So the VC-dimension is, indeed, different from the number of parameters (or entities) that is used to measure model complexity in statistical methods. We should also bear in mind that for many parameterizations analytic estimates of the VC-dimension are not known and hard to come by. In particular, the VC-dimension may depend on nonlinear optimization method used in many existing learning methods.

Further, we have only shown estimates for binary indicator functions, appropriate for *classification* problems. For *regression* problems, we need to estimate the VC-dimension for real-valued sets of functions. Fortunately, these estimates are identical to the corresponding estimates for indicator functions. For example, consider the set of third order polynomials for estimating univariate regression model: $f(x, \mathbf{w}, b) = w_3 x^3 + w_2 x^2 + w_1 x + b$. This parameterization is a linear combination of fixed basis functions (monomials) as in Example 4.7, so for this set of real-valued functions, the VC-dimension equals the number of free parameters, i.e. $h = 4$. Qualitatively, for regression problems, the VC-dimension measures the ability of real-valued functions to fit the training data. In the case of regression, the perfect fit corresponds to the model passing through all training samples. For example, 1-nearest neighbor regression always fits the training data, for any number of samples (see example in Fig. 2.10). Hence, for 1-nearest neighbor regression the VC-dimension is infinite.

4.4 GENERALIZATION BOUNDS

Recall that generalization is possible only if the empirical risk is sufficiently close to unknown prediction risk, as shown earlier in Fig. 4.4. The VC-theory quantifies this closeness, by providing different generalization bounds (aka VC-bounds). These bounds answer the following two questions:

- How close is the risk $R(\omega_n^*)$ to the minimal empirical risk $R_{emp}(\omega_n^*)$?
- How close is the risk $R(\omega_n^*)$ to the minimal possible risk $R(\omega_0) = \min_{\omega} R(\omega)$?

In this book, we describe only the first type of bounds that relate the unknown prediction risk $R(\omega_n^*)$ to the empirical risk $R_{emp}(\omega_n^*)$ and

the VC-dimension of a set of admissible models $f(\mathbf{x}, \omega)$. This bound can be visualized as the distance between the two curves $R(\omega_n^*) - R_{emp}(\omega_n^*)$, as shown in Fig. 4.4. These bounds may have both conceptual and practical value. Conceptually, the VC-bounds are important for understanding the connection between generalization and model complexity, described in Section 2.4. This connection leads to a new inductive principle called Structural Risk Minimization, described in Section 4.5. On a practical side, several examples of model selection for regression using VC-bounds are presented later in this section.

The particular form of VC-bounds depends on the properties of a loss function, the properties of a set of possible models $f(\mathbf{x}, \omega)$, and, sometimes, on the general properties of unknown distribution $P(\mathbf{x}, y)$. We are only interested in *distribution-independent* bounds, which do not incorporate any knowledge about unknown distribution. Next, we present two distribution-independent bounds:

- for classification problems with bounded loss (e.g., 0/1 loss);
- for regression problems with unbounded loss.

VC-bound for classification:

For the binary classification problem, the following bound for generalization ability of a learning method (implementing ERM) holds with probability of at least $1 - \eta$ for all admissible functions $f(\mathbf{x}, \omega)$, including the function $f(\mathbf{x}, \omega_n^*)$ that minimizes empirical risk:

$$R(\omega) \le R_{emp}(\omega) + \Phi\left(\frac{h}{n}, \frac{\ln \eta}{n}\right) \tag{4.11a}$$

where

$$\Phi\left(\frac{h}{n}, \frac{\ln \eta}{n}\right) = \sqrt{\frac{h\left(\ln \frac{2n}{h} + 1\right) - \ln(\eta/4)}{n}} \tag{4.11b}$$

The term Φ is called the *confidence interval*, since it estimates the difference between the training error and the true error of a classifier. Note that the bound is probabilistic, because it describes random quantities, training error $R_{emp}(\omega)$ and test error $R(\omega)$, that depend on a random realization of training data. The probabilistic *confidence level*

$1-\eta$ can be set high, but this will yield large value of Φ according to (4.11b), making the bound very crude. The probabilistic nature of VC-bounds reflects the trade-off between the accuracy provided by the bounds and the confidence in these bounds. Intuitively, the confidence level can be high only for *sufficiently large* sample size, because it is impossible to make a valid generalization from small samples. A practical prescription for setting the confidence level as a function of sample size is:

$$\eta = \min\left(\frac{4}{\sqrt{n}}, 1\right) \tag{4.12}$$

Let us analyze the behavior of Φ as a function of sample size n, with all other parameters fixed. Expression (4.11b) shows strong dependency of the confidence interval Φ on n/h, the ratio of the number of training samples to the VC-dimension. So we can distinguish two main regimes: (1) small (or finite) sample size, when the ratio n/h is small (e.g., less than 15), and (2) large sample size, when this ratio is large. For the large sample size, the value of the confidence interval Φ becomes small, and the empirical risk can be safely used as a measure of prediction risk. In this case, application of classical parametric estimation methods (using ERM) will work well. On the other hand, with small samples the value of the confidence interval cannot be ignored, and there is a need to match complexity (capacity) of admissible functions to available training data. This is achieved using the SRM inductive principle introduced in Section 4.5.

