Topics in Learning Theory

Lecture 2: Generalization Error

General Background

- Basic prediction problem:
  - known input X.
  - unknown output Y.
  - prediction function (classifier) f: $Y \approx f(X)$.
- Supervised learning:
  - learn f from training data $S_n = (X_i, Y_i)_{i=1,\ldots,n}$.
  - quality of prediction: measured by loss function $\phi(f(x), y)$.

Regression

- Predict real value $\boldsymbol{y} \in \boldsymbol{R}$
- Real-valued prediction rule $f(x)$.
- Squared error loss: $\phi(f(x), y) = (f(x) - y)^2$.

Binary Classification

- Predict binary label $y \in \{\pm 1\}$.
- Classifier $f(x)$:
  - binary valued: $f(x) \in \{\pm 1\}$
  - real valued $f(x)$, with decision rule: $\begin{cases} y = 1 & \text{if } f(x) > 0 \\ y = -1 & \text{if } f(x) \leq 0 \end{cases}$
- Classification error loss: $\phi(f(x), y) = I(f(X)Y \leq 0)$.
  - I : indicator function.

Training error and generalization error

- Prediction function $f(x)$.
- Loss function $\phi(f(x), y)$.
- Training error: $\hat{R}(f) = \frac{1}{n}\sum_{i=1}^{n} \phi(f(X_i), Y_i)$.
  - What we can observe.
- Generalization error (test error): $R(f) = \mathbf{E}_{X,Y}\phi(f(X), Y)$.
  - Prediction performance over unseen data: what we are interested in.

Learning Algorithm

- Learning algorithm $\mathcal{A}$
  - learn prediction rule $\hat{f} = \mathcal{A}(S_n)$ from training data $S_n = \{(X_i, Y_i)\}_{i=1,\ldots,n}$.
- Empirical risk minimization learner:

$$\hat{\boldsymbol{f}} = \arg\min_{\boldsymbol{f} \in \boldsymbol{\mathcal{H}}} \sum_{\boldsymbol{i}=1}^{\boldsymbol{n}} \boldsymbol{\phi}(\boldsymbol{f}(\boldsymbol{X}_{\boldsymbol{i}}), \boldsymbol{Y}_{\boldsymbol{i}}),$$

  $\mathcal{H}$: set of candidate prediction rules (e.g. linear combination of features).

- Complexity of learning algorithm $\mathcal{A}$
  - measured by how large and diversified the candidate rule set $\mathcal{H}$ is.

