University of Colorado **Boulder**

Department of Computer Science
CSCI 5622: Machine Learning
Chris Ketelsen

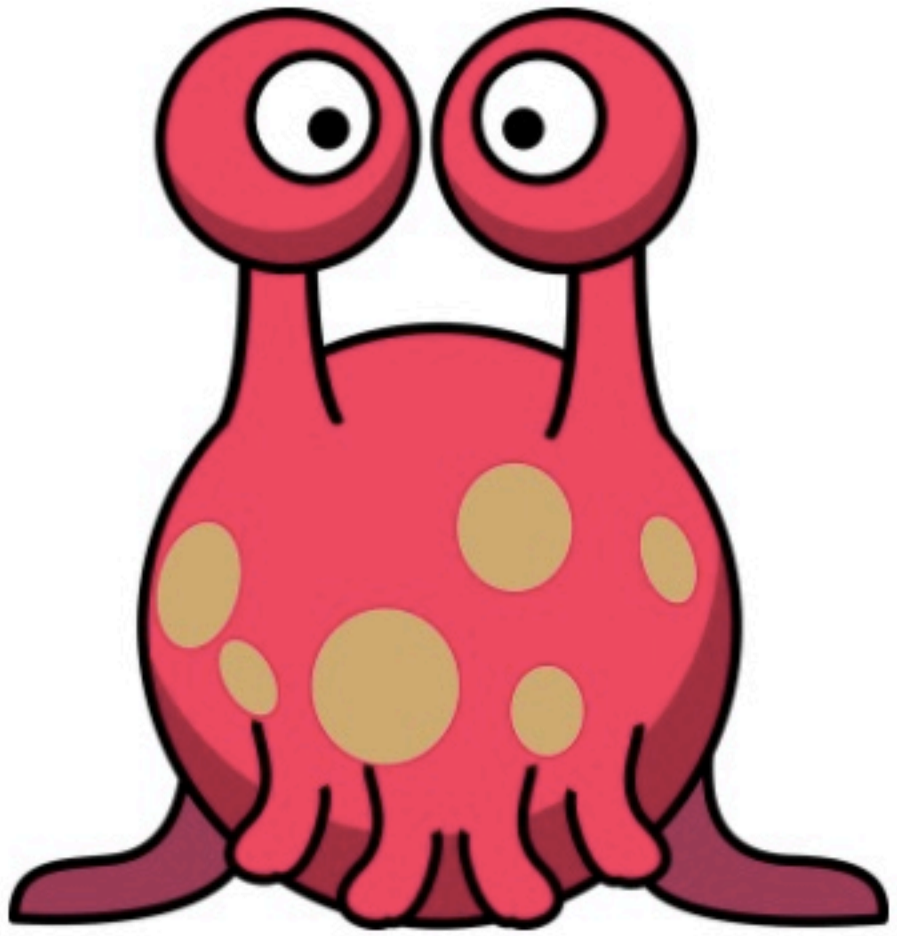
Lecture 10: Learning Theory Part 1
The PAC Framework

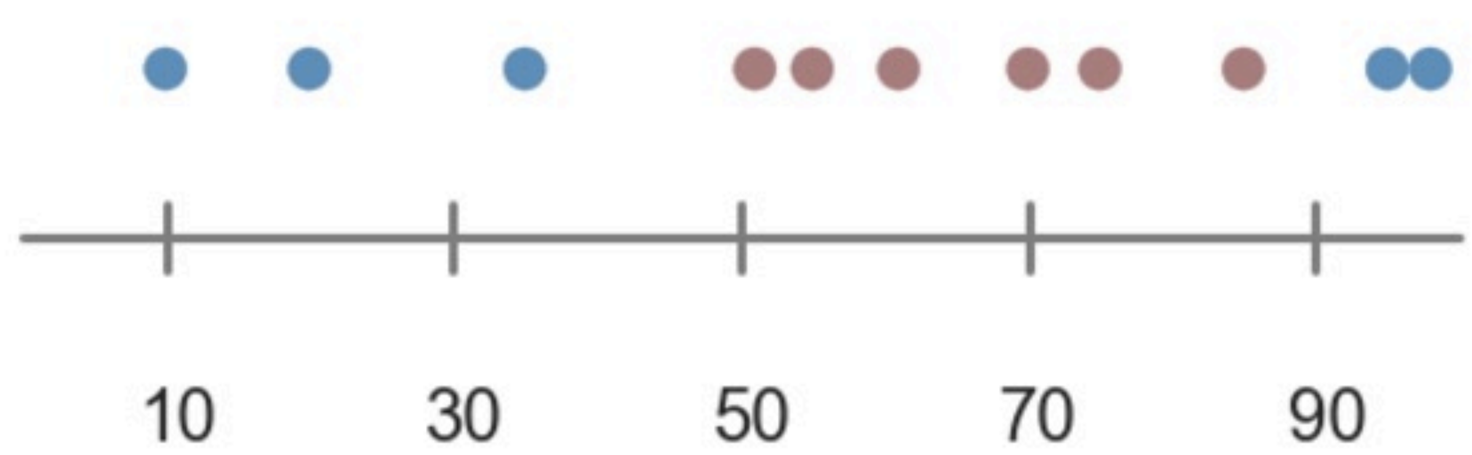
Learning Objectives

- [ ] Prove some simple bounds on errors and sample sizes
- [ ] Gain some intuition about complexity and overfitting
- [ ] See the Bias-Variance Trade-Off from a different perspective

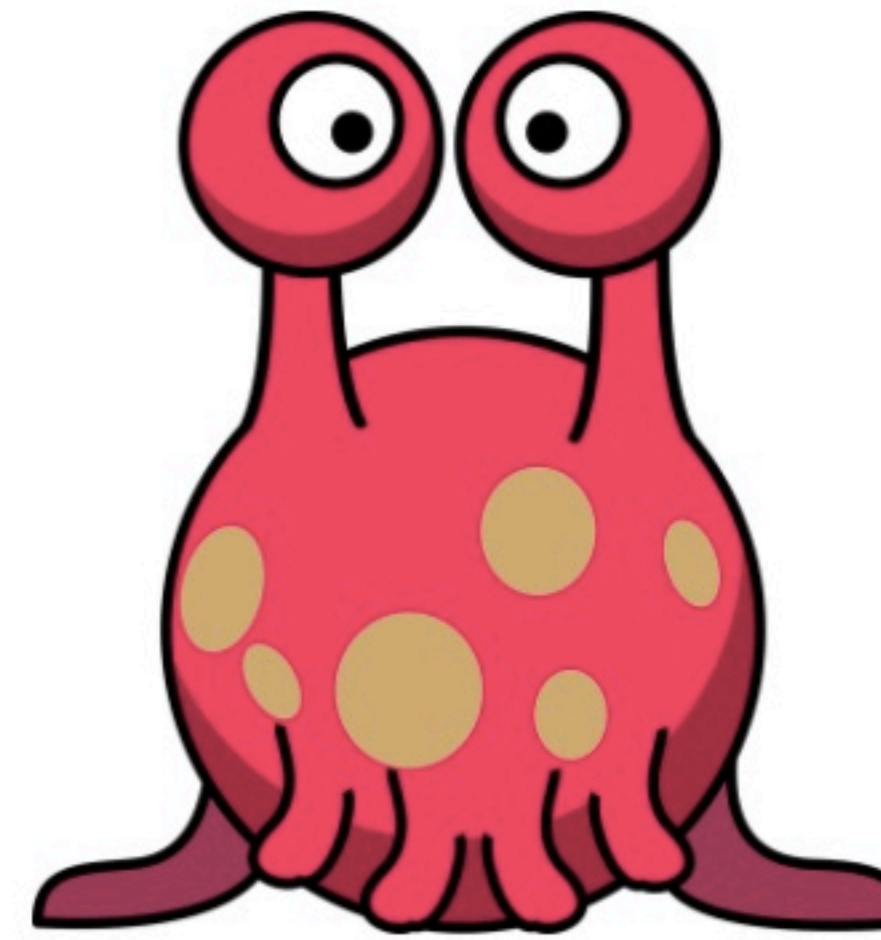
Intro to PAC

- Alien moves to Colorado
- Want to talk to locals about weather
- Specifically about when weather is *nice*
- Asks a bunch of locals if it's *nice* out
- Gets labeled observations $S = \{(x_i, y_i)\}_{i=1}^m$
- Coloradans have *concept* $c(x)$ of *nice*
- Alien wants to learn *hypothesis* $h(x)$





