10-701: Introduction to Machine Learning Lecture 21 – Learning Theory II

Pradeep Ravikumar & Geoff Gordon

Spring 2026

Front Matter

- Announcements
  - Project check-in is due next Wednesday, April 8.
  - You should complete a second 15-30 minute meeting with your project mentor before that date (April 8)
    - It is recommended to bring a (hopefully complete) draft of check-in to that meeting.

RECAP: Labelings

- Given some finite set of data points $S = \left(\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(M)}\right)$ and some hypothesis $h \in \mathcal{H}$, applying h to each point in S results in a **labelling**
  - $\left(h\left(\boldsymbol{x}^{(1)}\right), \ldots, h\left(\boldsymbol{x}^{(M)}\right)\right)$ is a vector of M +1's and -1's
- Insight: given $S = \left(\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(M)}\right)$, each hypothesis in $\mathcal{H}$ induces a labelling *but not necessarily a unique labelling*
  - The set of labellings induced by $\mathcal{H}$ on S is

$$\mathcal{H}(S) = \left\{\left(h\left(\boldsymbol{x}^{(1)}\right), \ldots, h\left(\boldsymbol{x}^{(M)}\right)\right) \,\middle|\, h \in \mathcal{H}\right\}$$

RECAP: Growth Function

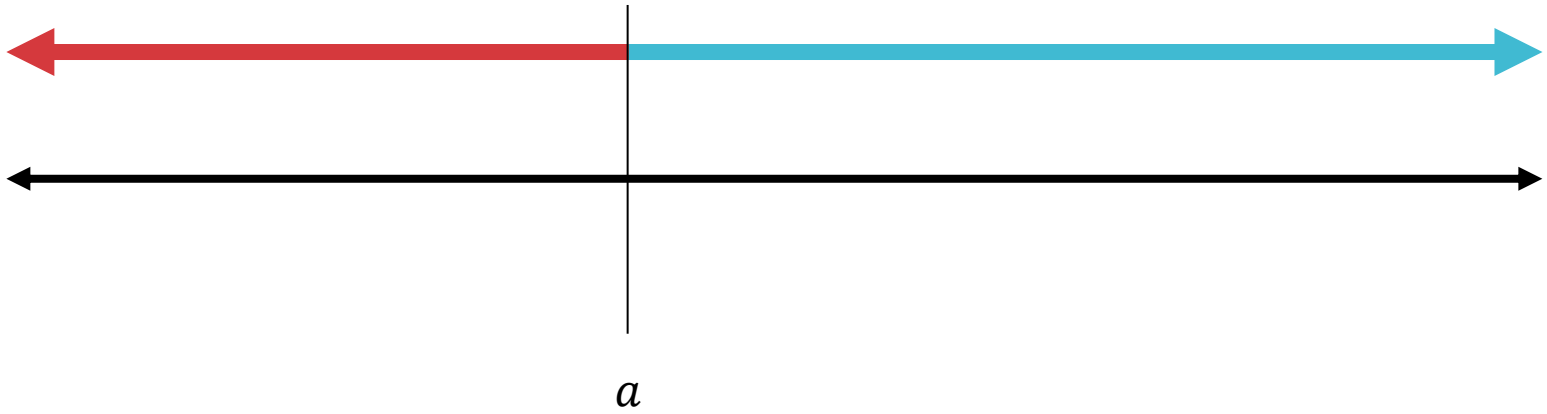
- The **growth function** of $\mathcal{H}$ is the maximum number of distinct labelings $\mathcal{H}$ can induce on ***any*** set of M data points:

$$g_{\mathcal{H}}(M) = \max_{S\,:\,|S|=M} |\mathcal{H}(S)|$$

- $g_{\mathcal{H}}(M) \leq 2^M \;\forall\; \mathcal{H}$ and M
- $\mathcal{H}$ **shatters** S if $|\mathcal{H}(S)| = 2^M$
- If $\exists\; S$ s.t. $|S| = M$ and $\mathcal{H}$ shatters S, then $g_{\mathcal{H}}(M) = 2^M$

VC-Dimension: Example

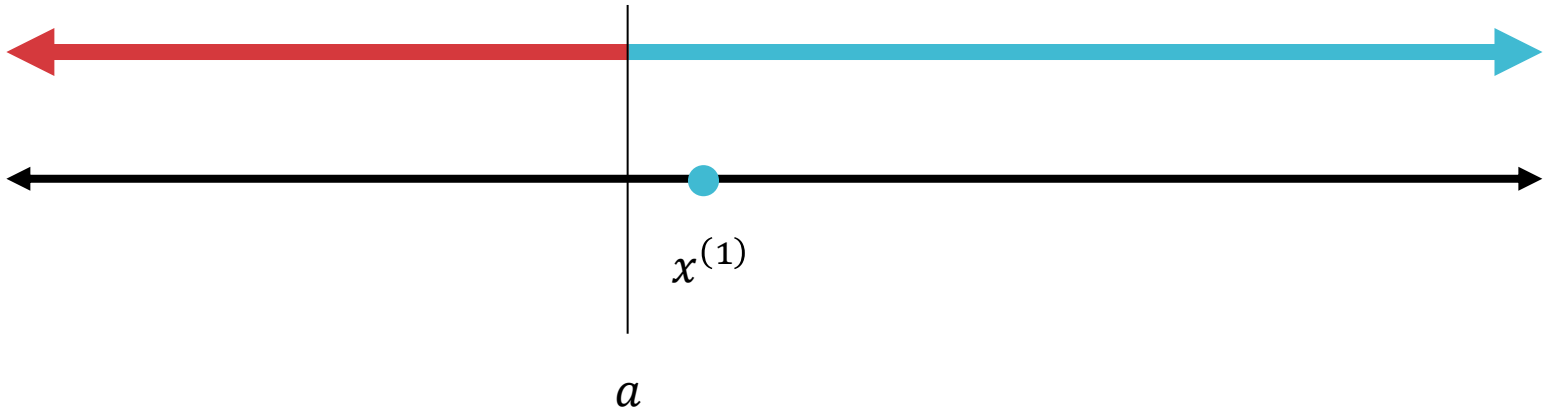
- $x^{(m)} \in \mathbb{R}$ and $\mathcal{H} =$ all 1-dimensional positive rays, i.e., all hypotheses of the form $h(x; a) = \text{sign}(x - a)$



- What is $d_{VC}(\mathcal{H})$?

