Learning theory
Lecture 9

David Sontag
New York University

Slides adapted from Carlos Guestrin & Luke Zettlemoyer

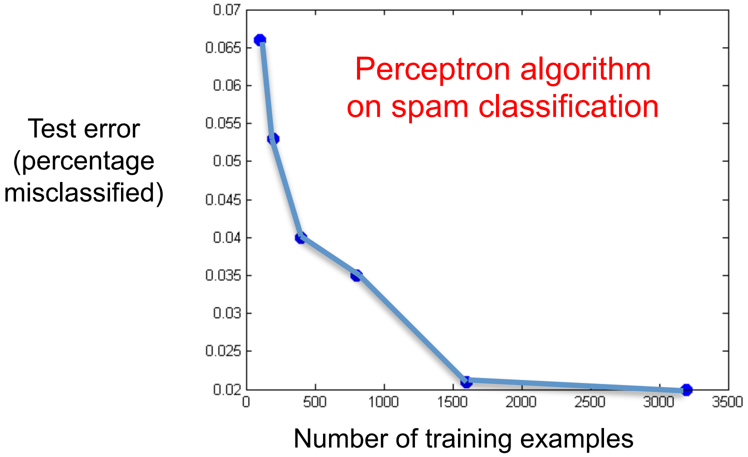
Roadmap of next lectures

1. Generalization of finite hypothesis spaces

2. VC-dimension

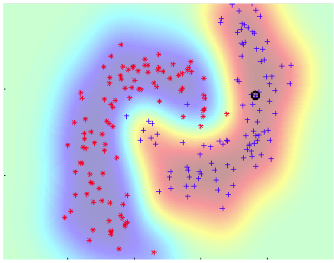
   - Will show that **linear** classifiers need to see approximately **d** training points, where **d** is the dimension of the feature vectors

   - Explains the good performance we obtained using perceptron!!!! (we had a few thousand features)



3. Margin based generalization

   - Applies to **infinite** dimensional feature vectors (e.g., Gaussian kernel)

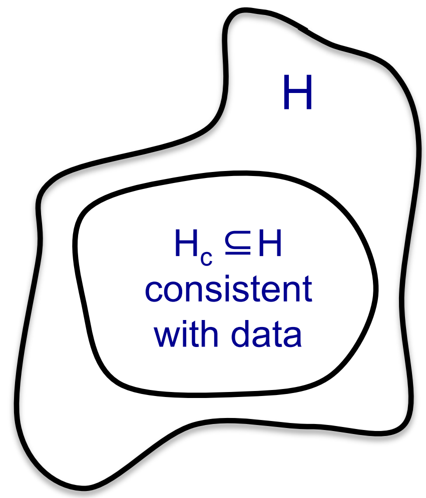


[Figure from Cynthia Rudin]

How big should your validation set be?

- In PS1, you tried many configurations of your algorithms (avg vs. regular perceptron, max # of iterations) and chose the one that had smallest validation error
- Suppose in total you tested **|H|=**40 different classifiers on the validation set of **m** held-out e-mails
- The best classifier obtains 98% accuracy on these **m** e-mails!!!
- But, what is the true classification accuracy?
- How large does **m** need to be so that we can guarantee that the best configuration (measured on validate) is truly good?

A simple setting…



- Classification
  - m data points
  - **Finite** number of possible hypothesis (e.g., 40 spam classifiers)

- A learner finds a hypothesis *h* that is **consistent** with training data
  - Gets zero error in training: $error_{train}(h) = 0$
  - I.e., assume for now that one of the classifiers gets 100% accuracy on the **m** e-mails (we'll handle the 98% case afterward)

- What is the probability that *h* has more than ε **true** error?
  - $error_{true}(h) \geq \varepsilon$

Intro to probability: outcomes

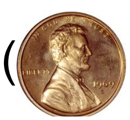
- An **outcome space** specifies the possible outcomes that we would like to reason about, e.g.

Ω = {  ,  }   Coin toss

Ω = { , , , , ,  }   Die toss

- We specify a **probability** p(**x**) for each outcome **x** such that

$$p(x) \geq 0, \qquad \sum_{x \in \Omega} p(x) = 1$$

E.g., p() = .6

p() = .4

Intro to probability: events

- An **event** is a subset of the outcome space, e.g.