Intro to PAC



- Alien moves to Colorado
- Want to talk to locals about weather
- Specifically about when weather is *nice*
- Asks a bunch of locals if it's *nice* out
- Gets labeled observations $S = \{(x_i, y_i)\}_{i=1}^{m}$
- Coloradans have *concept* $c(x)$ of *nice*
- Alien wants to learn *hypothesis* $h(x)$

- How many locals does he need to ask to get $h(x)$ that is 99% accurate with 98% confidence?

Intro to PAC

Assumptions:

- Data comes from distribution D
- Concept c comes from concept class C
- Hypothesis h comes from hypothesis class H

Def: **Generalization Error**

$$R(h) = Pr_{x \sim D}\left[h(x) \neq c(x)\right]$$

Goal: Given a set of data S of size m, can we learn a hypothesis h that we can say is *accurate* with high *confidence*?

Intro to PAC

Assumptions:

- Data comes from distribution D
- Concept c comes from concept class C
- Hypothesis h comes from hypothesis class H

Def: **Generalization Error**

$$R(h) = Pr_{x \sim D}\left[h(x) \neq c(x)\right]$$

Def: **Training Error** (sometimes called *Empirical Risk*)

$$\hat{R}(h) = \frac{1}{m} \sum_{i=1}^{m} I_{h(x_i) \neq c(x_i)}$$

Intro to PAC

We say that a concept is PAC-Learnable if we can find a hypothesis that is **P**robably **A**pproximately **C**orrect using a training set S of size m where m isn't too large.

$$R(h_S) \leq \epsilon$$

- **Approximately Correct**: Accuracy is $1 - \epsilon$

Intro to PAC

We say that a concept is PAC-Learnable if we can find a hypothesis that is **P**robably **A**pproximately **C**orrect using a training set S of size m where m isn't too large.

$$Pr_{S \sim \mathcal{D}^m}\left[R(h_S) \leq \epsilon\right] \geq 1 - \delta$$

- **Approximately Correct**: Accuracy is $1 - \epsilon$
- **Probably**: Confidence in hypothesis is $1 - \delta$

PAC = Probably Approximately Correct

Intro to PAC

Def: **PAC Learnability** - A concept from class C is PAC-Learnable if there exists an algorithm $\mathcal{A}$ and a polynomial function f such that for any $\epsilon > 0$ and any $\delta > 0$, for all distributions $\mathcal{D}$ and any concept $c \in C$, it holds that

$$Pr_{S \sim \mathcal{D}^m} [R(h_S) \leq \epsilon] \geq 1 - \delta$$

for any sample size $m \geq f(1/\epsilon, 1/\delta, n, \text{size}(c))$:

Intro to PAC

Def: **PAC Learnability** - A concept from class C is PAC-Learnable if there exists an algorithm $\mathcal{A}$ and a polynomial function f such that for any $\epsilon > 0$ and any $\delta > 0$, for all distributions $\mathcal{D}$ and any concept $c \in C$, it holds that

$$Pr_{S\sim\mathcal{D}^m}[R(h_S) \leq \epsilon] \geq 1 - \delta$$

for any sample size $m \geq f(1/\epsilon, 1/\delta, n, \text{size}(c))$:

- S: The training set we learn from
- $\mathcal{D}$: The distribution the data comes from
- h_S: The hypothesis we learn from training set

Intro to PAC

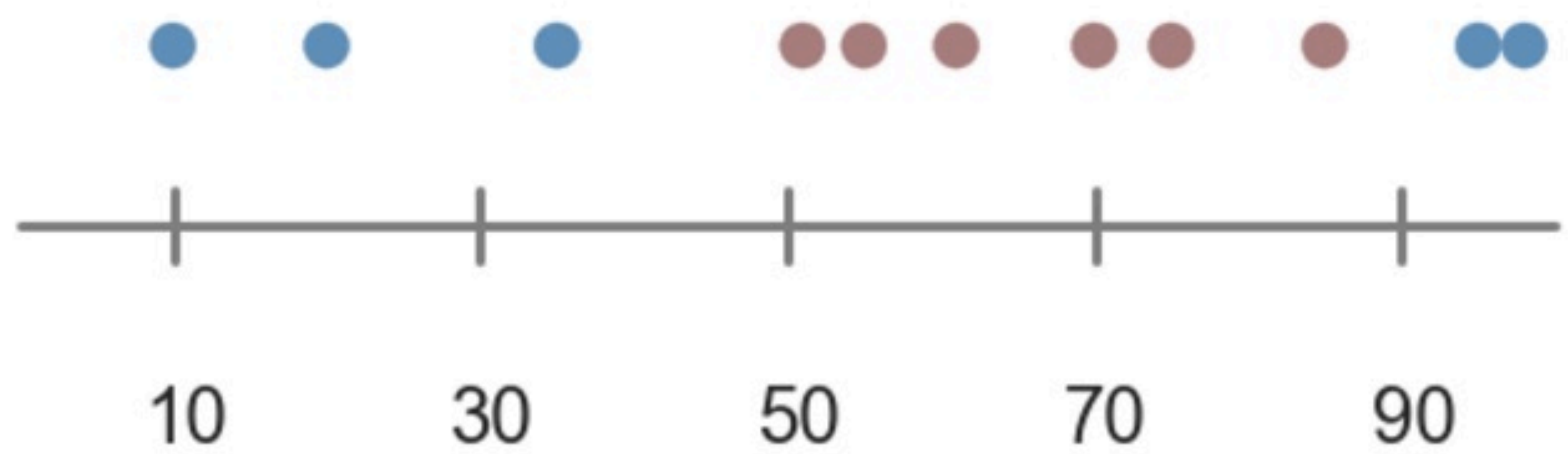
Def: **PAC Learnability** - A concept from class C is PAC-Learnable if there exists an algorithm $\mathcal{A}$ and a polynomial function f such that for any $\epsilon > 0$ and any $\delta > 0$, for all distributions $\mathcal{D}$ and any concept $c \in C$, it holds that

$$Pr_{S \sim \mathcal{D}^m} \left[R(h_S) \leq \epsilon \right] \geq 1 - \delta$$

for any sample size $m \geq f(1/\epsilon, 1/\delta, n, \text{size}(c))$:

- $R(h_S)$: The generalization error of h_S
- $1 - \epsilon$: The accuracy of h_s
- $1 - \delta$: The confidence the accuracy $1 - \epsilon$ is realized