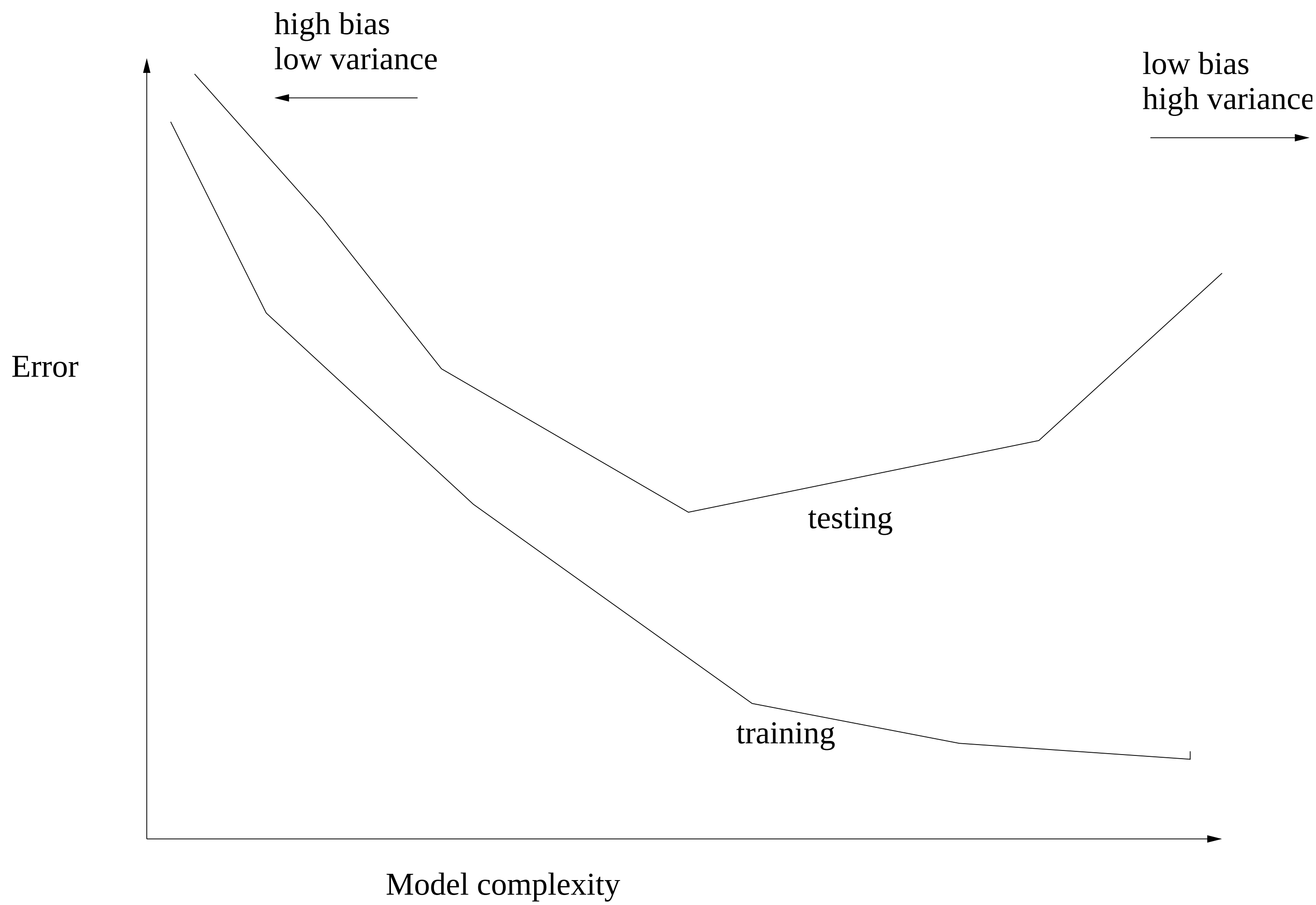
Overfitting and Model Complexity

- Over-fitting of data:
  - $f(x) = y_i$ if $x = x_i$ at a training point, and $f(x) = 0$ otherwise.
  - $f(x)$ that perfectly explains the data is not necessarily a good predictor.

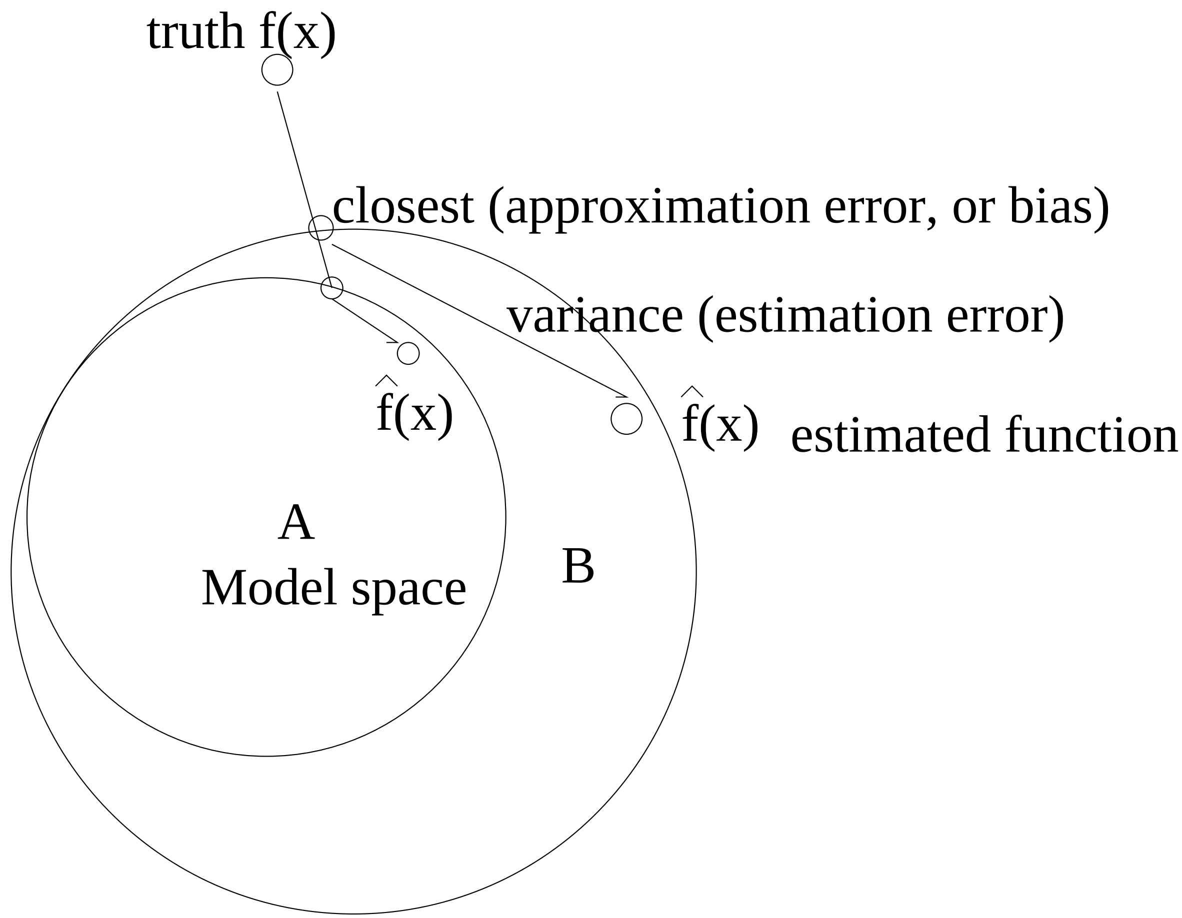
- Predictive ability:
  - fit well on the training data (small bias).
  - training performance resembles test performance (small variance).

