23 Generalization theory / Overfitting

Generalization

How well does a learned function predict on future test examples?

How to choose hypothesis space H?

If H is too complex

- over-fitting
- small training error
- large test error
- very different functions have similar training error
- perturbing training data slightly yields very different optimal hypotheses

If H is too restricted

- under-fitting
- large training error
- large test error

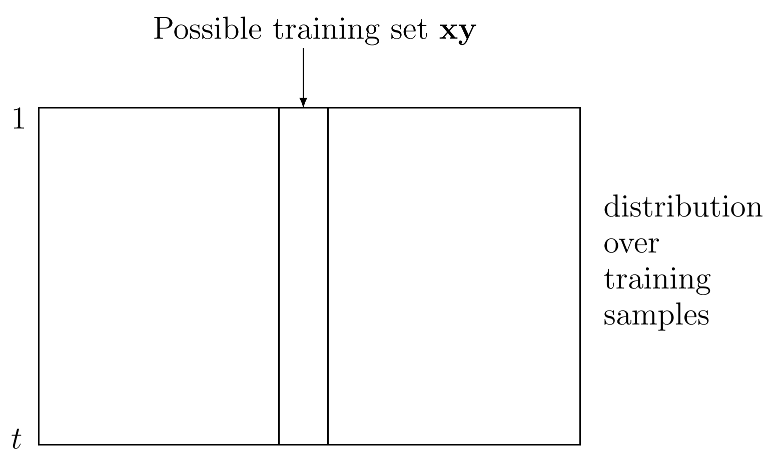
23.1 Introduction to statistical generalization theory

Mathematical model

independent identically distributed (IID) random examples

- Assume a fixed joint distribution P_{XY} over $X \times Y$
- Training examples $(x_1, y_1), \ldots, (x_t, y_t)$ independently drawn from P_{XY}
- Test examples independently drawn from same P_{XY}

Learner maps $(x_1, y_1) \ldots (x_t, y_t)$ to a hypothesis $h : X \to Y$



$\Downarrow$

$\boxed{\text{Learner}}$

$\Downarrow$

$h_{\mathbf{xy}}$ distribution over hypotheses

$\Downarrow$

$E_{xy}\,\mathrm{err}(h_{\mathbf{xy}}(x), y)$ expected test error

$\hat{E}_{x_i y_i}\mathrm{err}(h_{\mathbf{xy}}(x_i), y_i)$ training error

For squared prediction error

$$\mathrm{err}(\hat{y}, y) \;=\; (\hat{y} - y)^2$$

get

$$\begin{array}{lll}
E_{\mathbf{xy}} E_{xy} \left(h_{\mathbf{xy}}(x) - y\right)^2 & & \text{test error} \\[1ex]
\quad = & E_{\mathbf{xy}} \hat{E}_{x_i y_i} \left(h_{\mathbf{xy}}(x_i) - y_i\right)^2 & \text{train error} \\
\quad + & E_{\mathbf{xy}} \hat{E}_{x_i} \left(h_{\mathbf{xy}}(x_i) - h^*(x_i)\right)^2 & \text{train variance} \\
\quad + & E_{\mathbf{xy}} E_{x} \left(h_{\mathbf{xy}}(x) - h^*(x)\right)^2 & \text{variance}
\end{array}
\left.\begin{array}{l} \\ \left.\begin{array}{l}\\ \\ \end{array}\right\} \begin{array}{l}\text{opt test}\\ \text{err in } H\end{array} \\ \\ \end{array}\right\} \begin{array}{l}\text{hypothesis}\\ \text{test err}\end{array}$$

where

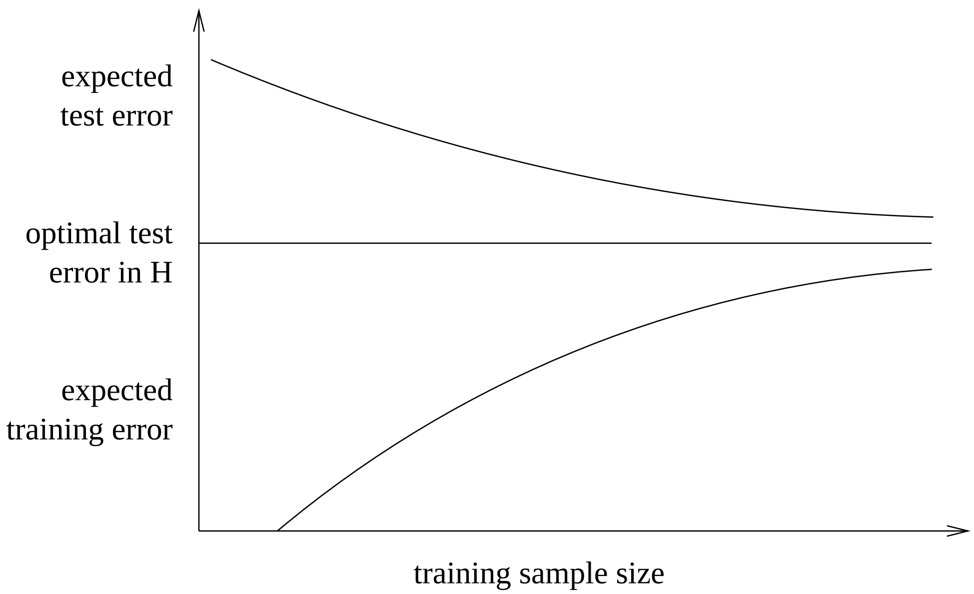
$$h^* \;=\; \arg\min_{h \in H} \; E_{xy} \left(h(x) - y\right)^2$$