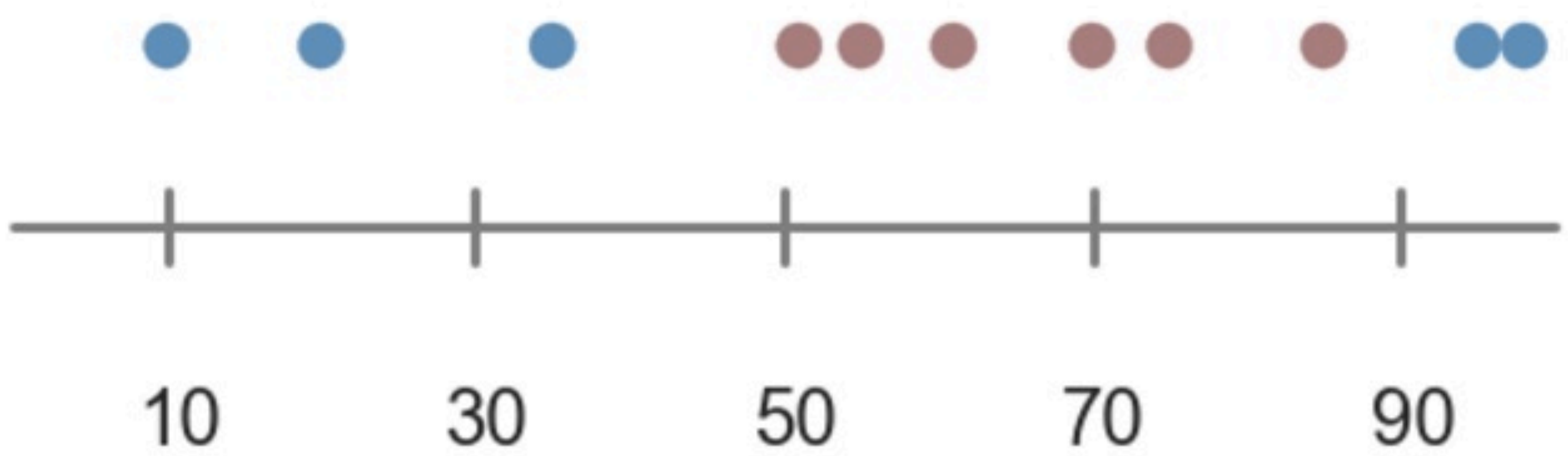
Frank the Alien Example



- Concept class $C =$ Intervals on Real Line
- Hypothesis class $H =$ Intervals on Real Line

Want to obtain bound on training examples needed to satisfy PAC

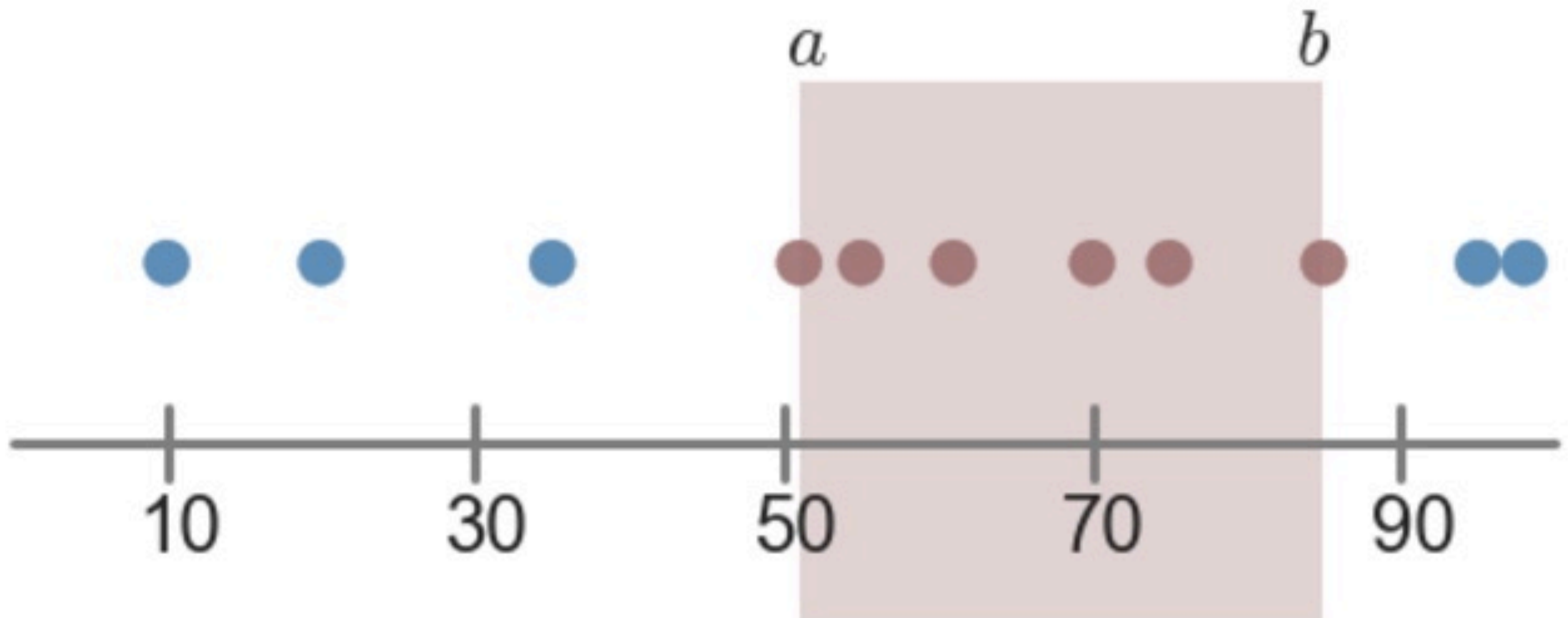
Frank the Alien Example



What is Algorithm $\mathcal{A}$?

Set hypothesis to smallest interval containing S

Frank the Alien Example

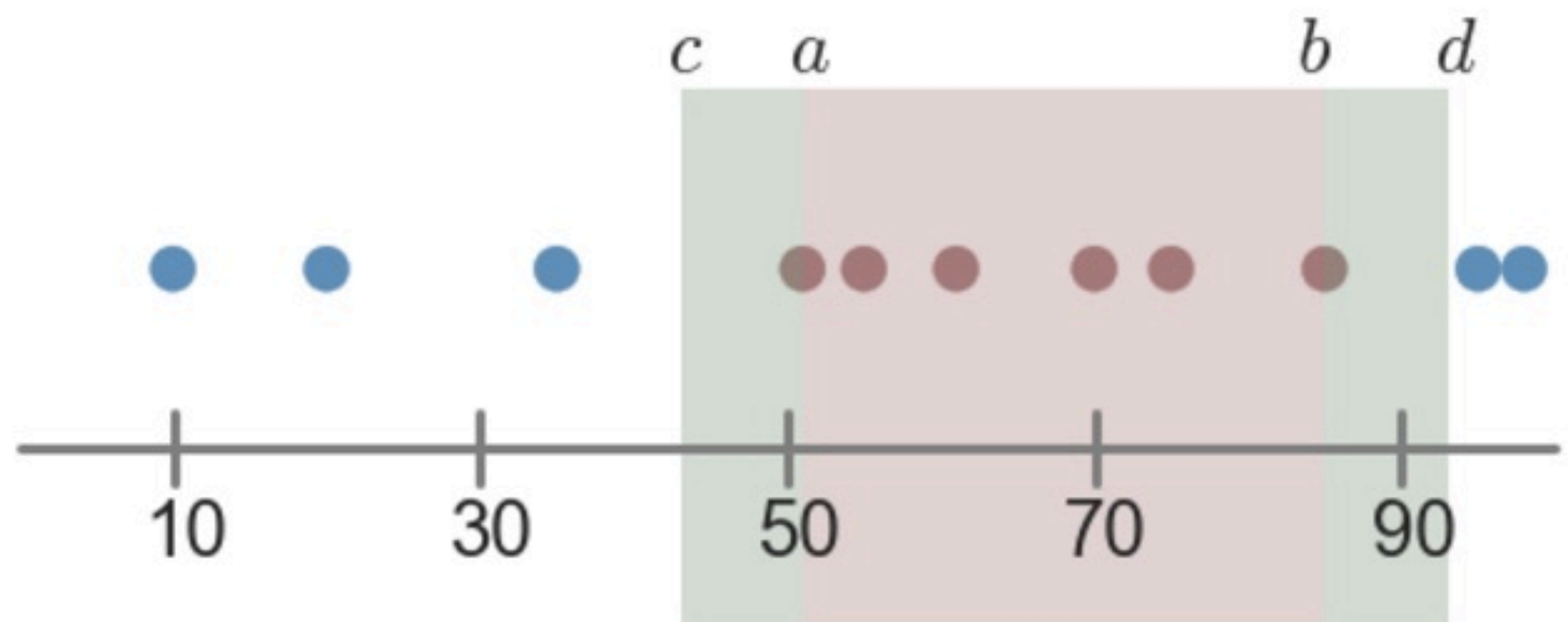


What is Algorithm $\mathcal{A}$?