E = { , ,  } Even die tosses

O = { , ,  } Odd die tosses

- The **probability** of an event is given by the sum of the probabilities of the outcomes it contains,

$$p(E) = \sum_{x \in E} p(x)$$

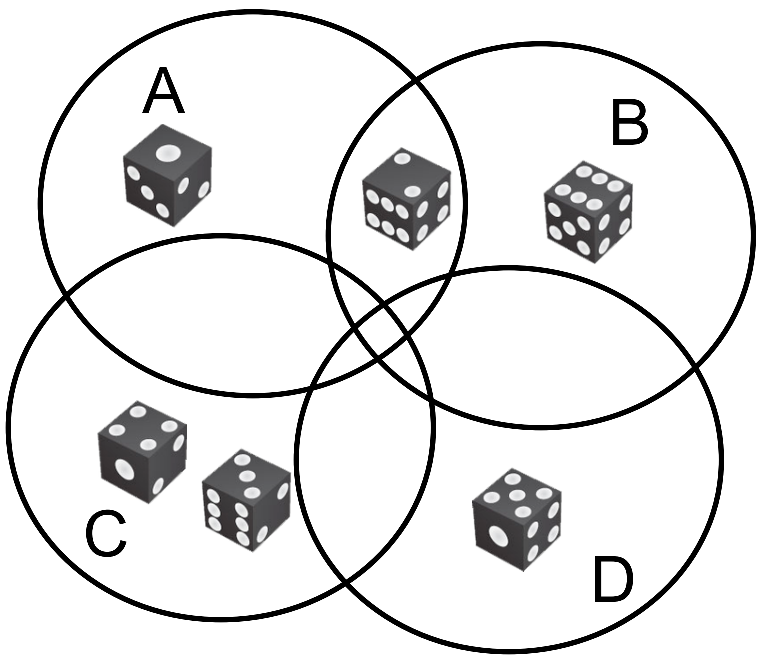
E.g., p(E) = p() + p() + p()

= 1/2, if fair die

Intro to probability: union bound

- P(A or B or C or D or …)

  ≤ P(A) + P(B) + P(C) + P(D) + …



$$p(A \cup B) = p(A) + p(B) - p(A \cap B)$$
$$\leq p(A) + p(B)$$

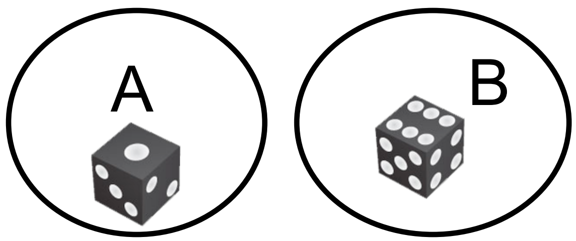
Q: When is this a tight bound?

A: For disjoint events
(i.e., non-overlapping circles)

Intro to probability: independence

- Two events A and B are **independent** if

$$p(A \cap B) = p(A)p(B)$$



Are these events independent?

No! $p(A \cap B) = 0$

$$p(A)p(B) = \left(\frac{1}{6}\right)^2$$

Intro to probability: independence

- Two events A and B are **independent** if

$$p(A \cap B) = p(A)p(B)$$

Analogy: outcome space defines all possible sequences of e-mails in training set

- Suppose our outcome space had two different die:

$\Omega = \{$  $\}$ 2 die tosses

$6^2 = 36$ outcomes

and the probability of each outcome is defined as

p() = a_1 b_1 p() = a_1 b_2 $\cdots$

a_1	a_2	a_3	a_4	a_5	a_6
.1	.12	.18	.2	.1	.3

$$\sum_{i=1}^{6} a_i = 1$$

b_1	b_2	b_3	b_4	b_5	b_6
.19	.11	.1	.22	.18	.2

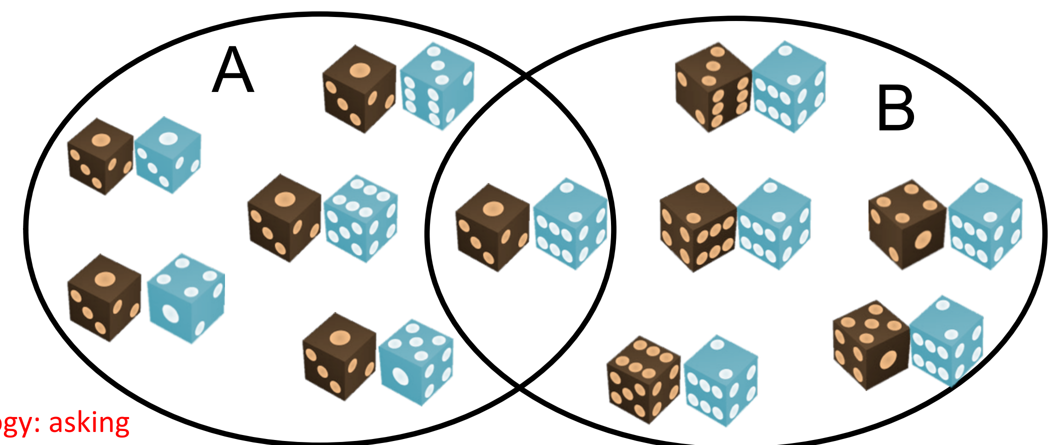
$$\sum_{j=1}^{6} b_j = 1$$

Intro to probability: independence

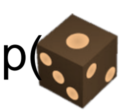
- Two events A and B are **independent** if

$$p(A \cap B) = p(A)p(B)$$

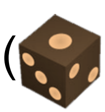
- Are these events independent?



Yes! $p(A \cap B) =$ p()

$p(A)p(B) =$ p() p()

Analogy: asking about first e-mail in training set

p(A) = p()

$$= \sum_{j=1}^{6} a_1 b_j = a_1 \sum_{j=1}^{6} b_j = a_1$$

p(B) = p() $= b_2$

Analogy: asking about second e-mail in training set

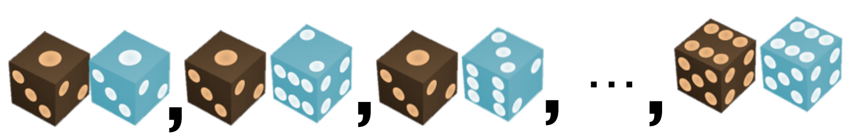
Intro to probability: discrete random variables

- A **random variable** X is a mapping $X : \Omega \to D$
  - D is some set (e.g., the integers)
  - Induces a partition of all outcomes Ω
- For some $x \in D$, we say

$$p(X = x) = p(\{\omega \in \Omega : X(\omega) = x\})$$

"probability that variable X assumes state x"

- Notation: $\mathrm{Val}(X)$ = set D of all values assumed by X
  (will interchangeably call these the "values" or "states" of variable X)

$\Omega = \{$  $\}$ 2 die tosses

Intro to probability: discrete random variables

- $p(X)$ is a distribution: $\sum_{x \in \mathrm{Val}(X)} p(X = x) = 1$
- E.g. X_1 may refer to the value of the first dice, and X_2 to the value of the second dice
- We call two random variables X and Y *identically distributed* if Val(X) = Val(Y) and p(X=s) = p(Y=s) for all s in Val(X)

p() = a_1 b_1 p() = a_1 b_2 $\cdots$

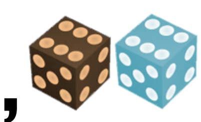
X_1 and X_2 NOT identically distributed

a_1	a_2	a_3	a_4	a_5	a_6
.1	.12	.18	.2	.1	.3

$$\sum_{i=1}^{6} a_i = 1$$

b_1	b_2	b_3	b_4	b_5	b_6
.19	.11	.1	.22	.18	.2

$$\sum_{j=1}^{6} b_j = 1$$

$\Omega = \{$ , , , $\cdots$,  $\}$ 2 die tosses

Intro to probability: discrete random variables

- $p(X)$ is a distribution: $\sum_{x \in \mathrm{Val}(X)} p(X = x) = 1$

- E.g. X_1 may refer to the value of the first dice, and X_2 to the value of the second dice

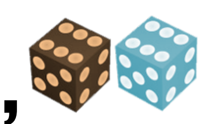
- We call two random variables X and Y *identically distributed* if Val(X) = Val(Y) and p(X=s) = p(Y=s) for all s in Val(X)

p() = $a_1\ a_1$ p() = $a_1\ a_2$ …

a_1	a_2	a_3	a_4	a_5	a_6
.1	.12	.18	.2	.1	.3

$$\sum_{i=1}^{6} a_i = 1$$

X_1 and X_2 identically distributed

$\Omega = \{$ , , , $\ldots$,  $\}$ 2 die tosses

Intro to probability: discrete random variables

- X=x is simply an event, so can apply union bound, etc.
- Two random variables **X** and **Y** are **independent** if:

$$p(X = x, Y = y) = p(X = x)p(Y = y) \quad \forall x \in \mathrm{Val}(X), y \in \mathrm{Val}(Y)$$