VC-Dimension: Example

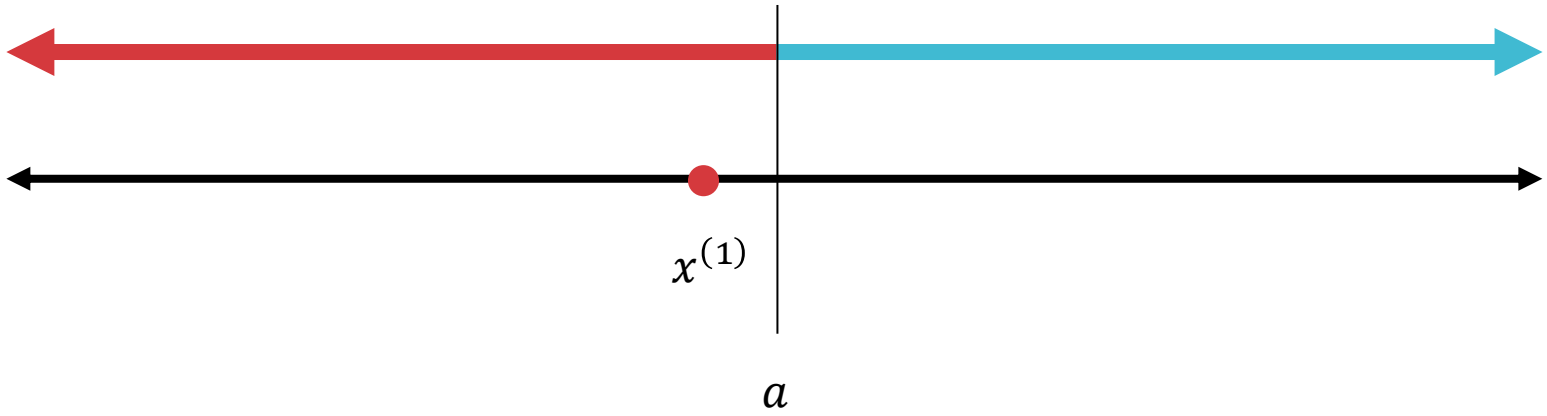
- $x^{(m)} \in \mathbb{R}$ and $\mathcal{H} =$ all 1-dimensional positive rays, i.e., all hypotheses of the form $h(x; a) = \text{sign}(x - a)$



- What is $d_{VC}(\mathcal{H})$?

VC-Dimension: Example

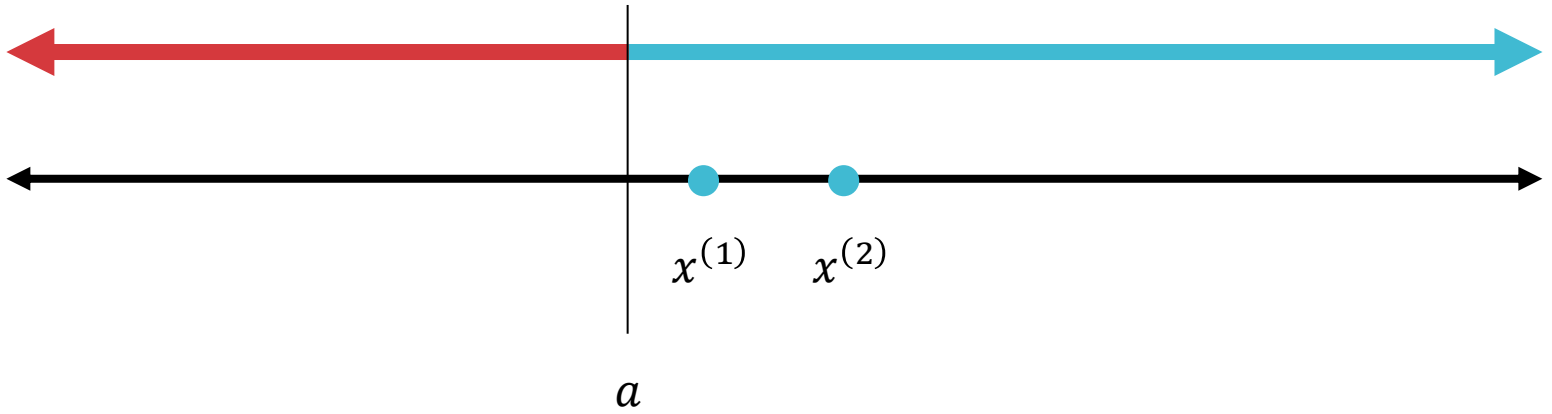
- $x^{(m)} \in \mathbb{R}$ and $\mathcal{H} =$ all 1-dimensional positive rays, i.e., all hypotheses of the form $h(x; a) = \text{sign}(x - a)$



- What is $d_{VC}(\mathcal{H})$?