$$H \quad \text{is} \quad \text{a closed linear space}$$

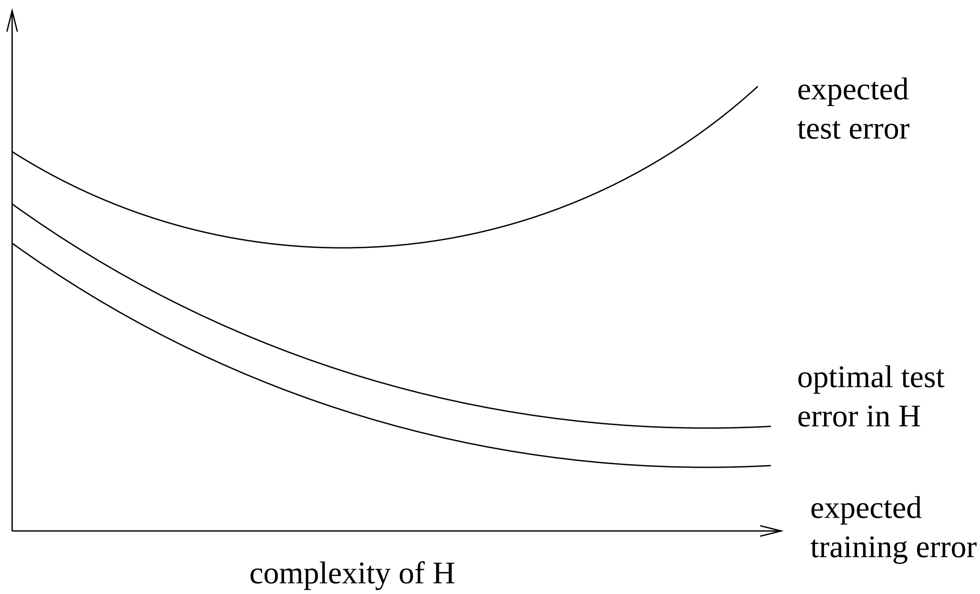
Immediate consequence

$$\begin{array}{lclcl}
\text{expected} & & \text{optimal} & & \text{expected} \\
\text{hypothesis} & \geq & \text{test} & \geq & \text{hypothesis} \\
\text{test} & & \text{error} & & \text{train} \\
\text{error} & & \text{in } H & & \text{error}
\end{array}$$