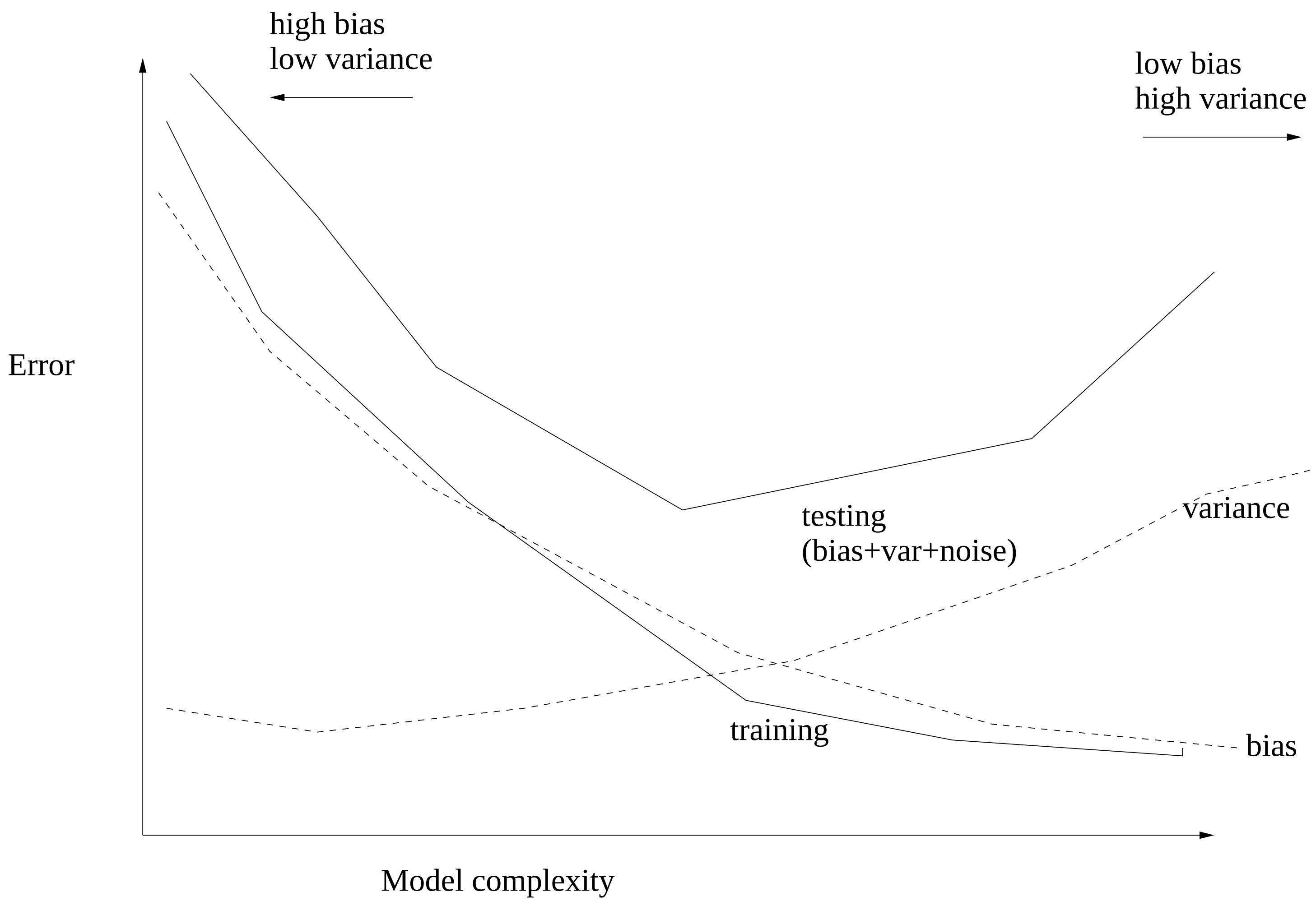
- Trade-off: expressively powerful model $\rightarrow$ poor generalization.