VC-Dimension: Example

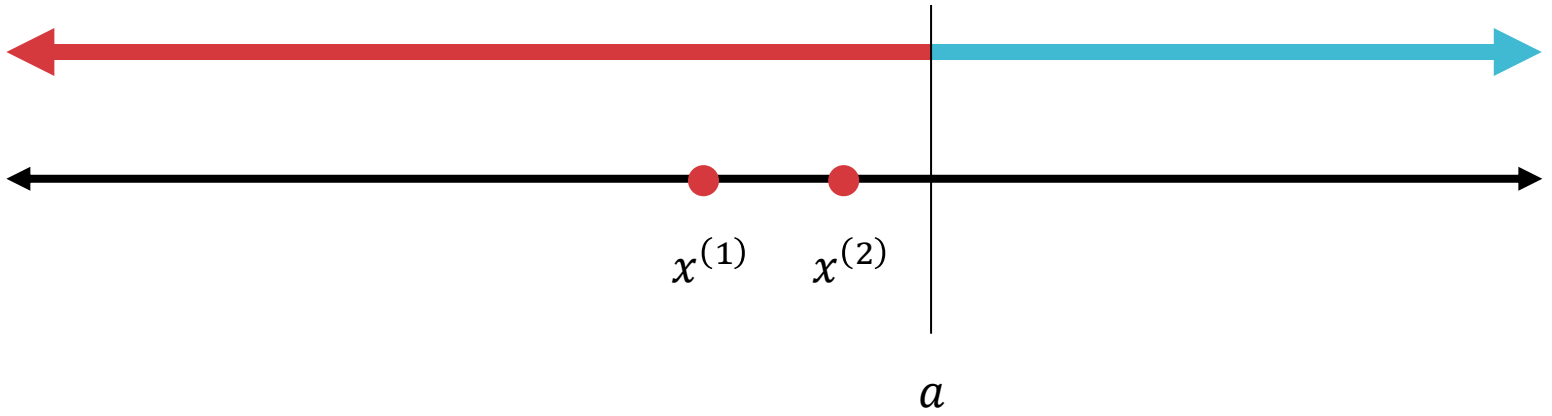
- $x^{(m)} \in \mathbb{R}$ and $\mathcal{H}$ = all 1-dimensional positive rays, i.e., all hypotheses of the form $h(x; a) = \text{sign}(x - a)$



- What is $d_{VC}(\mathcal{H})$?

VC-Dimension: Example

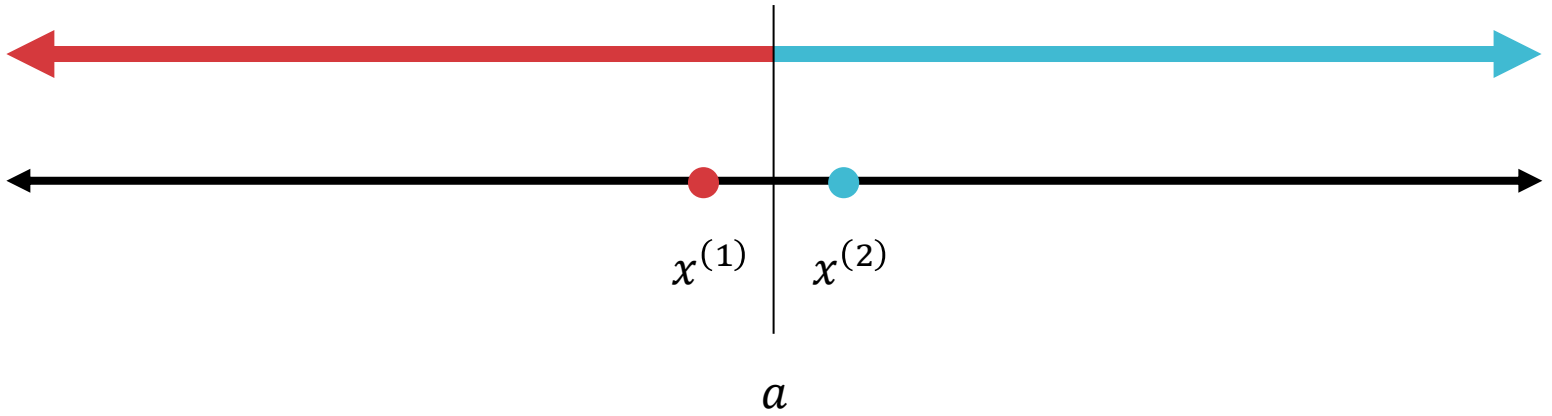
- $x^{(m)} \in \mathbb{R}$ and $\mathcal{H}$ = all 1-dimensional positive rays, i.e., all hypotheses of the form $h(x; a) = \text{sign}(x - a)$



- What is $d_{VC}(\mathcal{H})$?

VC-Dimension: Example

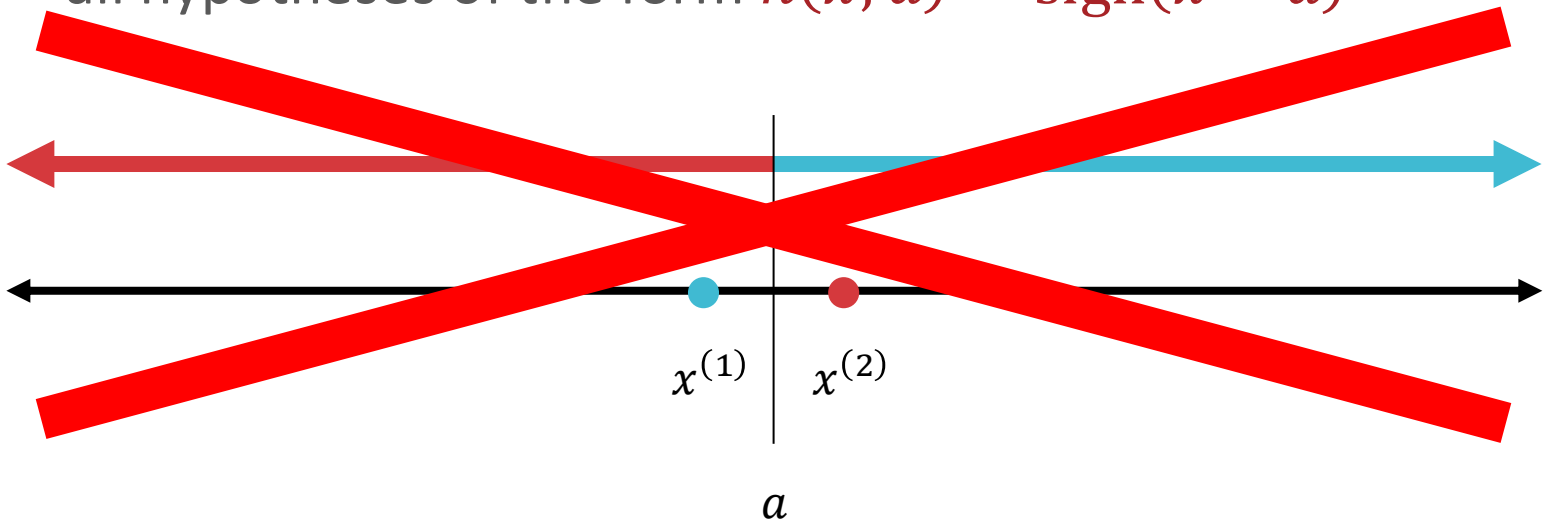
- $x^{(m)} \in \mathbb{R}$ and $\mathcal{H} =$ all 1-dimensional positive rays, i.e., all hypotheses of the form $h(x; a) = \text{sign}(x - a)$