Joint probability. Formally, given by the event $X = x \cap Y = y$

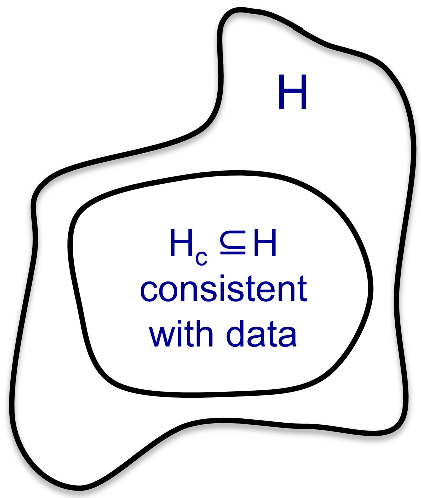
- The **expectation** of **X** is defined as: $E[X] = \sum_{x \in \mathrm{Val}(X)} p(X = x)x$

- If **X** is binary valued, i.e. x is either 0 or 1, then:

$$\begin{aligned} E[X] &= p(X = 0) \cdot 0 + p(X = 1) \cdot 1 \\ &= p(X = 1) \end{aligned}$$

- Linearity of expectations: $E\Big[\sum_{i=1}^{n} X_i\Big] = \sum_{i=1}^{n} E[X_i]$

A simple setting…



- Classification
  - m data points
  - **Finite** number of possible hypothesis (e.g., 40 spam classifiers)
- A learner finds a hypothesis *h* that is **consistent** with training data
  - Gets zero error in training: $error_{train}(h) = 0$
  - I.e., assume for now that one of the classifiers gets 100% accuracy on the **m** e-mails (we'll handle the 98% case afterward)
- What is the probability h correctly classifies all **m** data points given that *h* has more than ε **true** error?
  - $error_{true}(h) \geq \varepsilon$

How likely is a **single** hypothesis to get *m* data points right?

- The probability of a hypothesis *h* incorrectly classifying: $\epsilon_h = \sum_{(\vec{x},y)} p(\vec{x},y) 1[h(\vec{x}) \neq y]$
- Let Z_i^h be a random variable that takes two values: **1 if h correctly classifies** ith data point, and 0 otherwise
- The Z^h variables are **independent** and **identically distributed** (i.i.d.) with

$$\Pr(Z_i^h = 0) = \sum_{(\vec{x},y)} p(\vec{x},y) 1[h(\vec{x}) \neq y] \; = \epsilon_h$$

- What is the probability that *h* classifies *m* data points correctly?

Pr(h gets m *iid* data points right) $= (1 - \epsilon_h)^m \leq e^{-\epsilon_h m}$

Are we done?

$\Pr(\text{h gets m } \textit{iid} \text{ data points right} \mid \text{error}_{true}(h) \geq \varepsilon) \leq e^{-\varepsilon m}$

- Says "with probability > $1-e^{-\varepsilon m}$, if h gets m data points correct, then it is close to perfect (will have error $\leq \varepsilon$)"
- This only considers **one** hypothesis!
- Suppose 1 billion classifiers were tried, and each was a *random* function
- For **m** small enough, one of the functions will classify all points correctly – but all have very large true error

How likely is learner to pick a bad hypothesis?

Pr(h gets m *iid* data points right | $\text{error}_{\text{true}}(h) \geq \varepsilon) \leq e^{-\varepsilon m}$

Suppose there are $|H_c|$ hypotheses consistent with the training data

- How likely is learner to pick a bad one, i.e. with *true* error $\geq \varepsilon$?
- We need a bound that holds for all of them!

$$P(\text{error}_{\text{true}}(h_1) \geq \varepsilon \text{ OR } \text{error}_{\text{true}}(h_2) \geq \varepsilon \text{ OR } \ldots \text{ OR } \text{error}_{\text{true}}(h_{|H_c|}) \geq \varepsilon)$$