Set hypothesis to smallest interval containing S: $h_s = [a, b]$

Errors happen if a positive point falls outside of $h_s = [a, b]$

Frank the Alien Example



What is Algorithm $\mathcal{A}$?

Set hypothesis to smallest interval containing S: $h_s = [a, b]$

Errors happen if a positive point false outside of $h_s = [a, b]$

Suppose true concept is $c = [c, d]$

Frank the Alien Example

Want to define relationship between ϵ, δ, and m such that

$$Pr_{S\sim\mathcal{D}^m}\left[R(h_S) \leq \epsilon\right] \geq 1-\delta$$

Easier to prove things about the contrapositive statement

$$\begin{aligned} Pr_{S\sim\mathcal{D}^m}\left[R(h_S) \leq \epsilon\right] \geq 1-\delta & \quad \Leftrightarrow \\ 1 - Pr_{S\sim\mathcal{D}^m}\left[R(h_S) > \epsilon\right] \geq 1-\delta & \quad \Leftrightarrow \\ -Pr_{S\sim\mathcal{D}^m}\left[R(h_S) > \epsilon\right] \geq -\delta & \quad \Leftrightarrow \\ Pr_{S\sim\mathcal{D}^m}\left[R(h_S) > \epsilon\right] < \delta & \end{aligned}$$

So instead we'll try to prove something about

$$Pr_{S\sim\mathcal{D}^m}\left[R(h_S) > \epsilon\right] < \delta$$

Frank the Alien Example

We want find a bound on the probability that the generalization error h_S is greater than ϵ.

This is the probability that despite the fact that the true concept was $c = [c, d]$, we didn't observe any points in $[c, a]$ or $[b, d]$.

Let ϵ be probability that a point $x \in \mathcal{D}$ lands in the missed intervals. WLOG assume probability of landing in either missed interval is $\epsilon/2$

Useful Fact 1: **Union Bound**

$$Pr[A \cup B] \leq Pr[A] + Pr[B]$$

Simplify $Pr_{S \sim \mathcal{D}^m}\left[R(h_S) > \epsilon\right]$ to statement that $Pr[h_s \text{ is bad}]$

Frank the Alien Example

$$
\begin{aligned}
Pr[h_s \text{ is bad}] &= Pr[\text{no } x_i \text{ in } [c,a] \text{ or } [b,d]\ ] \\
&\leq Pr[\text{no } x_i \text{ in } [c,a]\ ] + Pr[\text{no } x_i \text{ in } [b,d]\ ]
\end{aligned}
$$

$$
\begin{aligned}
Pr[\text{ no } x_i \text{ in } [c,a]\ ] &= Pr[\text{all } x_i \text{ not in } [c,a]\ ] \\
&= \prod_{i=1}^{m} \left(1 - \frac{\epsilon}{2}\right) = \left(1 - \frac{\epsilon}{2}\right)^m
\end{aligned}
$$

$$
\begin{aligned}
Pr[h_s \text{ is bad}] &\leq \left(1 - \tfrac{\epsilon}{2}\right)^m + \left(1 - \tfrac{\epsilon}{2}\right)^m \\
&= 2\left(1 - \tfrac{\epsilon}{2}\right)^m
\end{aligned}
$$

Useful Fact 2: For any $z \in \mathbb{R}$, $1 + z \leq e^z$

Frank the Alien Example

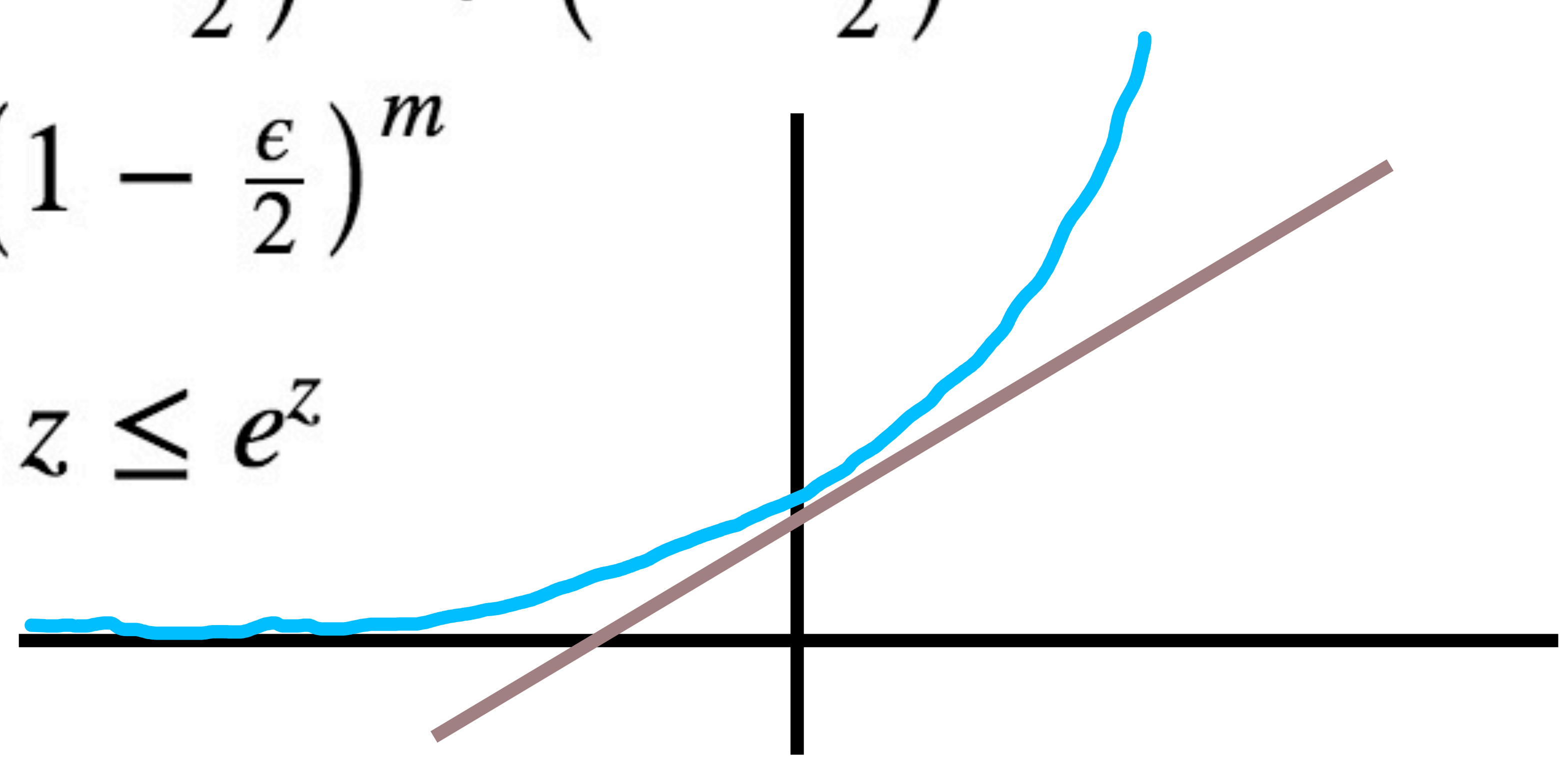
$$
\begin{aligned}
Pr[h_s \text{ is bad}] &= Pr[\text{no } x_i \text{ in } [c, a] \text{ or } [b, d]\ ] \\
&\leq Pr[\text{no } x_i \text{ in } [c, a]\ ] + Pr[\text{no } x_i \text{ in } [b, d]\ ]
\end{aligned}
$$

