

18

$$R_{tr}(\beta) = \frac{1}{N} \sum_{i=1}^N (y_i - \beta^T x_i)^2$$

$$R_{te}(\beta) = \frac{1}{M} \sum_{i=1}^M (\tilde{y}_i - \beta^T \tilde{x}_i)^2$$

a) $E[R_{te}(\hat{\beta})] = \frac{1}{M} \sum_{i=1}^M E[(\tilde{y}_i - \hat{\beta}^T \tilde{x}_i)^2] = E[(\tilde{y}_i - \hat{\beta}^T \tilde{x}_i)^2]$

← cancel each other out.

It does not matter if we have M or N observations as we are taking the expectation over on average, then M (or N) cancels out.

b) $\hat{\beta}_{te} = (X^T X)^{-1} X^T y$ and it is the solution of

argmin _{β} $\frac{1}{N} \sum_{i=1}^N (\tilde{y}_i - \beta^T \tilde{x}_i)^2$, so for any other $\hat{\beta}$,

eg. LS estimator on the training data

we will have

$$\frac{1}{N} \sum_{i=1}^N (\tilde{y}_i - \hat{\beta}_{te}^T \tilde{x}_i)^2 \leq \frac{1}{N} \sum_{i=1}^N (\tilde{y}_i - \hat{\beta}^T \tilde{x}_i)^2$$

c) $E[R_{tr}(\hat{\beta})] = E\left[\frac{1}{N} \sum_{i=1}^N (y_i - \hat{\beta}^T x_i)^2\right] = E\left[\frac{1}{N} \sum_{i=1}^N (\tilde{y}_i - \hat{\beta}_{te}^T \tilde{x}_i)^2\right] = E[R_{te}(\hat{\beta}_{te})]$

as training and test data are drawn independently from the same data distribution.

d) $E[R_{tr}(\hat{\beta})] = E\left[\frac{1}{N} \sum_{i=1}^N (\tilde{y}_i - \hat{\beta}_{te}^T \tilde{x}_i)^2\right] \leq E\left[\frac{1}{N} \sum_{i=1}^N (\tilde{y}_i - \hat{\beta}^T \tilde{x}_i)^2\right]$

$$= E[R_{te}(\hat{\beta})]$$

Interpretation: the expected training error is lower than the expected test error. Reasonable.