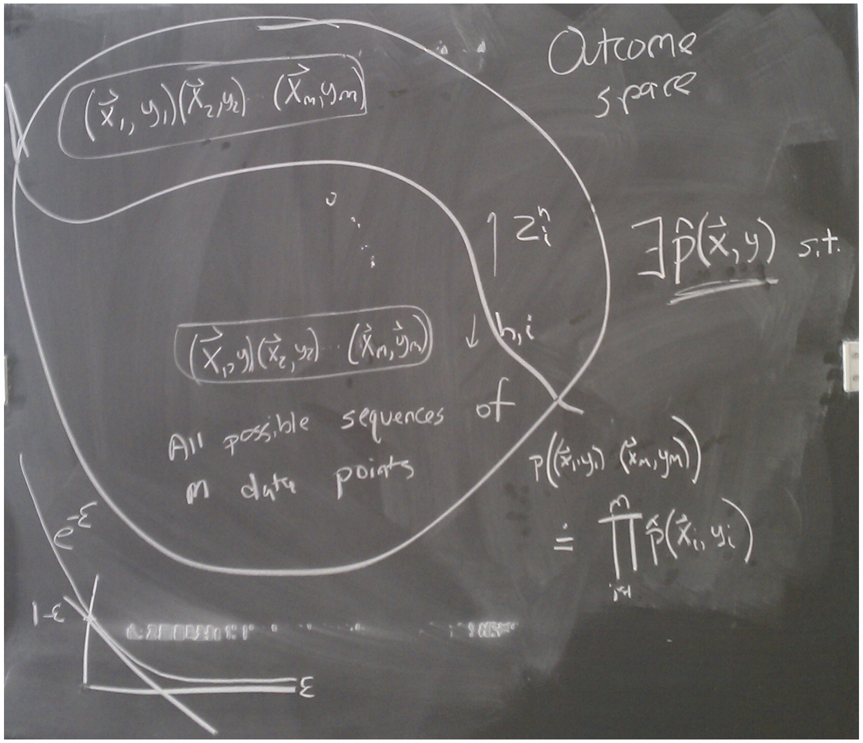
$\leq \sum_k P(\text{error}_{\text{true}}(h_k) \geq \varepsilon)$ ← Union bound

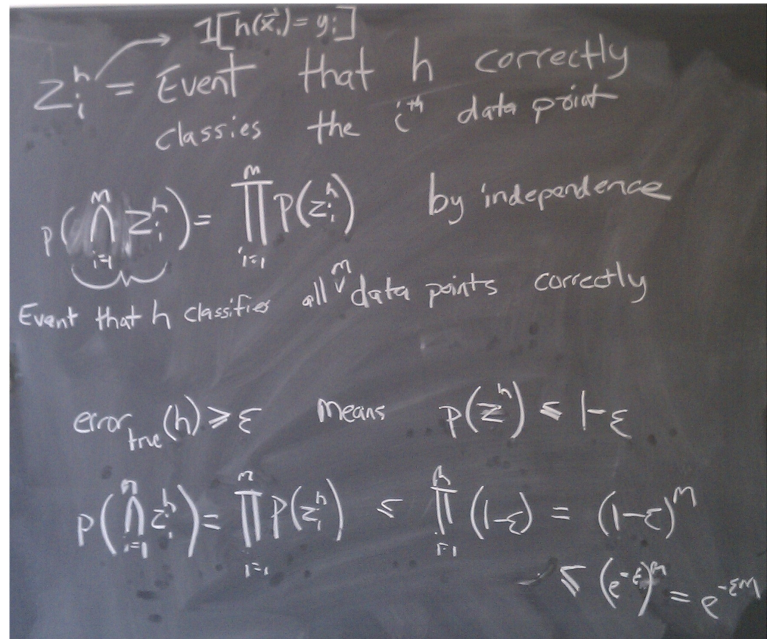
$\leq \sum_k (1-\varepsilon)^m$ ← bound on individual h_ks

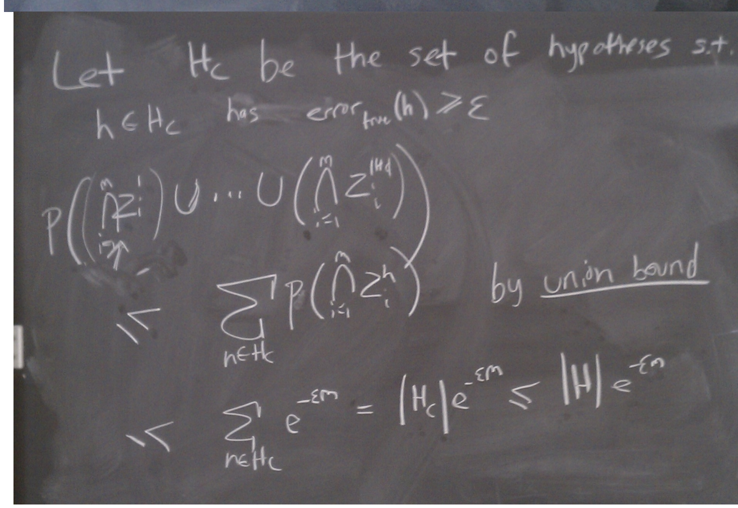
$\leq |H|(1-\varepsilon)^m$ ← $|H_c| \leq |H|$

