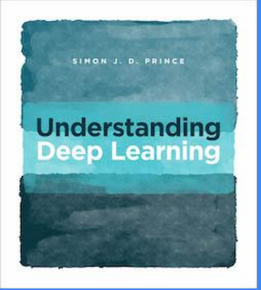
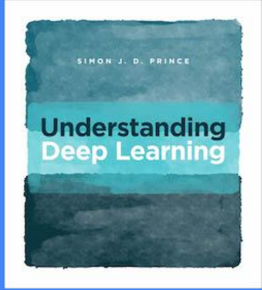
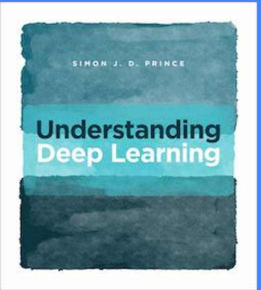
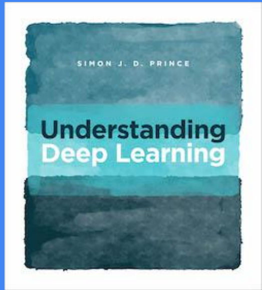




Measuring performance





Noise

"Noise may arise because there is a genuine stochastic element to the data generation process, because some of the data are mislabeled, or because there are further explanatory variables that were not observed."

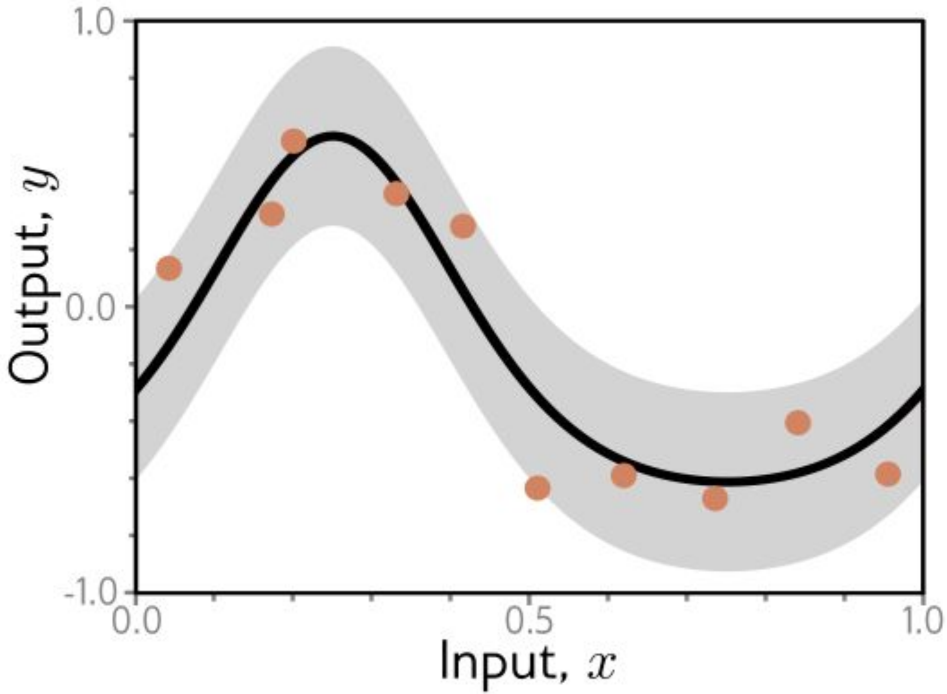
Bias

"A second potential source of error may occur because the model is not flexible enough to fit the true function perfectly."

Variance

"We have limited training examples, and there is no way to distinguish systematic changes in the underlying function from noise in the underlying data. When we fit a model, we do not get the closest possible approximation to the true underlying function. Indeed, for different training datasets, the result will be slightly different each time."

Figure 8.3 Regression function. Solid black line shows ground truth function. To generate I training examples $\{x_i, y_i\}$, the input space $x \in [0, 1]$ is divided into I equal segments and one sample x_i is drawn from a uniform distribution within each segment. The corresponding value y_i is created by evaluating the function at x_i and adding Gaussian noise (gray region shows ± 2 standard deviations). The test data are generated in the same way.