- Regularization: restrict the model expressiveness or statistical complexity.


high bias
low variance
low bias
high variance
Error
testing
training
Model complexity

Bias-variance trade-off





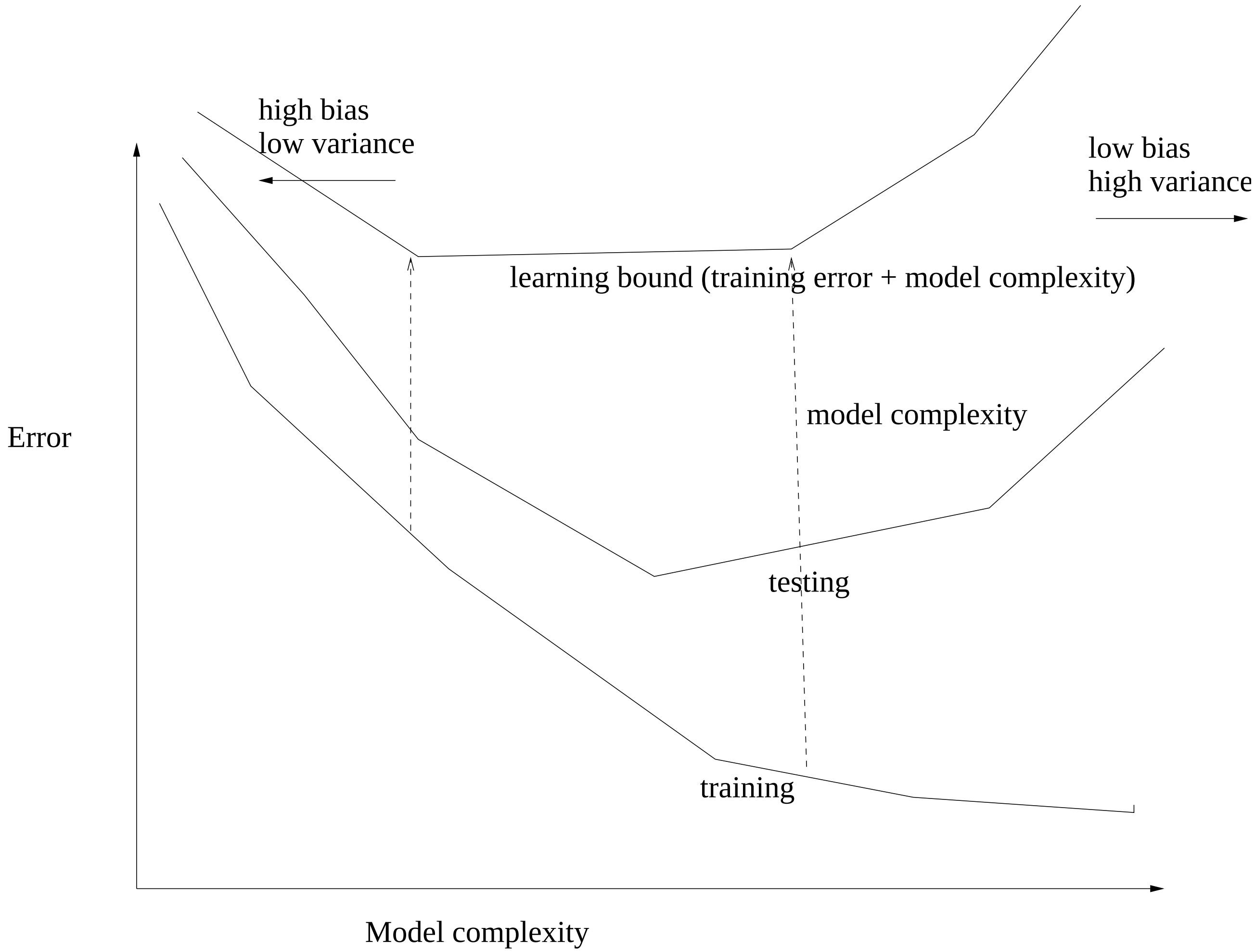
high bias
low variance
low bias
high variance
Error
testing
(bias+var+noise)
variance
training
bias
Model complexity