Further, note that VC-bounds for classification and regression hold for all admissible functions $f(\mathbf{x}, \omega)$, not just for models minimizing the empirical risk. Hence, these bounds can be used for many learning algorithms that do not explicitly minimize the empirical risk. This often happens with practical learning methods that:

- perform nonlinear optimization where only locally optimal solutions can be found;
- minimize a loss function that is different from the loss function used to define empirical risk.

VC-bound for regression:
For the regression learning task, the following bound for generalization ability of a learning method (implementing ERM) holds with probability

of at least $1-\eta$ for all admissible functions $f(\mathbf{x},\omega)$, including the function $f(\mathbf{x},\omega_n^*)$ that minimizes empirical risk:

$$R(\omega) \le \frac{R_{emp}(\omega)}{(1-\sqrt{\varepsilon})_+} \tag{4.13a}$$

where

$$\varepsilon\left(\frac{h}{n},\frac{\ln(\eta/4)}{n}\right)=\frac{h\left(\ln\frac{n}{h}+1\right)-\ln(\eta/4)}{n} \tag{4.13b}$$

Note that the bound for regression (4.13a) has a multiplicative form, in contrast to an additive form of the bound (4.12) for classification. Further, when the value of ε approaches 1, the denominator in (4.13a) becomes close to zero and the bound goes to infinity. Detailed analysis of expression (4.13b) shows that the bound goes to infinity for sufficiently large values of h/n, so that the condition $h \le 0.5n$ should hold for *any* learning method. Further, by setting the value of the confidence level $1-\eta$ in (4.13) according to prescription (4.12), we can obtain the following practical form of the VC-bound for regression:

$$R(\omega) \le R_{emp}(\omega)\left(1-\sqrt{p-p\ln p+\frac{\ln n}{2n}}\right)_+^{-1} \tag{4.14}$$

where $p = h/n$. Note that the VC-bound (4.14) has the same form as classical statistical methods for model selection presented in Section 2.4. Using notation from Section 2.4, the practical VC-bound (4.14) specifies the VC penalization factor:

$$r(p,n)=\left(1-\sqrt{p-p\ln p+\frac{\ln n}{2n}}\right)_+^{-1} \tag{4.15}$$

m	Empirical Risk	Penalization factor	VC-bound
1	0.2241	2.4307	0.5447
2	0.2178	2.9623	0.6451
3	0.2150	3.5643	0.7662
4	0.1038	4.2677	0.4430
5	0.1011	5.1097	0.5163
6	**0.0650**	**6.1403**	**0.3994**
7	0.0628	7.4320	0.4670
8	0.0627	9.0943	0.5698
9	0.0543	11.3037	0.6139
10	0.0490	14.3639	0.7045

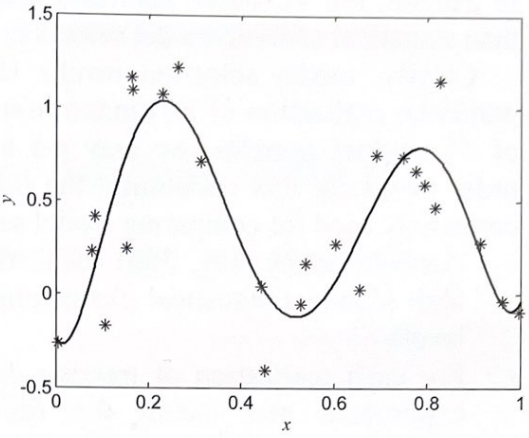


FIGURE 4.9 Polynomial model estimated from 25 samples using VC-bound. An optimal model is the 6-th degree polynomial.

