- to see how well any of these smoothing methods work, need a notion of 'long-run' performance
- e.g. if we assume $Y = f(X) + \epsilon$ and our method gives $\widehat{f}(\cdot)$ based on $(x_1, y_n), \ldots, (x_N, y_N)$:
  - Does $\widehat{f}(x_0) \to f(x_0), \quad N \to \infty$? all x_0?
  - Is $\sqrt{n}\{\widehat{f}(x_0) - f(x_0)\}$ asymptotically normal? variance?
  - Is $E\widehat{f}(x_0) = f(x_0)$? (unbiased?)
- assume we have a a loss function, i.e. a measure of distance from Y to $\widehat{f}(X)$

$$L(Y, \widehat{f}(X)) = (Y - \widehat{f}(X))^2$$

- **Test error, generalization error**:

$$\mathrm{Err} = E[L\{Y, \widehat{f}(X)\}]$$

over the distribution of Y, X, and $\widehat{f}$.

- to see how well any of these smoothing methods work, need a notion of 'long-run' performance
- e.g. if we assume $Y = f(X) + \epsilon$ and our method gives $\widehat{f}(\cdot)$ based on $(x_1, y_n), \dots, (x_N, y_N)$:
  - Does $\widehat{f}(x_0) \to f(x_0)$, $\quad N \to \infty$? all x_0?
  - Is $\sqrt{n}\{\widehat{f}(x_0) - f(x_0)\}$ asymptotically normal? variance?
  - Is $E\widehat{f}(x_0) = f(x_0)$? (unbiased?)
- assume we have a a loss function, i.e. a measure of distance from Y to $\widehat{f}(X)$

$$L(Y, \widehat{f}(X)) = (Y - \widehat{f}(X))^2$$

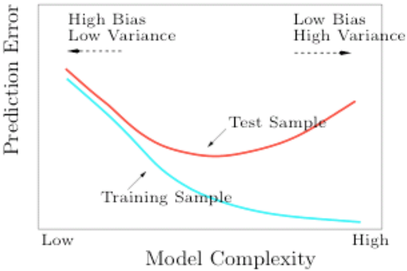
- **Test error, generalization error**:

$$\mathrm{Err} \ = E[L\{Y, \widehat{f}(X)\}]$$

over the distribution of Y, X, and $\widehat{f}$.

- **Test error**: $\mathrm{Err} \ = EL\{Y, \widehat{f}(X)\}$:
  $\widehat{f}(X) = \widehat{f}(X; x_1, y_1, \dots, x_N, y_N) = \widehat{f}(X, t_N)$, say
- distribution of $\widehat{f}(X)$ depends on distribution of X and t_N
- $\mathrm{Err} \ = E_{X,Y,t_N} L\{Y, \widehat{f}(X)\}$

- **Training error**: average loss in training sample
  $\overline{\text{err}} = \frac{1}{N}\sum_{i=1}^{N} L\{y_i, \widehat{f}(x_i)\}$
- As $\widehat{f}(\cdot)$ becomes more complex, training error will decrease (eventually to 0) but test error will increase, because $\widehat{f}(\cdot)$ fits observed data exactly



- Test error at a fixed x_0: $\mathrm{Err}\,(x_0) = E[L\{Y, \widehat{f}(x_0)\}]$
- depends on distribution of Y and $\widehat{f}(x_0) = \widehat{f}(x_0; t_N)$
- under squared error loss

$$\mathrm{Err}\,(x_0) = \sigma^2 + \mathrm{Bias}^2 \widehat{f}(x_0) + \mathrm{var}\,\widehat{f}(x_0)$$

- Example: k-nearest neighbour estimate $\widehat{f}(x_0) = \frac{1}{k}\sum_{i=1}^{N} y_i 1\{x_i \in N_k(x_0)\}$
- $E\widehat{f}(x_0) = \frac{1}{k}\sum E[y_i 1\{x_i \in N_k(x_0)\}]$
- assume x_i are fixed $= \frac{1}{k}\sum_{i=1}^{N} E(y_i) 1\{x_i \in N_k(x_0)\} = \frac{1}{k}\sum_{\ell=1}^{k} f(x_{(\ell)})$
- $\mathrm{var}\,\widehat{f}(x_0) = \sigma^2/k$
- see Eq. (7.9) ; note have assumed training x_i are fixed