- What is $d_{VC}(\mathcal{H})$?

VC-Dimension: Example

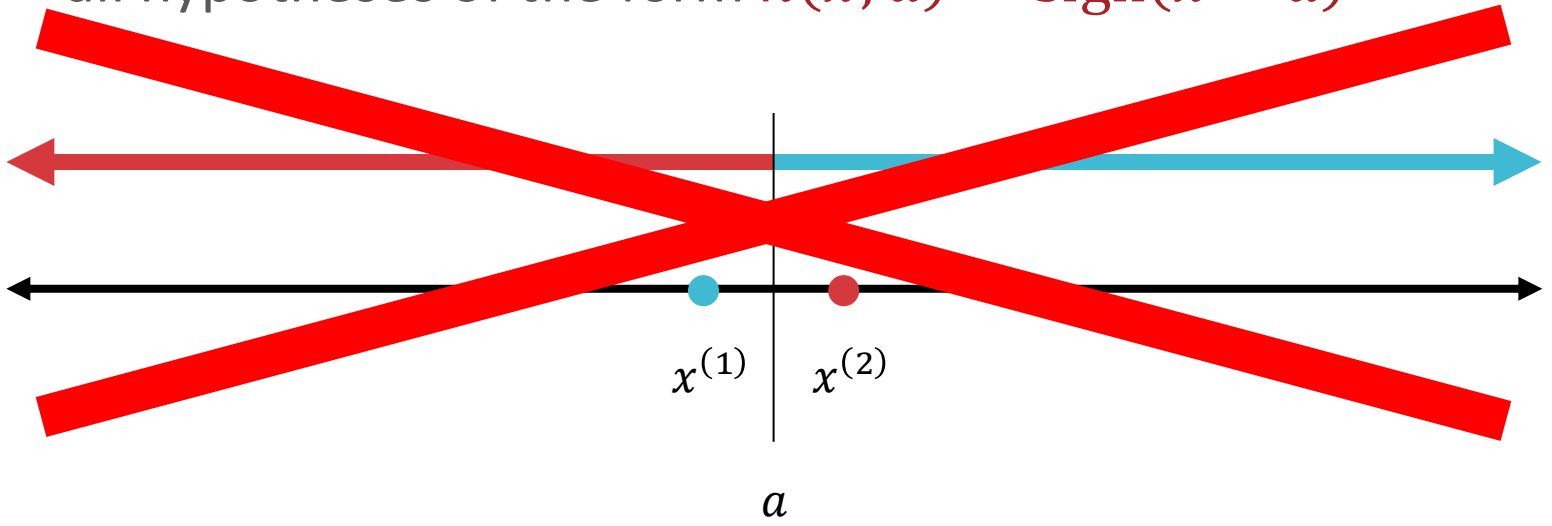
- $x^{(m)} \in \mathbb{R}$ and $\mathcal{H}$ = all 1-dimensional positive rays, i.e., all hypotheses of the form $h(x; a) = \text{sign}(x - a)$



- What is $d_{VC}(\mathcal{H})$?

VC-Dimension: Example

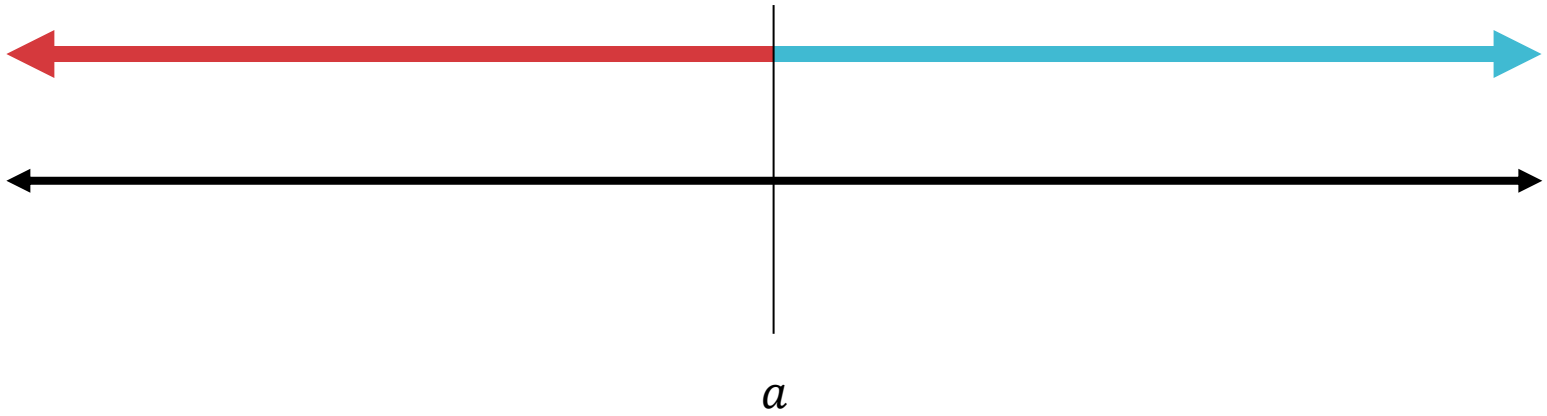
- $x^{(m)} \in \mathbb{R}$ and $\mathcal{H}$ = all 1-dimensional positive rays, i.e., all hypotheses of the form $h(x; a) = \text{sign}(x - a)$