Let's simulate some data

Mathematical formulation of test error

we can observe different outputs y for the same input x, so for each x, there is a distribution P r(y|x) with expected value (mean) μ[x]

$$\mu[x] = \mathbb{E}_y[y[x]] = \int y[x] Pr(y|x) dy$$

$$\sigma^2 = \mathbb{E}_y\left[(\mu[x] - y[x])^2\right]$$

$$
\begin{aligned}
L[x] &= \left(\mathrm{f}[x,\boldsymbol{\phi}] - y[x]\right)^2 \\
&= \Big(\left(\mathrm{f}[x,\boldsymbol{\phi}] - \mu[x]\right) + \left(\mu[x] - y[x]\right)\Big)^2 \\
&= \left(\mathrm{f}[x,\boldsymbol{\phi}] - \mu[x]\right)^2 + 2\left(\mathrm{f}[x,\boldsymbol{\phi}]) - \mu[x]\right)\left(\mu[x] - y[x]\right) + \left(\mu[x] - y[x]\right)^2
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{E}_y\left[L[x]\right] &= \mathbb{E}_y\Big[\left(\mathrm{f}[x,\boldsymbol{\phi}]-\mu[x]\right)^2 + 2\left(\mathrm{f}[x,\boldsymbol{\phi}]-\mu[x]\right)\left(\mu[x]-y[x]\right) + \left(\mu[x]-y[x]\right)^2\Big] \\
&= \left(\mathrm{f}[x,\boldsymbol{\phi}]-\mu[x]\right)^2 + 2\left(\mathrm{f}[x,\boldsymbol{\phi}] - \mu[x]\right)\left(\mu[x]-\mathbb{E}_y\left[y[x]\right]\right) + \mathbb{E}_y\left[\left(\mu[x]-y[x]\right)^2\right] \\
&= \left(\mathrm{f}[x,\boldsymbol{\phi}]-\mu[x]\right)^2 + 2\left(\mathrm{f}[x,\boldsymbol{\phi}]-\mu[x]\right)\cdot 0 + \mathbb{E}_y\left[\left(\mu[x]-y[x]\right)^2\right] \\
&= \left(\mathrm{f}[x,\boldsymbol{\phi}] - \mu[x]\right)^2 + \sigma^2,
\end{aligned}
\tag{8.3}
$$

Expected loss = squared **deviation between the model and the true function mean**, and the second term is the **noise**

The expected model output $\mathrm{f}_\mu[x]$ with respect to all possible datasets $\mathcal{D}$ is hence:

$$\mathrm{f}_\mu[x] = \mathbb{E}_{\mathcal{D}}\Big[\mathrm{f}\big[x, \boldsymbol{\phi}[\mathcal{D}]\big]\Big]. \tag{8.4}$$

$$\begin{aligned}
&\big(\mathrm{f}[x, \boldsymbol{\phi}[\mathcal{D}]] - \mu[x]\big)^2 \\
&\quad= \Big(\big(\mathrm{f}[x, \boldsymbol{\phi}[\mathcal{D}]] - \mathrm{f}_\mu[x]\big) + \big(\mathrm{f}_\mu[x] - \mu[x]\big)\Big)^2 \\
&\quad= \big(\mathrm{f}[x, \boldsymbol{\phi}[\mathcal{D}]] - \mathrm{f}_\mu[x]\big)^2 + 2\big(\mathrm{f}[x, \boldsymbol{\phi}[\mathcal{D}]] - \mathrm{f}_\mu[x]\big)\big(\mathrm{f}_\mu[x] - \mu[x]\big) + \big(\mathrm{f}_\mu[x] - \mu[x]\big)^2
\end{aligned}$$

$$\mathbb{E}_{\mathcal{D}}\Big[\big(\mathrm{f}[x, \boldsymbol{\phi}[\mathcal{D}]] - \mu[x]\big)^2\Big] = \mathbb{E}_{\mathcal{D}}\Big[\big(\mathrm{f}[x, \boldsymbol{\phi}[\mathcal{D}]] - \mathrm{f}_\mu[x]\big)^2\Big] + \big(\mathrm{f}_\mu[x] - \mu[x]\big)^2$$

$$\mathbb{E}_{\mathcal{D}}\Big[\mathbb{E}_y[L[x]]\Big] = \underbrace{\mathbb{E}_{\mathcal{D}}\Big[\big(\mathrm{f}[x,\boldsymbol{\phi}[\mathcal{D}]] - \mathrm{f}_{\mu}[x]\big)^2\Big]}_{\text{variance}} + \underbrace{\big(\mathrm{f}_{\mu}[x] - \mu[x]\big)^2}_{\text{bias}} + \underbrace{\sigma^2}_{\text{noise}}. \tag{8.7}$$

Bias-variance Trade-off

Double Descent