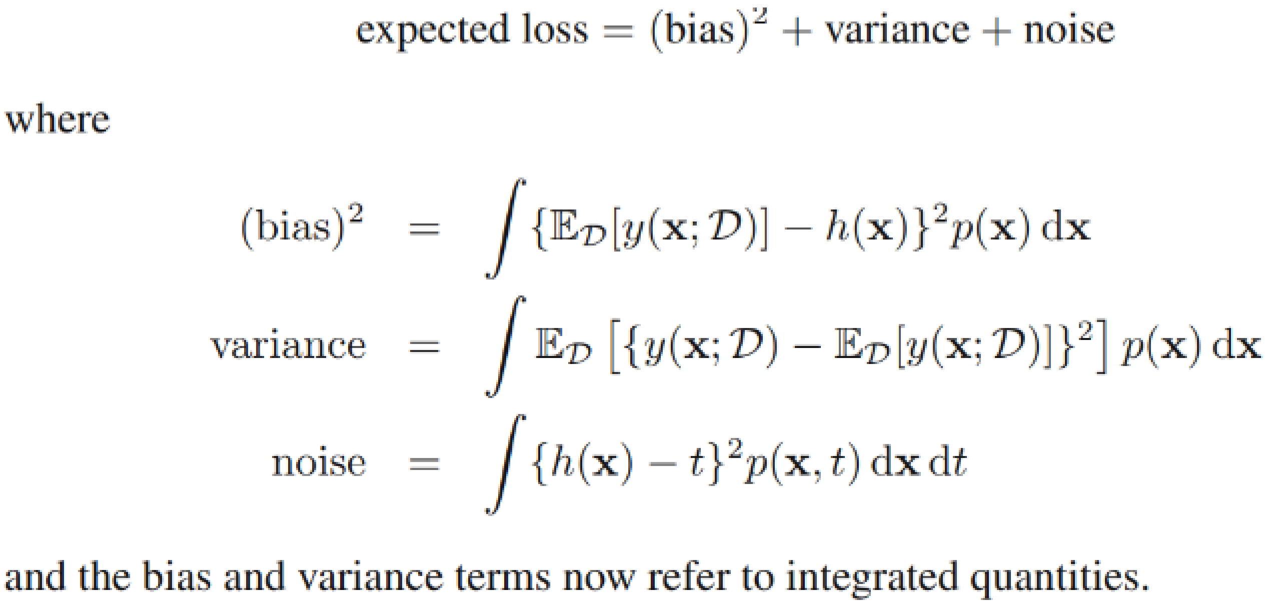
03.Model Evaluation,Selection and Ensembles

Bias-Variance Tradeoff

the use of maximum likelihood, or equivalently least squares, can lead to severe over-fitting if complex models are trained using data sets of limited size. However, limiting the number of basis functions in order to avoid over-fitting has the side effect of limiting the flexibility of the model to capture interesting and important trends in the data. Introduction of regularization terms can control over-fitting for models with many parameters but how can we determine a suitable value for the regularization coefficient λ?
Consider a frequentist viewpoint of the model complexity issue, known as the bias□variance trade-off. Given the conditional distribution $p(t|\mathbf{x})$ a popular choice for the loss function is the squared loss function for which the optimal prediction is given by the conditional expectation denoted by $h(\mathbf{x}) = E[t|\mathbf{x}] = \int tp(t|\mathbf{x})dt$. The squared loss can be written as $E[L] = \int\{y(\mathbf{x}) - h(\mathbf{x})\}^2 p(\mathbf{x})dx + \int\{h(\mathbf{x} - t\}^2 p(\mathbf{x}, t)d\mathbf{x}dt$ (The second term represent the minimum achievable value of the expected loss and the first term depends on the choice of y(x) and we will seek a solution to make it minimum. Being a non negative term the smallest we can achieve is zero).
We model h(x) as a parametric function $y(\mathbf{x}, \mathbf{w})$ governed by a parameter vector $\mathbf{w}$ and, using a frequentist treatment, we will make a point estimate of $\mathbf{w}$ based on the dataset $\mathcal{D}$ and tries instead to interpret the uncertainty of this estimate through the following thought experi□ment. Suppose to have a large number of data sets each of size N and each drawn indipendently from the distribution $p(t, \mathbf{x})$. For each data set $\mathcal{D}$ we can run our learning algorithm and obtain a prediction function $y(\mathbf{x}; \mathcal{D})$. Different data sets will give different functions and consequently different values of the squared loss. The performance of a particular learning algorithm is then assessed by taking the average over this ensemble of data sets. For a paritcular data set $\mathcal{D}$ $E[L] = \{y(\mathbf{x}; \mathcal{D}) - h(\mathbf{x})\}^2$, as it depends on the dataset we will take its average over the ensemble of data sets.