Estimating generalization error from training error

- generalization-error = training-error + model-complexity
  - training-error: measuring bias of the learning algorithm
  - model-complexity: measuring variance of the learning algorithm

- Generalization analysis: let $\hat{f} = \mathcal{A}(S_n)$, then with probability at least $1 - \eta$,

$$\underbrace{\mathbf{E}_{X,Y} \phi(\hat{f}(X), Y)}_{\text{generalization error}} \leq \underbrace{\frac{1}{n} \sum_{i=1}^{n} \phi(\hat{f}(X_i), Y_i)}_{\text{training error}} + \underbrace{Q_n(\mathcal{H}, \eta)}_{\text{model complexity} \to 0 \text{ as } n \to \infty}.$$


high bias
low variance
low bias
high variance
learning bound (training error + model complexity)
model complexity
Error
testing
training
Model complexity

Uniform Convergence

- Recall Empirical risk minimization learner:

$$\boldsymbol{\hat{f} = \arg\min_{f \in \mathcal{H}} \sum_{i=1}^{n} \phi(f(X_i), Y_i),}$$

$\mathcal{H}$: set of candidate prediction rules (e.g. linear combination of features).

- We have $\forall \epsilon > 0$:

$$P\left(\mathbf{E}_{X,Y}\phi(\hat{f}(X), Y) > \frac{1}{n}\sum_{i=1}^{n}\phi(\hat{f}(X_i), Y_i) + \epsilon\right) \leq \eta_{\phi(\mathcal{H})}(\epsilon),$$

where $\phi(\mathcal{H}) = \{\phi(f(X), Y) : f \in \mathcal{H})\}$, and for any function class $\mathcal{H}$:

$$\eta_{\phi(\mathcal{H})}(\epsilon) = \underbrace{P\left(\sup_{f \in \mathcal{H}}\left(\mathbf{E}_{X,Y}\phi(f(X), Y) - \frac{1}{n}\sum_{i=1}^{n}\phi(f(X_i), Y_i)\right) > \epsilon\right)}_{\text{one-sided uniform convergence over family } \phi(\mathcal{H})}.$$

- Given uniform convergence bound, model complexity can be estimated as:

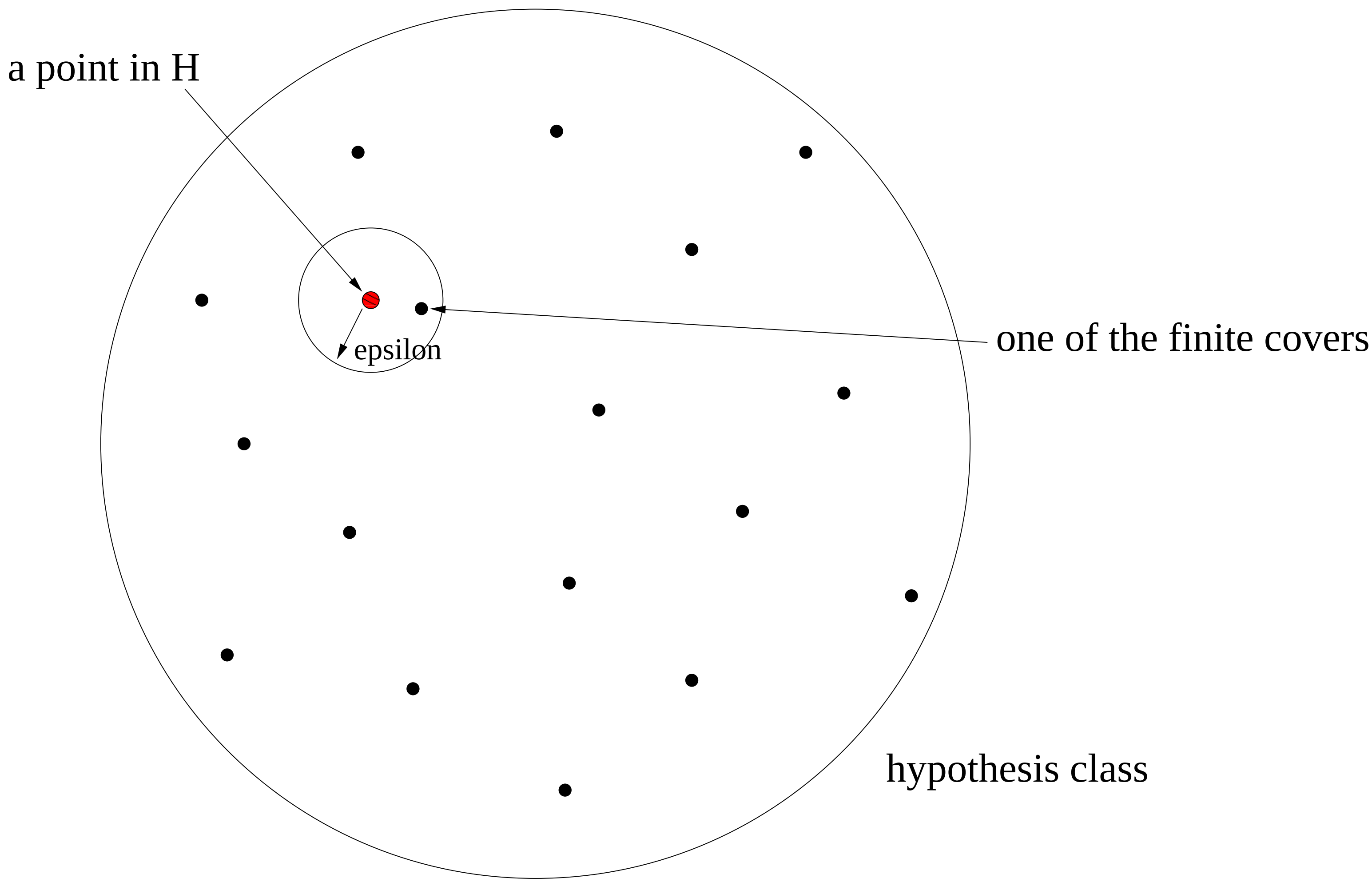
$$Q_n(\mathcal{H}, \eta) = \eta_{\phi(\mathcal{H})}^{-1}(\eta) = \inf\{\epsilon : \eta_{\phi(\mathcal{H})}(\epsilon) \le \eta\}.$$

Covering numbers: size of function family

- If $\mathcal{H}$ is finite, then (union bound)

$$\eta_{\phi(\mathcal{H})}(\epsilon) = |\mathcal{H}| \sup_{f \in \mathcal{H}} \underbrace{P\left(\mathbf{E}_{X,Y}\phi(f(X),Y) - \frac{1}{n}\sum_{i=1}^{n}\phi(f(X_i),Y_i) > \epsilon\right)}_{\text{convergence of single hypothesis}}.$$