$$
Pr[\text{ no } x_i \text{ in } [c, a]\ ] = Pr[\text{all } x_i \text{ not in } [c, a]\ ]
$$

$$
= \prod_{i=1}^{m} \left(1 - \frac{\epsilon}{2}\right) = \left(1 - \frac{\epsilon}{2}\right)^m
$$

$$
\begin{aligned}
Pr[h_s \text{ is bad}] &\leq \left(1 - \tfrac{\epsilon}{2}\right)^m + \left(1 - \tfrac{\epsilon}{2}\right)^m \\
&= 2\left(1 - \tfrac{\epsilon}{2}\right)^m
\end{aligned}
$$

Useful Fact 2: For any $z \in \mathbb{R}$, $1 + z \leq e^z$



Frank the Alien Example

$$
\begin{aligned}
Pr[h_s \text{ is bad}] &= Pr[\text{no } x_i \text{ in } [c,a] \text{ or } [b,d]\ ] \\
&\leq Pr[\text{no } x_i \text{ in } [c,a]\ ] + Pr[\text{no } x_i \text{ in } [b,d]\ ]
\end{aligned}
$$

$$
\begin{aligned}
Pr[\text{ no } x_i \text{ in } [c,a]\ ] &= Pr[\text{all } x_i \text{ not in } [c,a]\ ] \\
&= \prod_{i=1}^{m}\left(1-\frac{\epsilon}{2}\right) = \left(1-\frac{\epsilon}{2}\right)^m
\end{aligned}
$$

$$
\begin{aligned}
Pr[h_s \text{ is bad}] &\leq \left(1-\tfrac{\epsilon}{2}\right)^m + \left(1-\tfrac{\epsilon}{2}\right)^m \\
&= 2\left(1-\tfrac{\epsilon}{2}\right)^m \\
&\leq 2e^{-\epsilon m/2}
\end{aligned}
$$

Frank the Alien Example

OK, we've bounded the probability that the generalization error for h_S is greater than ϵ.

$$Pr_{S \sim \mathcal{D}^m} [R(h_S) > \epsilon] \leq 2e^{-\epsilon m/2}$$

Now, want to choose m so that this is less than δ. For a fixed $\delta > 0$

$$2e^{-\epsilon m/2} < \delta \;\; \Leftrightarrow \;\; \frac{-\epsilon m}{2} < \ln \frac{\delta}{2} \;\; \Leftrightarrow \;\; m > \frac{2}{\epsilon} \ln \frac{2}{\delta}$$

Punchline: For any choice of $\epsilon > 0$ and $\delta > 0$, hypothesis h_S is probably approximately correct if

$$m > \frac{2}{\epsilon} \ln \frac{2}{\delta}$$

Frank the Alien Example

Punchline: For any choice of $\epsilon > 0$ and $\delta > 0$, hypothesis h_S is probably approximately correct if

$$m > \frac{2}{\epsilon} \ln \frac{2}{\delta}$$

Example: Want 99% accuracy ($\epsilon = 0.01$) with 99% confidence ($\delta = 0.01$) then need

$$m > \frac{2}{.01} \ln \frac{2}{.01} \approx 1060 \text{ training examples}$$

Important: The lower bound on m is bounded above by a polynomial in $1/\epsilon$ and $1/\delta$, thus this problem is PAC Learnable.

General Case, Finite Hypothesis Class

OK, so we saw an example proving PAC Learnability for a specific problem with specific hypothesis and specific algorithm.

Can we be more general than this?

The answer is Yes!

- Today, the case when H is finite
- Next Time, the case when H is infinite

Distinction:

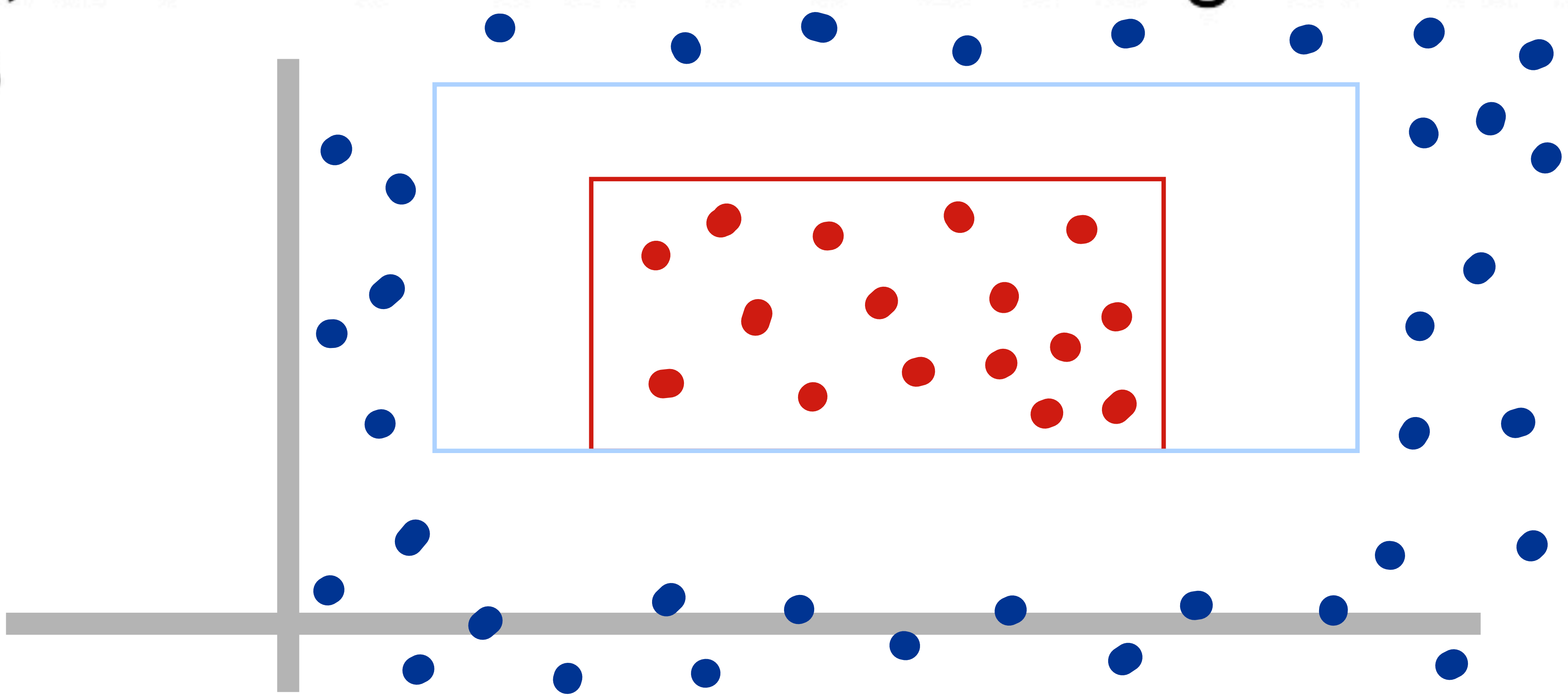
- H is finite and $c \in H$
- H is finite and $c \notin H$