- $d_{VC}(\mathcal{H}) = 1$

VC-Dimension: Example

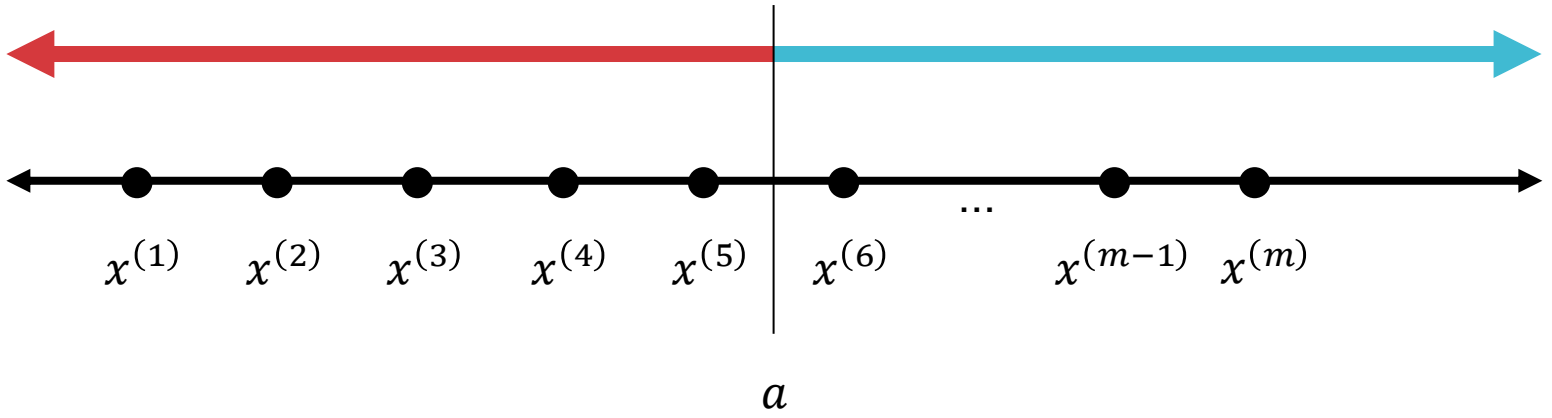
- $x^{(i)} \in \mathbb{R}$ and $\mathcal{H} =$ all 1-dimensional positive rays, i.e., all hypotheses of the form $h(x; a) = \text{sign}(x - a)$



- What is $g_{\mathcal{H}}(m)$?

VC-Dimension: Example

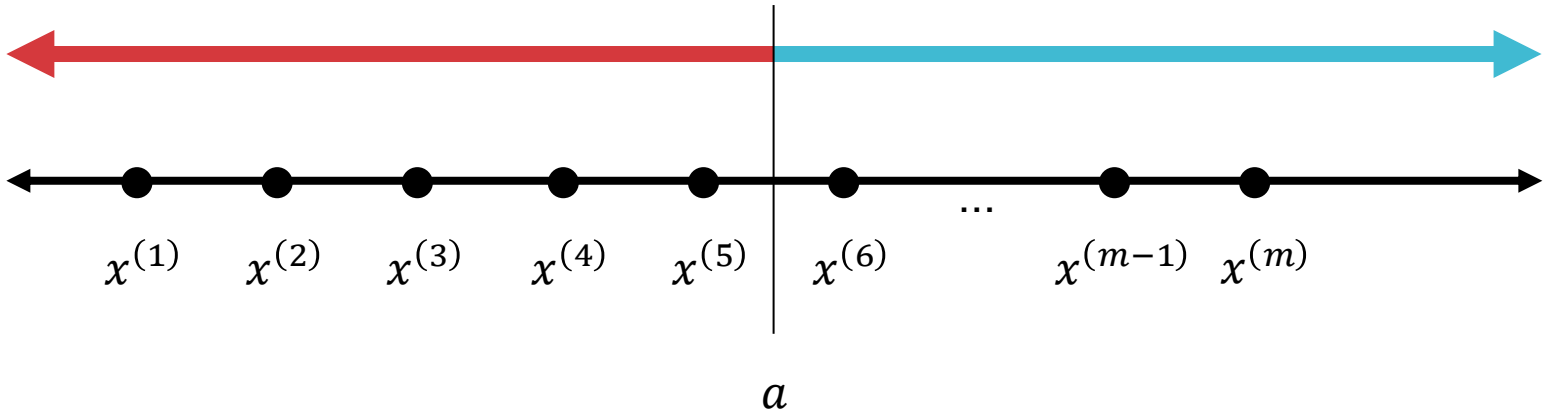
- $x^{(i)} \in \mathbb{R}$ and $\mathcal{H} =$ all 1-dimensional positive rays, i.e., all hypotheses of the form $h(x; a) = \text{sign}(x - a)$



- What is $g_{\mathcal{H}}(m)$?