The bound (4.14) can be immediately used for analytic model selection (if the VC-dimension is known). Next we show several examples of model selection for polynomial regression, where a set of models is specified as polynomials $f_m(x, \mathbf{w}) = w_0 + \sum_{i=1}^{m} w_i x^i$ and the goal is to choose the optimal polynomial degree m from available training data (x_i, y_i), $i = 1, ..., n$. We have already discussed this problem in Chapter 2 where resampling was used to choose the optimal polynomial degree. Let us use the same data set as in Example 2.7, but now we select optimal m using analytic VC bound (4.14) to estimate prediction risk. Note that the penalization factor depends on the VC-dimension. For polynomials of degree m, the VC-dimension equals the number of free parameters: $h = m + 1$. In the statistical literature, free parameters of a linear estimator are often denoted DoF (degrees-of-freedom). To apply the VC-bound, we perform polynomial fitting of the training data, calculate $R_{emp}(\omega)$ as MSE fitting error, and then multiply this fitting error by the VC penalization factor (4.15), in order to obtain an upper bound on risk. The best polynomial model corresponds to the smallest estimated bound on risk $R(\omega)$. Figure 4.9 shows the process of analytic model selection via (9.14), along with the final model. Note that the optimal model is the 6-th degree polynomial, which is different from the 5-th degree polynomial selected via 5-fold cross-validation in Fig. 2.14.

In general, the VC-based approach selects simpler (lower DoF) models than statistical analytic model selection methods.

Clearly, model selection results shown in Fig. 4.9 depend on a particular realization of 25 random training samples. For a different set of 25 random samples, we may get a different polynomial model. In order to handle this variability, the following experimental protocol is commonly used for comparing model selection methods:

- Generate many (say, 300) realizations of random training samples with identical statistical characteristics (i.e., sample size and noise level).
- For each realization of training data set, perform model selection experiment and record the fitting MSE and selected model complexity (in our case, polynomial degree). Also, for each selected model, estimate the prediction risk (MSE) using independent test set.
- Compile the values of fitting MSE, prediction MSE, and model complexity, obtained from 300 experiments, and show their empirical distributions using box plots. The standard box plot notation specifies marks at 95, 75, 50, 25, and 5th percentile of an empirical distribution. These box plots enable objective comparison between different model selection methods.

Of course, this protocol can be used *only* with synthetic data, as it uses multiple realizations of training data, and a large independent test set.

The next example shows model selection comparisons for polynomial regression when training data (30 samples) is pure noise. That is, the y-values of training data represent Gaussian noise with a standard deviation of one, and the x-values are uniformly distributed in a [0,1] interval. Model selection methods include several statistical methods (described in Section 2.4), VC-method (*vc*) and leave-one-out cross validation (*cv*). Comparison results in Fig. 4.10 show the box plots for the empirical distribution of the prediction RISK(MSE) for each model selection method. Note that the RISK (MSE) is presented in logarithmic scale. Relative performance of various model selection criteria can be evaluated by comparing their box plots. Box plots showing lower values of RISK correspond to better model selection. In particular, better model selection approaches select models that provide the lowest *guaranteed* prediction risk (i.e. with lowest Risk at the 95% mark), and also smallest variation of the risk (i.e., narrow box plots). Typically, model selection methods providing the *lowest guaranteed* prediction risk also yield the *lowest average* risk (i.e., lowest Risk at the 50% mark). Another performance index, DoF , shows the model complexity (degrees of freedom) chosen by a given method. The DoF box plot, in combination