$$\begin{aligned}&\mathbb{E}_{\mathcal{D}}\left[\{y(\mathbf{x};\mathcal{D}) - h(\mathbf{x})\}^2\right] \\ &\quad = \underbrace{\{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})] - h(\mathbf{x})\}^2}_{(\text{bias})^2} + \underbrace{\mathbb{E}_{\mathcal{D}}\left[\{y(\mathbf{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})]\}^2\right]}_{\text{variance}}.\end{aligned}$$

The first term, called the squared bias, represents the extent to which the average prediction over all data sets differs from the desired regression function. The second term, called the variance, measures the extent to which the solutions for individual data sets vary around their average, and hence this measures the extent to which the function $y(\mathbf{x}; \mathcal{D})$ is sensitive to the particular choice of data set.



$$\text{expected loss} = (\text{bias})^2 + \text{variance} + \text{noise}$$

where

$$(\text{bias})^2 = \int \{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})] - h(\mathbf{x})\}^2 p(\mathbf{x})\,\mathrm{d}\mathbf{x}$$

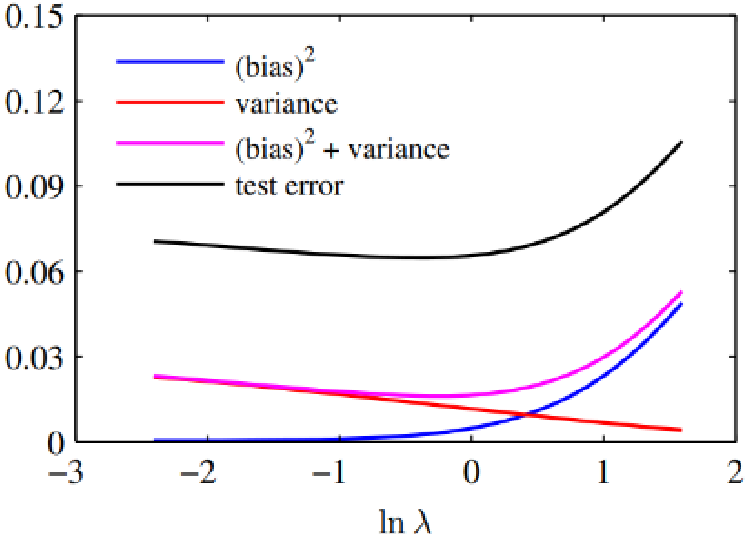
$$\text{variance} = \int \mathbb{E}_{\mathcal{D}}\left[\{y(\mathbf{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})]\}^2\right] p(\mathbf{x})\,\mathrm{d}\mathbf{x}$$