- What if family not finite?
  - Approximate by a finite number of functions (covering)
  - Let $\mathcal{H}$ be a hypothesis family, and $\epsilon > 0$, then the covering number $N(\mathcal{H}, \epsilon)$ is the smallest number of functions $\{f_j\}$ such that $\forall f \in \mathcal{H}$, then $\min_j d(f_j, f) \leq \epsilon$.



a point in H
epsilon
one of the finite covers
hypothesis class

L_∞ covering number

- Define $d(f, f') = \sup_{X,Y} |f(X) - f'(X)|$
  - the covering number is denoted as $N_\infty(\mathcal{H}, \epsilon)$
- Similarly, $d(f, f') = \sup_{X,Y} |\phi(f(X), Y) - \phi(f'(X), Y)|$
  - the coverining number is denoted as $N_\infty(\phi(\mathcal{H}), \epsilon)$
- If $|\phi(f, y) - \phi(f', y)| \leq \gamma |f - f'|$ (Lipschitz in f), then $N_\infty(\phi(\mathcal{H}), \gamma\epsilon) \leq N_\infty(\mathcal{H}, \epsilon)$.

Uniform Convergence Bound Using L_∞ covering number

Let f_j be a $N_\infty(\phi(\mathcal{H}), \epsilon/4)$ cover of $\phi(\mathcal{H})$, then

$$P\left(\sup_{f\in\mathcal{H}}\left(\mathbf{E}_{X,Y}\phi(f(X),Y) - \frac{1}{n}\sum_{i=1}^{n}\phi(f(X_i),Y_i)\right) > \epsilon\right)$$

$$\leq P\left(\exists f \in \mathcal{H}; \forall j: \text{if } |\mathbf{E}_{X,Y}\phi(f_j(X),Y) - \mathbf{E}_{X,Y}\phi(f(X),Y)| \leq \epsilon/4\right.$$

$$\left|\frac{1}{n}\sum_i \phi(f_j(X_i),Y_i) - \sum_i \phi(f(X_i),Y_i)\right| \leq \epsilon/4$$

$$\left.\text{then } \left(\mathbf{E}_{X,Y}\phi(f_j(X),Y) - \frac{1}{n}\sum_{i=1}^{n}\phi(f_j(X_i),Y_i)\right) > \epsilon/2\right)$$

$$\leq P\left(\sup_j\left(\mathbf{E}_{X,Y}\phi(f_j(X),Y) - \frac{1}{n}\sum_{i=1}^{n}\phi(f_j(X_i),Y_i)\right) > \epsilon/2\right)$$

$$\leq N_\infty(\phi(\mathcal{H}), \epsilon/4)\sup_j P\left(\left(\mathbf{E}_{X,Y}\phi(f_j(X),Y) - \frac{1}{n}\sum_{i=1}^{n}\phi(f_j(X_i),Y_i)\right) > \epsilon/2\right).$$

Exponential Tail Bound

- Estimating $P\left(\left(\mathbf{E}_{X,Y}\phi(f_j(X),Y) - \frac{1}{n}\sum_{i=1}^n \phi(f_j(X_i),Y_i)\right) > \epsilon/2\right)$

- Let $z = \phi(f_j(x),y)$ and $z_i = \phi(f_j(X_i),Y_i)$ be iid (independent, identically, distributed) random variables, we want to estimate $\mathbf{E}z - \frac{1}{n}\sum_{i=1}^n z_i$
  - convergence of empirical mean of a random variable to its true mean
  - law of large numbers

- Want a bound of the form (for all j):

$$P(\mathbf{E}z - \frac{1}{n}\sum_{i=1}^n z_i \geq \epsilon/2) \leq ae^{-n\epsilon^2/b^2},$$

implying

$$\eta_{\phi(\mathcal{H})}(\epsilon) \leq a N_\infty(\phi(\mathcal{H}), \epsilon/4) e^{-n\epsilon^2/b^2},$$

and thus

$$Q_n(\mathcal{H}, \eta) = \eta^{-1}_{\phi(\mathcal{H})}(\eta) \leq \inf\{\epsilon : \epsilon \geq b\sqrt{\ln[a N_\infty(\phi(\mathcal{H}), \epsilon/4)/\eta]/n}\}.$$