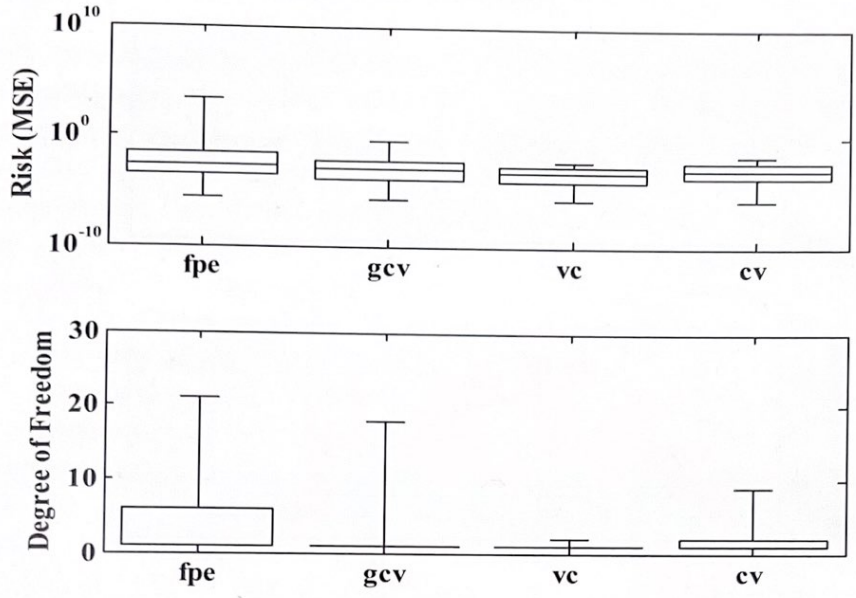
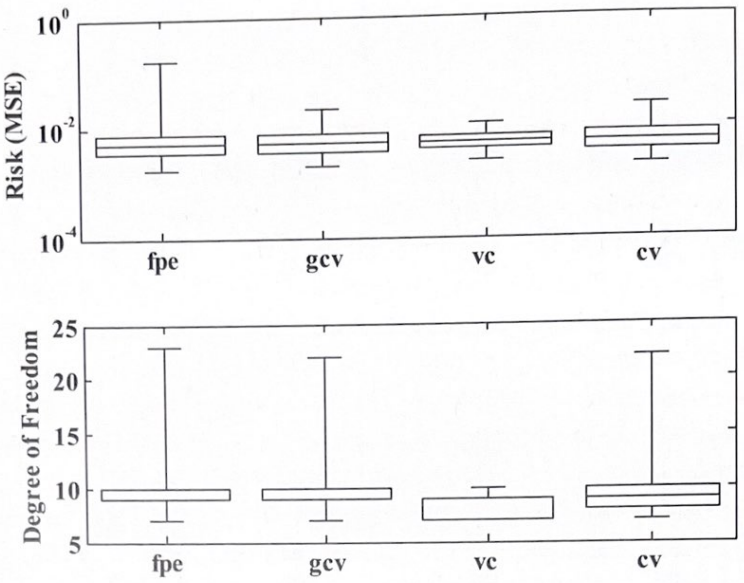


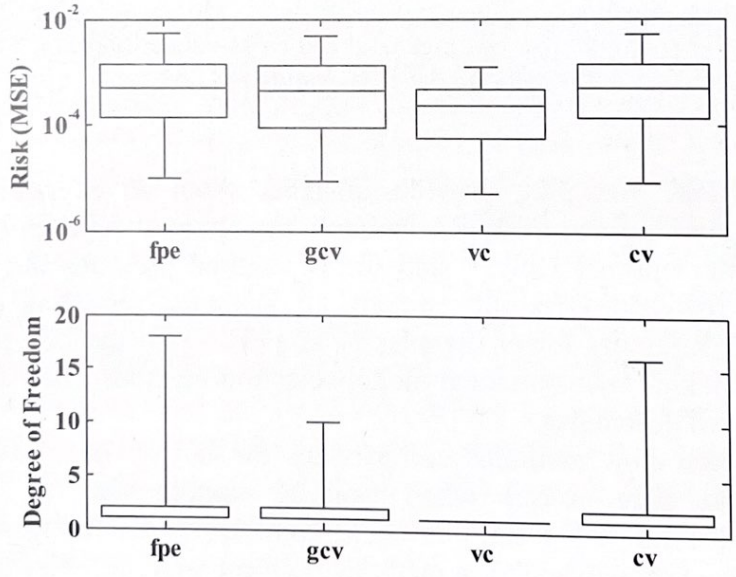
FIGURE 4.10 Model selection results for Gaussian noise, $\sigma = 1$, with sample size 30, using algebraic polynomial estimators. Model selection methods include: final prediction error (*fpe*), generalized cross-validation(*gcv*), VC-based model selection (*vc*) and leave-one-out cross-validation (*cv*).

with the RISK box plot, provides insights about an overfitting (or underfitting) of a given method, relative to the optimally chosen DoF. In this example, optimal DoF=1, and the *vc* method provides the lowest prediction risk and variability, among all methods (including *cv*), by consistently selecting lower complexity models. From the box plots of DoF we conclude that statistical model selection methods often detect a structure from pure noise.

Finally, we show methods' comparisons for the two data sets (sine-squared and pure noise) when training sample size is 'large'. Comparison results using 100 training samples are shown in Fig. 4.11. In this case, most model selection methods perform well, i.e., they provide small prediction risk (MSE). However, the VC-based model selection still shows the lowest variability with respect to random realizations of the training data.



(a) Comparisons for the sine-squared target function, standard deviation of additive noise $\sigma = 0.2$.



(b) Comparisons for pure Gaussian noise, $\sigma = 1$.

FIGURE 4.11 Model selection comparisons for the sine-squared target function and pure Gaussian noise with sample size 100, using algebraic polynomial estimators.