- Example: linear regression: $\widehat{f}(x_0) = x_0^T \hat{\beta}$
- $\mathrm{var}\widehat{f}(x_0) = \mathrm{var}(x_0^T \hat{\beta}) = \mathrm{var} x_0^T \{(X^T X)^{-1} X^T y\}$ where X and y refer to training data
- $= \mathrm{var}(a^T y)$, say,
  $= \sigma^2 a^T a = \sigma^2 ||a||^2 = \sigma^2 ||x_0 (X^T X)^{-1} X^T x_0||^2 = \sigma^2 ||h(x_0)||^2$
- note have assumed training x_i are fixed
- $\mathrm{Err}(x_0) = \sigma^2 + \mathrm{Bias}^2 \widehat{f}(x_0) + \sigma^2 ||h(x_0)||^2$
- a rough guide to $\mathrm{Err}(x_0)$ is
  $\frac{1}{N} \sum \mathrm{Err}(x_i) = \sigma^2 + \frac{1}{N} \sum \{f(x_i) - E\widehat{f}(x_i)\}^2 + \sigma^2 p/N$
- shows that Err increases as p increases
- similarly for ridge regression
  $\mathrm{Err}(x_0) = \sigma^2 + \mathrm{Bias}^2 \widehat{f}^{ridge}(x_0) + \sigma^2 ||h^{ridge}(x_0)||^2$
- $h^{ridge}(x_0) = x_0^T (X^T X + \lambda I)^{-1} x_0$

- §7.3 and 7.4 discuss instead the estimation of "in-sample error", not quite the same as test error
- $\mathrm{Err}_{in} = \frac{1}{N}\sum_{i=1}^{N} E_y E_{Y^{new}}[L\{Y_i^{new}, \widehat{f}(x_i)\}]$
- test values Y_i^{new} observed at training points x_i
- Claim $\mathrm{Err}_{in} = E_y\overline{err} + (2/N)\sum_{i=1}^{N}\mathrm{cov}(\hat{y}_i, y_i)$ (7.18)
- For squared error loss, a vague sketch

$$
\begin{aligned}
\overline{err} &= \frac{1}{N}\sum(y_i - \hat{y}_i)^2 \\
&= \frac{1}{N}\sum(y_i - f(x_i) + f(x_i) - \hat{y}_i)^2 \\
&= \frac{1}{N}\sum(y_i - f(x_i))^2 + \frac{1}{N}\sum\{\hat{y}_i - f(x_i)\}^2 \\
&\qquad -\frac{2}{N}\sum\{y_i - f(x_i)\}\{\hat{y}_i - f(x_i)\}
\end{aligned}
$$

- If $\hat{y} = Sy$, where S has d degrees of freedom, then $\sum \mathrm{cov}(\hat{y}_i, y_i) = d\sigma^2$
- $\mathrm{Err}_{in} = E_y \overline{err} + \frac{2}{N} d\sigma^2$
- Err_{in} relevant for model selection, and easier to analyse than Err
- Estimating Err_{in}: for example $\overline{err} + \frac{2}{N} d\sigma^2$: this is C_p
- AIC replaces $\frac{1}{N}\sum(y_i - \hat{y}_i)^2$ with $-\frac{2}{N}\log(\hat{\theta}; y)$
- see Figure 7.4

- Generalization/test error $\mathrm{Err} = E_{X,Y,\hat{f}}[L\{Y, \widehat{f}(X)\}]$
- Cross-validation attempts to estimate this directly
- $CV = \frac{1}{N}\sum L\{y_i, \widehat{f}^{\kappa(i)}(x_i)\}$ (7.42)
- $\kappa(i)$ indexes which of K partitions observation i is in (K-fold CV)
- If $\widehat{f}$ depends on a tuning parameter, α, then we compute
- $CV(\alpha) = \frac{1}{N}\sum L\{y_i, \widehat{f}^{\kappa(i)}(x_i, \alpha)\}$ for a variety of choices
- $K = 1$ has low bias but high variance; large K the opposite; $K = 5$ or 10 recommended
- *generalized* CV is an approximation to CV with $K = 1$ used in linear fitting methods with squared error loss
- (GCV is used by the lm.ridge program)