23.2 Learning curves

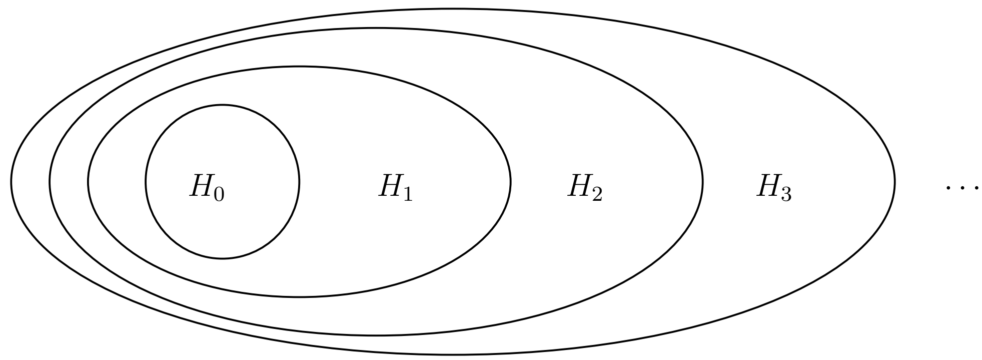


23.3 Overfitting curves



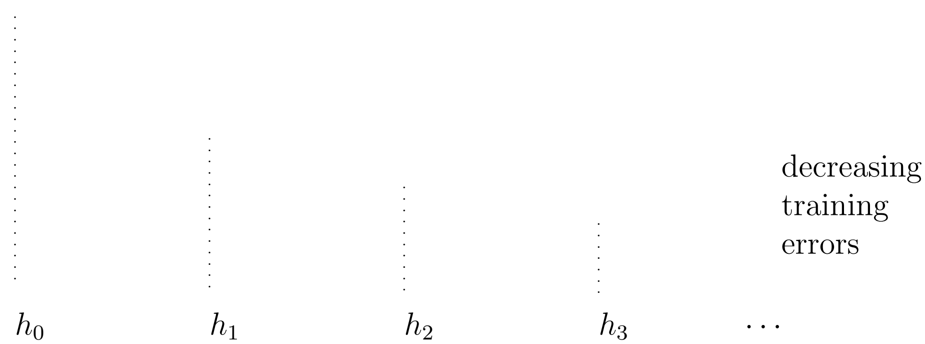
23.4 Automatic complexity control

Model selection



How to choose the right complexity level?

Given data, get



which hypothesis to choose?

- choose too early: under-fit
- choose too late: over-fit

strategy 1: complexity penalization

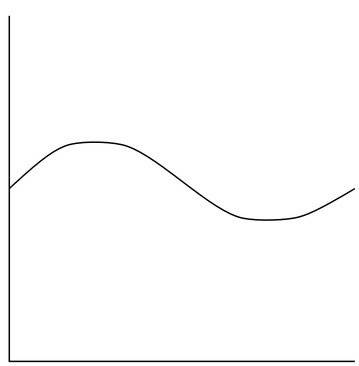
- guess at variances
- training errors say nothing about variances
- penalty(i) approximates variance at complexity level i
- minimize: training_error(i) + penalty(i)

Strategy 2: Hold out testing

- Split training data into pseudo-train and pseudo-test set
- Train on pseudo-train and test each hypothesis $h_0, h_1, \ldots$ on the held-out pseudo-test
- Hold-out test gives an *unbiased* estimate of test error
- Pick i with best hold-out test
- Re-train at complexity level i on all the data

Strategy 3: Metric space

$\mathrm{P}_{Y|X}$



P_X

$f : X \to Y$

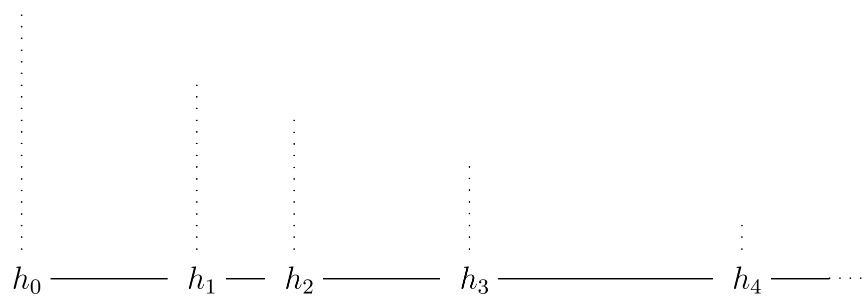
Assume we know P_X (which can be estimated from unlabeled data $x_1, x_2, \ldots$)

Defines a *metric* on H

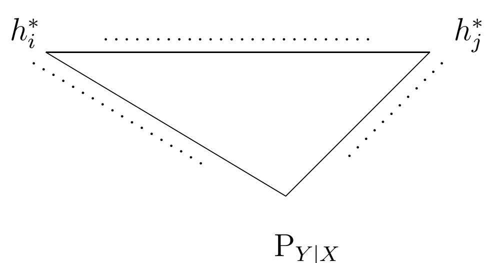
$$
\begin{aligned}
d(h,g) &= \sqrt{\int_x \left(h(x)-g(x)\right)^2 \; d\mathrm{P}_X} \\
d(h,\mathrm{P}_{Y|X}) &= \sqrt{\int_x \int_y \left(h(x)-y\right)^2 \; d\mathrm{P}_{Y|x} d\mathrm{P}_X}
\end{aligned}
$$

Goal is to minimize $d(h,\mathrm{P}_{Y|X})$

Given data, get



Have 2 metrics, real and estimated



Adjust $\hat{d}(h_i, P_{Y|X})$ by multiplying it by $\displaystyle \max_{j<i} \frac{d(h_i,h_j)}{\hat{d}(h_i,h_j)}$

Readings

Hastie, Tibshirani, Friedman: Sections 2.9, 5.1–5.5

Schuurmans, D. and Southey, F. (2001) Metric-based methods for adaptive model selection and regularization. *Machine Learning*, 48(1-3): 51–84.