General Case, Finite Consistent Hypothesis Class

Def: We say that a hypothesis h is **consistent** if it admits no error on the training sample S, or in other words, $\hat{R}(h) = 0$

Note: If we suppose that our algorithm $\mathcal{A}$ can always find a consistent hypothesis, then we suppose that $c \in H$

Example: Suppose c is the interior of an axis-aligned rectangle with integer vertices, and H is the set of all axis-aligned rectangles with integer vertices

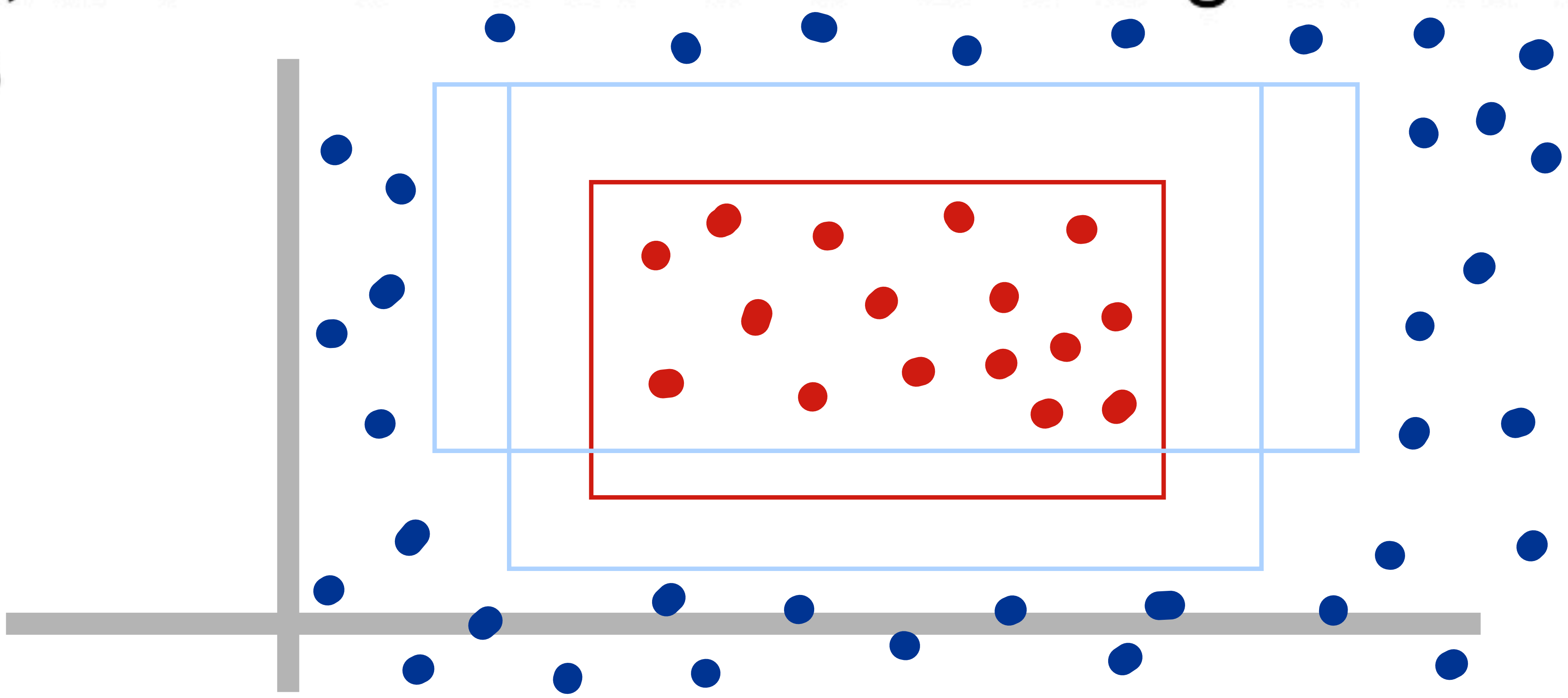


General Case, Finite Consistent Hypothesis Class

Def: We say that a hypothesis h is **consistent** if it admits no error on the training sample S, or in other words, $\hat{R}(h) = 0$

Note: If we suppose that our algorithm $\mathcal{A}$ can always find a consistent hypothesis, then we suppose that $c \in H$

Example: Suppose c is the interior of an axis-aligned rectangle with integer vertices, and H is the set of all axis-aligned rectangles with integer vertices



Note: There may be *multiple* consistent hypotheses in H

General Case, Finite Consistent Hypothesis Class

Suppose our algorithm $\mathcal{A}$ can find a consistent hypothesis

Theorem: Let H be a finite set of functions mapping $\mathcal{X}$ to $\mathcal{Y}$. Let $\mathcal{A}$ be an algorithm that for an i.i.d. sample S returns a consistent hypothesis, then for any $\epsilon, \ \delta > 0$, the concept c is PAC Learnable with

$$m \geq \frac{1}{\epsilon}\left(\ln|H| + \ln\frac{1}{\delta}\right)$$

Proof: Like before, we want to bound the probability that some $h \in H$ is consistent and has generalization error more than ϵ

General Case, Finite Consistent Hypothesis Class

Proof: We want to bound the probability that some $h \in H$ is consistent and has generalization error more than ϵ

$$Pr[\ \exists\, h \in H \text{ s.t. } \hat{R}(h) = 0 \text{ and } R(h) > \epsilon\ ] \quad =$$

$$Pr[(h_1 \in H \text{ and } \hat{R}(h_1) = 0 \text{ and } R(h_1) > \epsilon) \text{ or } \ldots$$

$$\ldots \text{ or } (h_{|H|} \in H \text{ and } \hat{R}(h_1) = 0 \text{ and } R(h_1) > \epsilon)]$$

Probability of at least one of at least one of all consistent $h \in H$ having generalization error greater than ϵ

General Case, Finite Consistent Hypothesis Class

Proof: We want to bound the probability that some $h \in H$ is consistent and has generalization error more than ϵ

$$
\begin{aligned}
Pr[\ \exists\, h \in H \text{ s.t. } \hat{R}(h) = 0 \text{ and } R(h) > \epsilon\ ] \quad &= \\
Pr[(h_1 \in H \text{ and } \hat{R}(h_1) = 0 \text{ and } R(h_1) > \epsilon) \text{ or} \ldots \\
Pr[(h_2 \in H \text{ and } \hat{R}(h_2) = 0 \text{ and } R(h_2) > \epsilon) \text{ or} \ldots \\
\ldots \text{ or } (h_{|H|} \in H \text{ and } \hat{R}(h_1) = 0 \text{ and } R(h_1) > \epsilon)] \quad &\leq \\
\textstyle\sum_h Pr[\ \hat{R}(h) = 0 \text{ and } R(h) > \epsilon]
\end{aligned}
$$

Using the Union Bound

General Case, Finite Consistent Hypothesis Class

Proof: We want to bound the probability that some $h \in H$ is consistent and has generalization error more than ϵ

$$
\begin{aligned}
Pr[\ \exists\, h \in H \text{ s.t. } \hat{R}(h) = 0 \text{ and } R(h) > \epsilon\ ] \quad &= \\
Pr[(h_1 \in H \text{ and } \hat{R}(h_1) = 0 \text{ and } R(h_1) > \epsilon) \text{ or} \ldots \\
Pr[(h_2 \in H \text{ and } \hat{R}(h_2) = 0 \text{ and } R(h_2) > \epsilon) \text{ or} \ldots \\
\ldots \text{ or } (h_{|H|} \in H \text{ and } \hat{R}(h_1) = 0 \text{ and } R(h_1) > \epsilon)] \quad &\leq \\
\textstyle\sum_h Pr[\ \hat{R}(h) = 0 \text{ and } R(h) > \epsilon] \quad &= \\
\textstyle\sum_h Pr[\ \hat{R}(h) = 0 \mid R(h) > \epsilon]\ Pr[\ R(h) > \epsilon\ ] \quad &\leq \\
\textstyle\sum_h Pr[\ \hat{R}(h) = 0 \mid R(h) > \epsilon]
\end{aligned}
$$