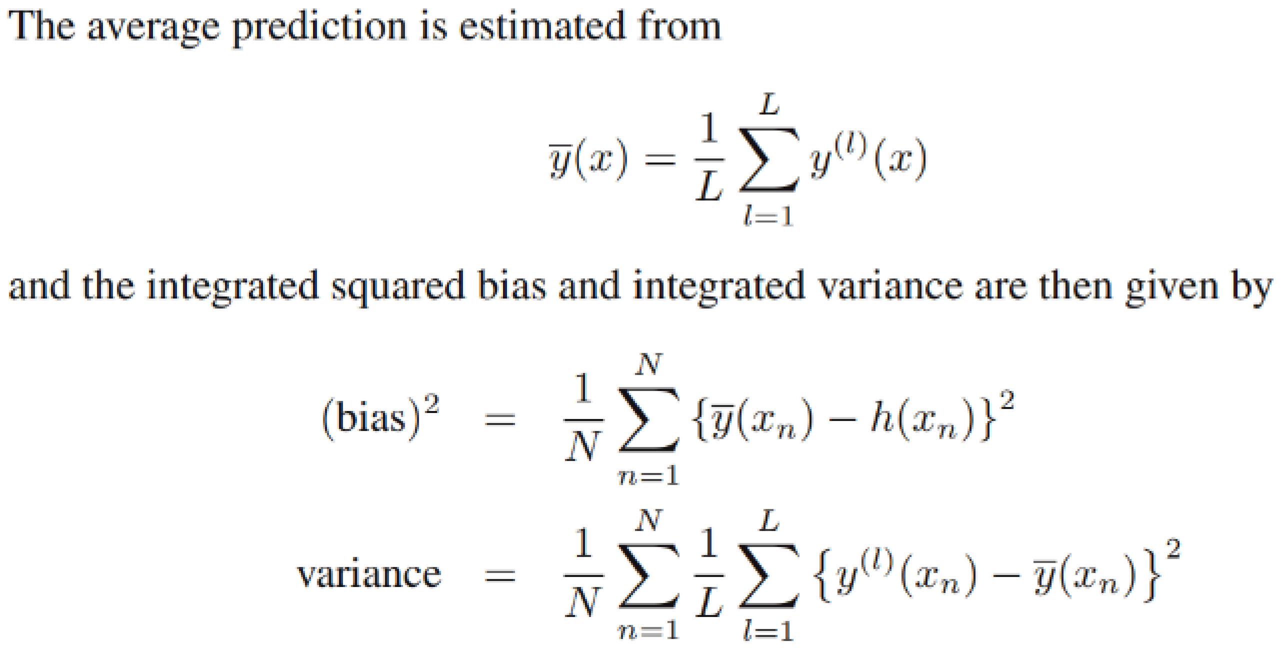
$$\text{noise} = \int \{h(\mathbf{x}) - t\}^2 p(\mathbf{x},t)\,\mathrm{d}\mathbf{x}\,\mathrm{d}t$$

and the bias and variance terms now refer to integrated quantities.

Our goal is to minimize the expected loss, which we have decomposed into the sum of a (squared) bias, a variance, and a constant noise term. There is a trade-off between bias and variance, with very flexible models having low bias and high variance, and relatively rigid models having high bias and low variance. The model with the optimal predictive capability is the one that leads to the best balance between bias and variance.

Plot of squared bias and variance, together with their sum, corresponding to the results shown in Figure 3.5. Also shown is the average test set error for a test data set size of 1000 points. The minimum value of $(\text{bias})^2 + \text{variance}$ occurs around $\ln\lambda = -0.31$, which is close to the value that gives the minimum error on the test data.





The average prediction is estimated from

$$\overline{y}(x) = \frac{1}{L}\sum_{l=1}^{L} y^{(l)}(x)$$

and the integrated squared bias and integrated variance are then given by

$$\begin{aligned}(\text{bias})^2 &= \frac{1}{N}\sum_{n=1}^{N}\left\{\overline{y}(x_n) - h(x_n)\right\}^2 \\ \text{variance} &= \frac{1}{N}\sum_{n=1}^{N}\frac{1}{L}\sum_{l=1}^{L}\left\{y^{(l)}(x_n) - \overline{y}(x_n)\right\}^2\end{aligned}$$

where the integral over x weighted by the distribution p(x) is approximated by a finite sum over data points drawn from that distribution.

We see that small values of λ allow the model to become finely tuned to the noise on each individual data set leading to large variance. Conversely, a large value of λ pulls the weight parameters towards zero leading to large bias.

If we had a large number of independent training sets of a given size, we would be better off combining them into a single large training set, which of course would reduce the level of over-fitting for a given model complexity.