VC-Dimension: Example

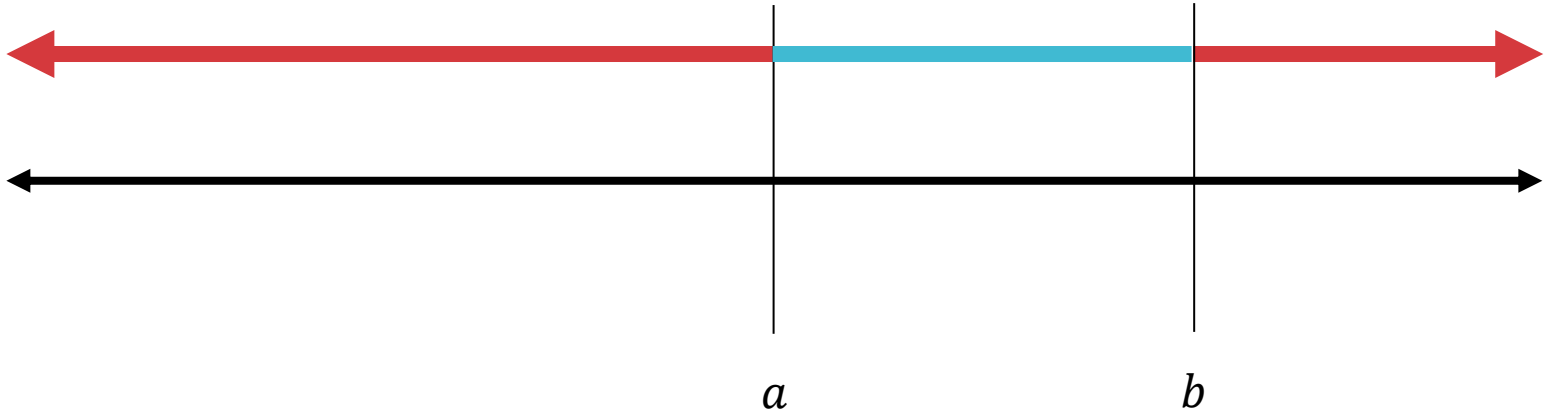
- $x^{(m)} \in \mathbb{R}$ and $\mathcal{H} =$ all 1-dimensional positive rays, i.e., all hypotheses of the form $h(x; a) = \text{sign}(x - a)$



- $g_{\mathcal{H}}(m) = m + 1 = O(m^1)$

VC-Dimension: Example

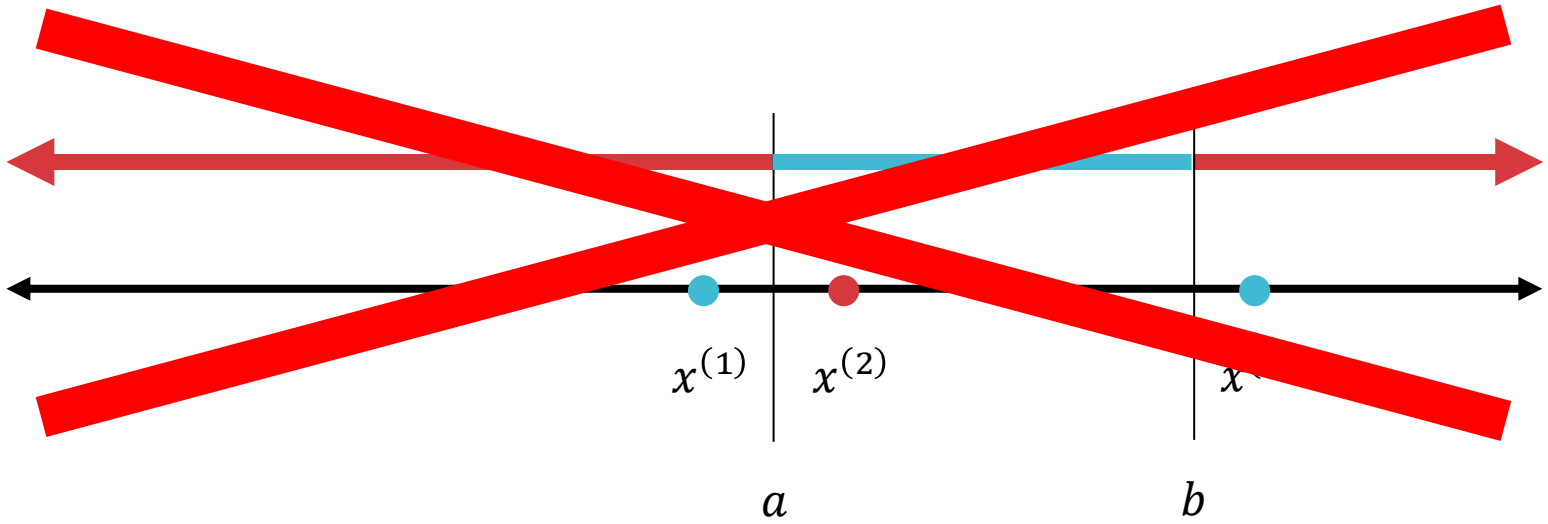
- $x^{(m)} \in \mathbb{R}$ and $\mathcal{H} =$ all 1-dimensional positive intervals



- What are $d_{VC}(\mathcal{H})$ and $g_{\mathcal{H}}(m)$?

VC-Dimension: Example

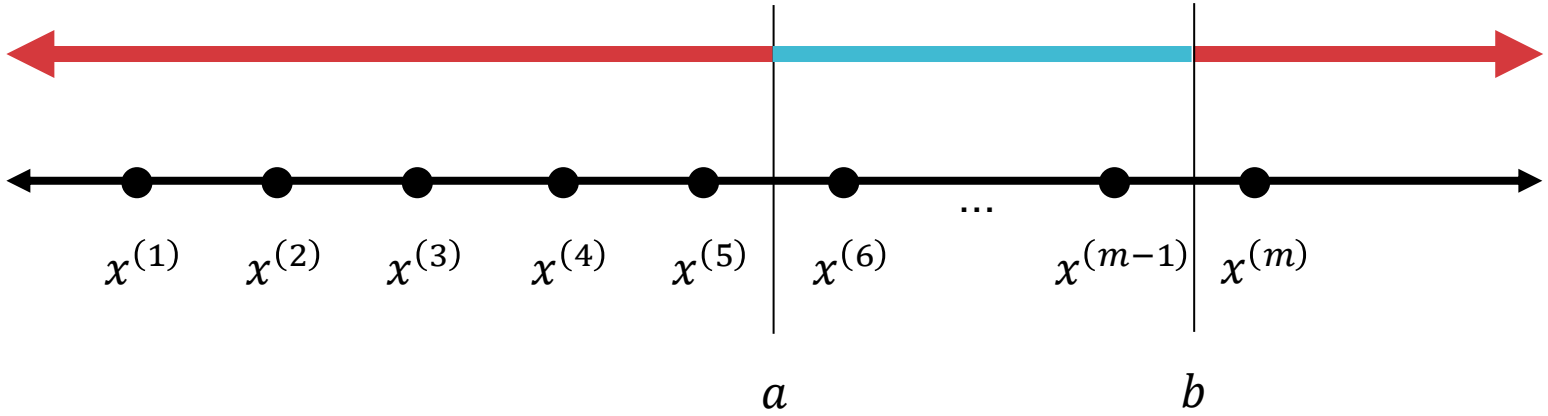
- $x^{(m)} \in \mathbb{R}$ and $\mathcal{H}$ = all 1-dimensional positive intervals



- What are $d_{VC}(\mathcal{H})$ and $g_{\mathcal{H}}(m)$?

