (Kearns *et al.* 1995). Note that the decomposition of prediction error into additive bias and variance components is not so obvious for classification however (Kohavi & Wolpert 1996).

Acknowledgements

Thanks to Adam Grove for many important suggestions that improved this paper and the ideas therein. Helpful comments were also received by two anonymous referees. The first author gratefully acknowledges the support of IRCS.

References

Ash, R. B. 1972. *Real Analysis and Probability*. San Diego: Academic Press.

Cherkassky, V.; Mulier, F.; and Vapnik, V. 1996. Comparison of VC-method with classical methods for model selection. Preprint.

Craven, P., and Wahba, G. 1979. Smoothing noisy data with spline functions. *Numer. Math.* 31:377–403.

Efron, B. 1979. Computers and the theory of statistics. *SIAM Review* 21:460–80.

Foster, D., and George, E. 1994. The risk inflation criterion for multiple regression. *Ann. Statist.* 22:1947–75.

Geman, S.; Bienenstock, E.; and Doursat, R. 1992. Neural networks and the bias/variance dilemma. *Neural Comp.* 4:1–58.

Hoeffding, W. 1963. Probability inequalities for sums of bounded random variables. *JASA* 58(301):13–30.

Kearns, M.; Mansour, Y.; Ng, A.; and Ron, D. 1995. An experimental and theoretical comparison of model selection methods. In *Proceedings COLT-95*.

Kohavi, R., and Wolpert, D. 1996. Bias plus variance decomposition for zero-one loss functions. In *Proceedings ML-96*.

Kohavi, R. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings IJCAI-95*.

Moody, J., and Utans, J. 1992. Principled architecture selection for neural networks: Application to corporate bond rating prediction. In *Proceedings NIPS-4*.

Moody, J. 1992. The effective number of parameters: An analysis of generalization and regularization in nonlinear learning systems. In *Proceedings NIPS-4*.

Rissanen, J. 1986. Stochastic complexity and modeling. *Ann. Statist.* 14:1080–100.

Schaffer, C. 1993. Overfitting avoidance as bias. *Mach. Learn.* 10(2):153–78.

Schuurmans, D. 1997. A new metric-based approach to model selection. In *Proceedings AAAI-97*. To appear.

Vapnik, V. 1982. *Estimation of Dependences Based on Empirical Data*. New York: Springer-Verlag.

Vapnik, V. 1996. *The Nature of Statistical Learning Theory*. New York: Springer-Verlag.

Weiss, S. M., and Kulikowski, C. A. 1991. *Computer Systems that Learn*. San Mateo: Morgan Kaufmann.