- Learning Bound for empirical risk minimization

$$\underbrace{\mathbf{E}_{X,Y}\phi(\hat{f}(X), Y)}_{\text{generalization error}} \leq \underbrace{\frac{1}{n}\sum_{i=1}^{n} \phi(\hat{f}(X_i), Y_i)}_{\text{training error}} + \underbrace{\inf\{\epsilon : \epsilon \geq b\sqrt{\ln[a N_\infty(\phi(\mathcal{H}), \epsilon/4)/\eta]/n}\}}_{\text{model complexity}}.$$

  - why exponential tail bound?
  - learning complexity of form $n^{-1} \ln N_\infty(\phi(\mathcal{H}), \epsilon/4)$.
  - test error similar to training error as long as the number of functions in the family is sub-exponential in n

Hoeffding Inequality

Assume $z \in [0,1]$, then

$$P(\mathbf{E}z - \frac{1}{n}\sum_{i=1}^{n} z_i \geq \epsilon) \leq \exp(-2\epsilon^2)$$

$$P(\mathbf{E}z - \frac{1}{n}\sum_{i=1}^{n} z_i \leq -\epsilon) \leq \exp(-2\epsilon^2).$$

Thus

$$P(\mathbf{E}z - \frac{1}{n}\sum_{i=1}^{n} z_i \geq \epsilon/2) \leq \exp(-\epsilon^2/2).$$

exponential inequality holds with $a = 1$ and $b = \sqrt{2}$.

Hoeffding Inequality: Proof

$$
\begin{aligned}
&P(\mathbf{E}z - \frac{1}{n}\sum_{i=1}^{n} z_i \geq \epsilon)e^{\lambda\epsilon} \leq \mathbf{E}e^{\lambda(\mathbf{E}z - \frac{1}{n}\sum_{i=1}^{n} z_i)} \\
\leq &\mathbf{E}\prod_{i=1}^{n} e^{\lambda/n(\mathbf{E}z - z_i)} = \underbrace{[\mathbf{E}_{z_i} e^{\lambda/n(\mathbf{E}z - z_i)}]^n}_{\text{max achieved at } z_i = 0 \text{ or } 1} \\
\leq &[e^{\lambda/n(\mathbf{E}z-1)}\mathbf{E}z + (1 - \mathbf{E}z)e^{\lambda/n\mathbf{E}z}]^n \quad (*) \\
\leq &[e^{(\lambda/n)^2/8}]^n. \quad (**)
\end{aligned}
$$

Taking $\lambda = 4n\epsilon$, we have $P(\mathbf{E}z - \frac{1}{n}\sum_{i=1}^{n} z_i \geq \epsilon) \leq e^{-2n\epsilon^2}$.

Details

- Proof of $(*)$: using Jensens: $e^x \leq (1 - x/a) + x/ae^a$ for all $0 \leq x \leq 1$, we have

$$\mathbf{E}_{z_i} e^{\lambda/n(\mathbf{E}z - z_i)} \leq e^{\lambda/n(\mathbf{E}z-1)} \mathbf{E}_{z_i}[(1-(1-z_i)) + (1-z_i)e^{\lambda/n}]$$

- Proof of $(**)$: need to show that when $x \in [0,1]$

$$e^{\alpha(x-1)}x + (1-x)e^{\alpha x} \leq \underbrace{0.5[e^{-\alpha/2} + e^{\alpha/2}]}_{\text{achieved at } x = 0.5} \leq e^{\alpha^2/8}.$$

The last inequlity follows from comparing Taylor expansion of

$$\ln[0.5e^{-\alpha/2} + 0.5e^{\alpha/2}] \leq \alpha^2/8.$$

Empirical (sample dependent) covering numbers

- distance d is data dependent.
- e.g. empirical L_∞ covering number: distance is $d(f, f'|S_n) = \sup_{(X_i,Y_i)} |\phi(f(X_i), Y_i) - \phi(f'(X_i), Y_i)|$
- empirical L_∞ cover can be finite when L_∞ cover is infinite.
- Learning bound can be obtained using empirical covering numbers (shown in later lectures)

References

- Books on empirical process:
  - A. W. van der Vaart and J. A. Wellner. *Weak convergence and empirical processes. Springer Series in Statistics. Springer-Verlag, New York, 1996.*
- Hoeffding Inequality:
  - W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, March 1963.