VC-Dimension: Example

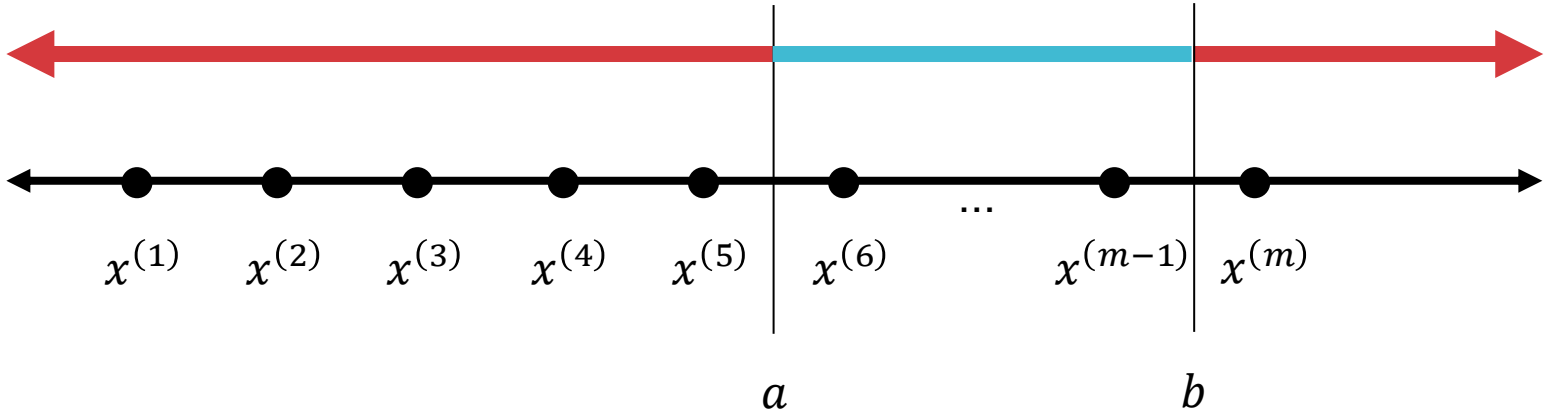
- $x^{(m)} \in \mathbb{R}$ and $\mathcal{H} =$ all 1-dimensional positive intervals



- What are $d_{VC}(\mathcal{H})$ and $g_{\mathcal{H}}(m)$?

VC-Dimension: Example

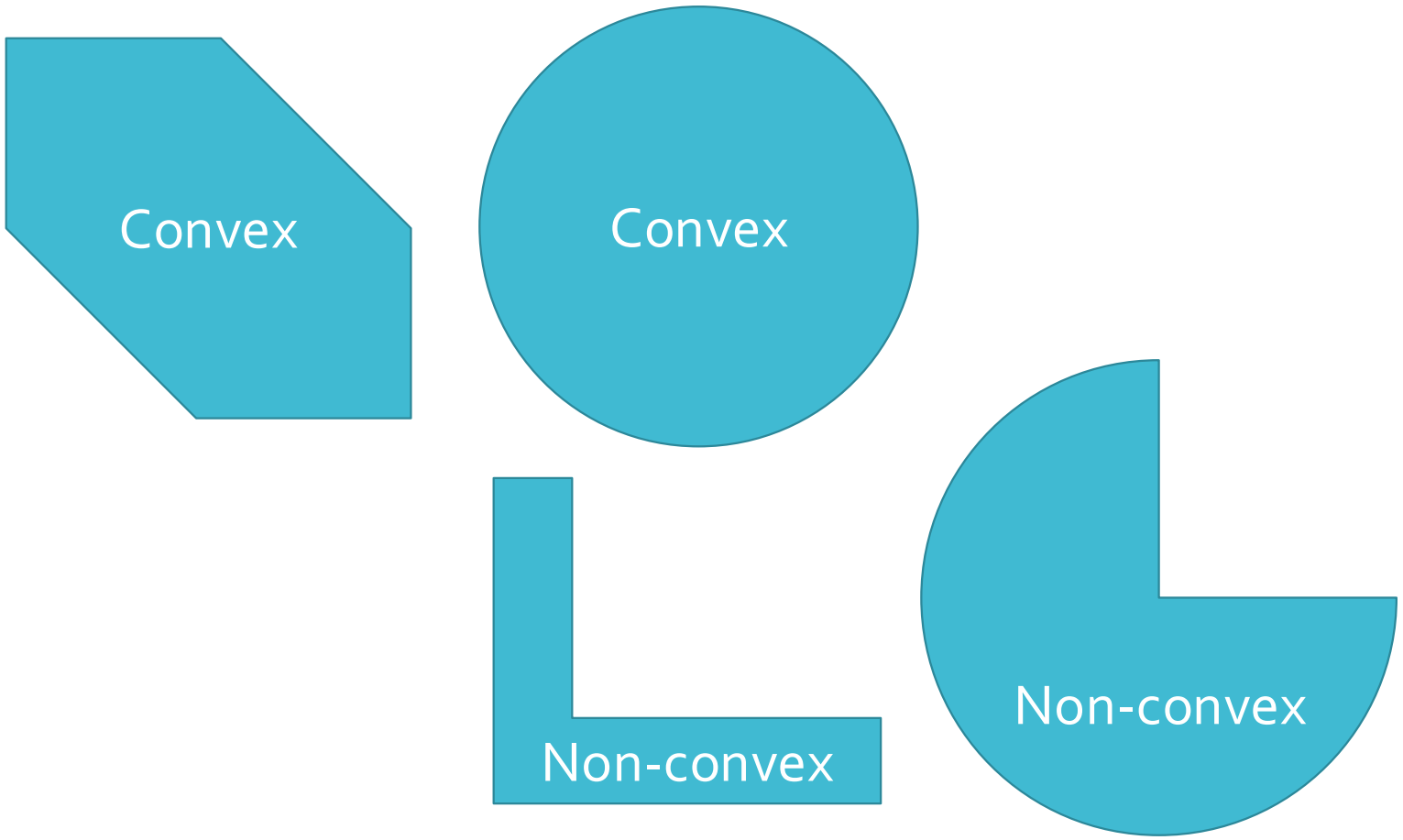
- $x^{(m)} \in \mathbb{R}$ and $\mathcal{H} =$ all 1-dimensional positive intervals