Using the product rule and fact that $Pr[\ R(h) > \epsilon\ ] \leq 1$

General Case, Finite Consistent Hypothesis Class

The generalization error is greater than ϵ, so we bound the probability that we found **no** bad points in training set for a single hypothesis h as

$$Pr[\ \hat{R}(h) = 0 \mid R(h) > \epsilon] \leq (1 - \epsilon)^m$$

General Case, Finite Consistent Hypothesis Class

The generalization error is greater than ϵ, so we bound the probability that we found **no** bad points in training set for a single hypothesis h as

$$Pr[\ \hat{R}(h) = 0 \mid R(h) > \epsilon] \leq (1 - \epsilon)^m$$

But this must be true for all of the hypotheses in H, so

$$Pr[\ \exists\ h \in H \text{ s.t. } \hat{R}(h) = 0 \text{ and } R(h) > \epsilon\ ] \leq |H|(1 - \epsilon)^m$$

General Case, Finite Consistent Hypothesis Class

The generalization error is greater than ϵ, so we bound the probability that we found **no** bad points in training set for a single hypothesis h as

$$Pr[\ \hat{R}(h) = 0 \mid R(h) > \epsilon] \leq (1 - \epsilon)^m$$

But this must be true for all of the hypotheses in H, so

$$Pr[\ \exists\ h \in H \text{ s.t. } \hat{R}(h) = 0 \text{ and } R(h) > \epsilon\ ] \leq |H|(1 - \epsilon)^m$$

Using our exponential trick again

$$Pr[\ \exists\ h \in H \text{ s.t. } \hat{R}(h) = 0 \text{ and } R(h) > \epsilon\ ] \leq |H|e^{-m\epsilon}$$

General Case, Finite Consistent Hypothesis Class

Have our bound on $Pr_{S \sim \mathcal{D}^m}[R(h_S) > \epsilon]$. Now for any $\delta > 0$

$$
\begin{aligned}
Pr_{S \sim \mathcal{D}^m}[R(h_S) > \epsilon] \leq |H| e^{-m\epsilon} < \delta \quad &\Leftrightarrow \\
\ln |H| - m\epsilon < \ln \delta \quad &\Leftrightarrow \\
\ln |H| - \ln \delta < m\epsilon \quad &\Leftrightarrow \\
\ln |H| + \ln \tfrac{1}{\delta} < m\epsilon &
\end{aligned}
$$

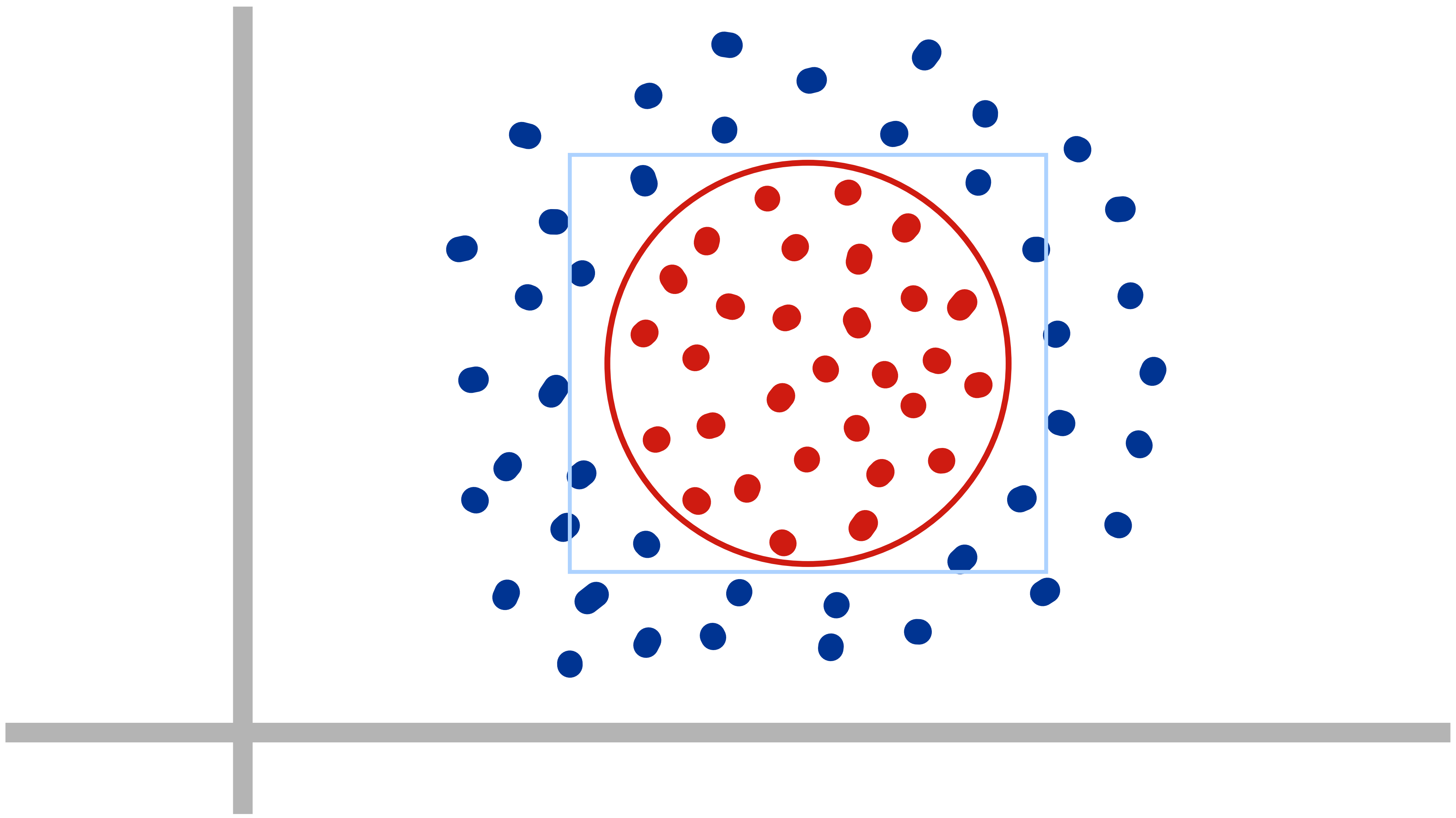
And solving for m gives us the bound we're after

$$
m \geq \frac{1}{\epsilon}\left(\ln |H| + \ln \frac{1}{\delta} \right)
$$

Finite Inconsistent Hypothesis Class

The more common case occurs when the true concept c does not occur in our hypothesis class H and so we can find no hypothesis h that admits a zero training error.

Example: Hypothesis class H is axis aligned rectangles, but true concept is a circle



Finite Inconsistent Hypothesis Class

The more common case occurs when the true concept c does not occur in our hypothesis class H and so we can find no hypothesis h that admits a zero training error.