$\leq |H|\, e^{-m\varepsilon}$ ← $(1-\varepsilon) \leq e^{-\varepsilon}$ for $0 \leq \varepsilon \leq 1$

Extra analysis







Generalization error of finite hypothesis spaces [Haussler '88]

We just proved the following result:

Theorem: Hypothesis space H finite, dataset D with m i.i.d. samples, $0 < \varepsilon < 1$: for any learned hypothesis h that is consistent on the training data:

$$P(\text{error}_{true}(h) > \epsilon) \leq |H|e^{-m\epsilon}$$

Using a PAC bound

Argument: Since for all *h* we know that

$$P(\text{error}_{true}(h) > \epsilon) \leq |H|e^{-m\epsilon}$$

… with probability 1-δ the following holds… (either case 1 or case 2)

Typically, 2 use cases:

- 1: Pick ε and δ, compute *m*
- 2: Pick m and δ, compute ε

$$\left.\begin{array}{l} p(\text{error}_{true}(h) \geq \epsilon) \leq |H|e^{-m\epsilon} \leq \delta \end{array}\right\}$$

Says: we are willing to tolerate a δ probability of having ≥ ε error

$\epsilon = \delta = .01, |H| = 40$

Need $m \geq 830$

$$\ln\left(|H|e^{-m\epsilon}\right) \leq \ln\delta$$

$$\ln|H| - m\epsilon \leq \ln\delta$$

Case 1

$$m \geq \frac{\ln|H| + \ln\frac{1}{\delta}}{\epsilon}$$

Case 2

$$\epsilon \geq \frac{\ln|H| + \ln\frac{1}{\delta}}{m}$$

Log dependence on |H|, OK if exponential size (but not doubly)

ε has stronger influence than δ

ε shrinks at rate O(1/m)

Limitations of Haussler '88 bound

- There may be no consistent hypothesis h (where $error_{train}(h)$=0)

- Size of hypothesis space
  - What if |H| is really big?
  - What if it is continuous?

- First Goal: Can we get a bound for a learner with $error_{train}(h)$ in the data set?

Question: What's the expected error of a hypothesis?

- The probability of a hypothesis incorrectly classifying: $\sum_{(\vec{x},y)} p(\vec{x},y) 1[h(\vec{x}) \neq y]$
- Let's now let Z_i^h be a random variable that takes two values, 1 if h correctly classifies i^{th} data point, and 0 otherwise
- The Z variables are **independent** and **identically distributed** (i.i.d.) with

$$\Pr(Z_i^h = 0) = \sum_{(\vec{x},y)} p(\vec{x},y) 1[h(\vec{x}) \neq y]$$

- Estimating the true error probability is like estimating the parameter of a coin!

- **Chernoff bound**: for *m* i.i.d. coin flips, $X_1,\ldots,X_m$, where $X_i \in \{0,1\}$. For $0<\varepsilon<1$:

$$p(X_i = 1) = \theta$$

$$\boxed{P\left(\theta - \frac{1}{m}\sum_i x_i > \epsilon\right) \leq e^{-2m\epsilon^2}}$$

True error probability

Observed fraction of points incorrectly classified

$$E[\frac{1}{m}\sum_{i=1}^m X_i] = \frac{1}{m}\sum_{i=1}^m E[X_i] = \theta$$

(by linearity of expectation)

Generalization bound for |H| hypothesis

Theorem: Hypothesis space *H* finite, dataset *D* with *m* i.i.d. samples, 0 < ε < 1 : for any learned hypothesis *h*:

$$\Pr(\mathrm{error}_{true}(h) - \mathrm{error}_D(h) > \epsilon) \leq |H| e^{-2m\epsilon^2}$$

Why? Same reasoning as before. Use the Union bound over individual Chernoff bounds

PAC bound and Bias-Variance tradeoff

for all h, with probability at least 1-δ:

$$\underbrace{\text{error}_{true}(h) \leq \text{error}_D(h)}_{\text{“bias”}} + \underbrace{\sqrt{\frac{\ln|H| + \ln\frac{1}{\delta}}{2m}}}_{\text{“variance”}}$$

- For large |H|
  - low bias (assuming we can find a good h)
  - high variance (because bound is looser)
- For small |H|
  - high bias (is there a good h?)
  - low variance (tighter bound)