Example: Hypothesis class H is axis aligned rectangles, but true concept is a circle

To handle this case we have to borrow a theorem of analysis

Theorem: Hoeffding's Inequality: - Fix $\epsilon > 0$ and let S denote i.i.d. same of size m. Then, for any hypothesis $h : \mathcal{X} \rightarrow \{0, 1\}$, the following holds

$$Pr_{S \sim \mathcal{D}^m}[\ |\hat{R}(h) - R(h)| > \epsilon\ ] \leq 2\exp[-2m\epsilon^2]$$

Finite Inconsistent Hypothesis Class

Setting $\delta = 2\exp[-2m\epsilon^2]$, solving for $\epsilon = \epsilon(\delta)$ and plugging back in yields, for a single hypothesis h

$$R(h) \leq \hat{R}(h) + \sqrt{\frac{\ln(2/\delta)}{2m}}$$

But this is just for a single h. We have

Theorem: Let H be a finite hypothesis set. Then, for any $\delta > 0$, with probability at least $1 - \delta$, we have

$$\forall h \in H, \;\; R(h) \leq \hat{R}(h) + \sqrt{\frac{\ln|H| + \ln(2/\delta)}{2m}}$$

Finite Inconsistent Hypothesis Class

Proof: (Very similar to before). Let $h_1, \ldots, h_{|H|}$ be the elements of H. Then

$$\begin{aligned}
Pr[\ \exists h \in H \text{ s.t. } |\hat{R}(h) - R(h)| > \epsilon\ ] &= \\
Pr\left[\bigvee_{h_i \in H} |\hat{R}(h_i) - R(h_i)| > \epsilon\right] &\leq \\
\sum_{h_i \in H} Pr\left[|\hat{R}(h_i) - R(h_i)| > \epsilon\right] &\leq \\
2|H| \exp[-2m\epsilon^2]
\end{aligned}$$

Finite Inconsistent Hypothesis Class

Proof:

If we fix $\epsilon > 0$ and set $\delta = 2|H| \exp[-2m\epsilon^2]$, we can choose m large enough such that with confidence $1 - \delta$

$$\forall h \in H \quad |\hat{R}(h) - R(h)| \leq \epsilon \leq \sqrt{\frac{\ln|H| + \ln(2/\delta)}{2m}}$$

which implies that

$$\forall h \in H \quad R(h) \leq \hat{R}(h) + \sqrt{\frac{\ln|H| + \ln(2/\delta)}{2m}}$$

Finite Inconsistent Hypothesis Class

$$\forall h \in H \quad R(h) \leq \hat{R}(h) + \sqrt{\frac{\ln |H| + \ln(2/\delta)}{2m}}$$

Could talk about this one for an hour

- The larger m is, better training error predicts gen. error

What about the case that we consider making H more complex?

- Training error would go down
- Bound term would go up ...

Finite Inconsistent Hypothesis Class

$$\forall h \in H \quad R(h) \leq \hat{R}(h) + \sqrt{\frac{\ln |H| + \ln(2/\delta)}{2m}}$$

Could talk about this one for an hour

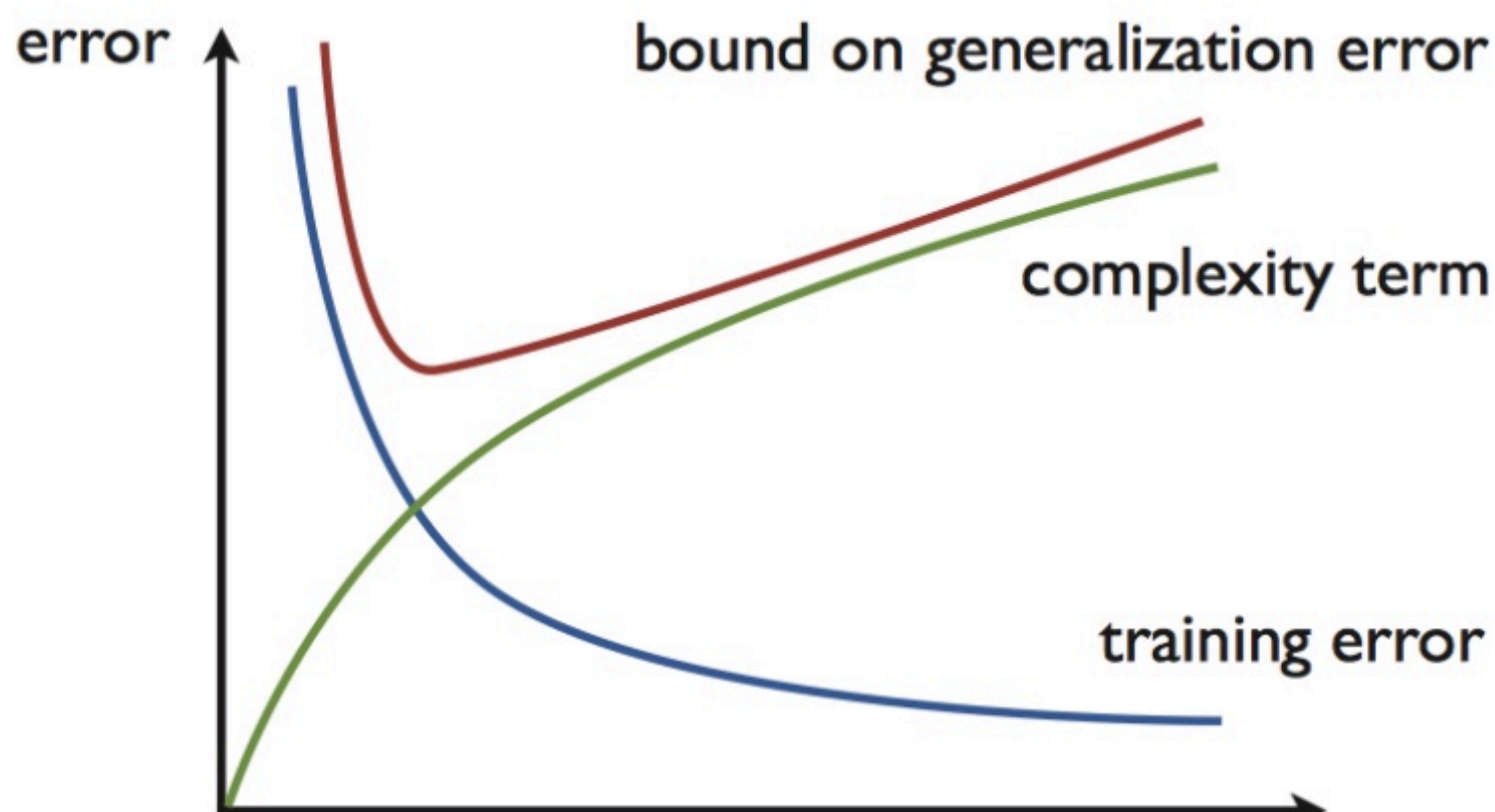
- The larger m is, better training error predicts gen. error

What about the case that we consider making H more complex?

- Training error would go down
- Bound term would go up ...
- **Bias-Variance Trade-Off!**

Finite Inconsistent Hypothesis Class

$$\forall h \in H \quad R(h) \le \hat{R}(h) + \sqrt{\frac{\ln |H| + \ln(2/\delta)}{2m}}$$



PAC Learnability

OK, that was a lot of Math!

I Expect You to ...

- Know the bounds we've proved
- Know which bounds apply to which situations
- Know how to apply a particular bound to a problem

I Do Not Expect You to ...

- Know the details of the proofs
- Be able to prove PAC bounds yourself

PAC Learnability

In Class:

- **Your Questions!**
- Talk about how to use these bounds in practice
- Work through some simple examples

Next Time:

- See what happens when H is infinite
- Introduce the important idea of the VC Dimension

Acknowledgments

Many of these slides were adopted from Jordan Boyd-Graber

Some of the figures in this presentation were adopted from *Foundations of Machine Learning* by Mohri, et. al.

In Class

In Class

In Class