- $d_{VC}(\mathcal{H}) = 2$ and $g_{\mathcal{H}}(m) = \binom{m+1}{2} + 1 = O(m^2)$

Growth Function: Example

- $x^{(m)} \in \mathbb{R}^2$ and $\mathcal{H}$ = all 2-dimensional positive convex sets

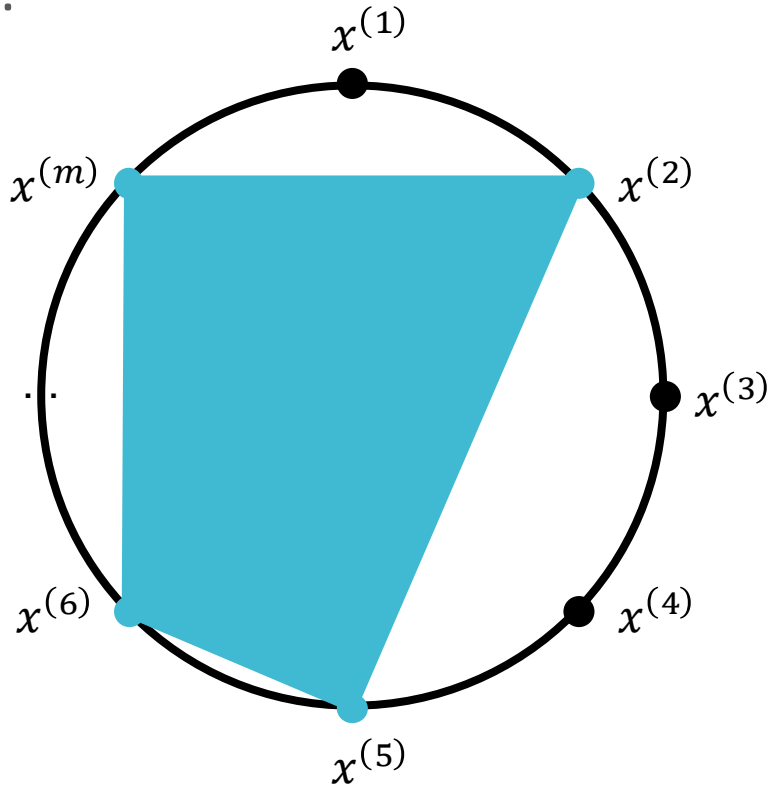


Growth Function: Example

- $x^{(m)} \in \mathbb{R}^2$ and $\mathcal{H} =$ all 2-dimensional positive convex sets
- What are $d_{VC}(\mathcal{H})$ and $g_{\mathcal{H}}(M)$?

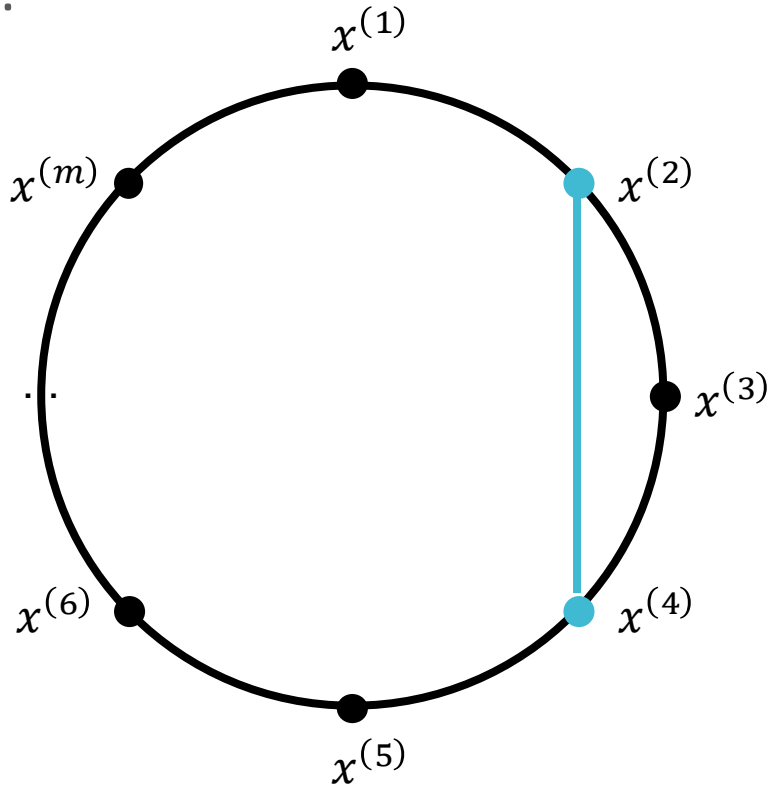
Growth Function: Example

- $x^{(m)} \in \mathbb{R}^2$ and $\mathcal{H} =$ all 2-dimensional positive convex sets
- What are $d_{VC}(\mathcal{H})$ and $g_{\mathcal{H}}(M)$?



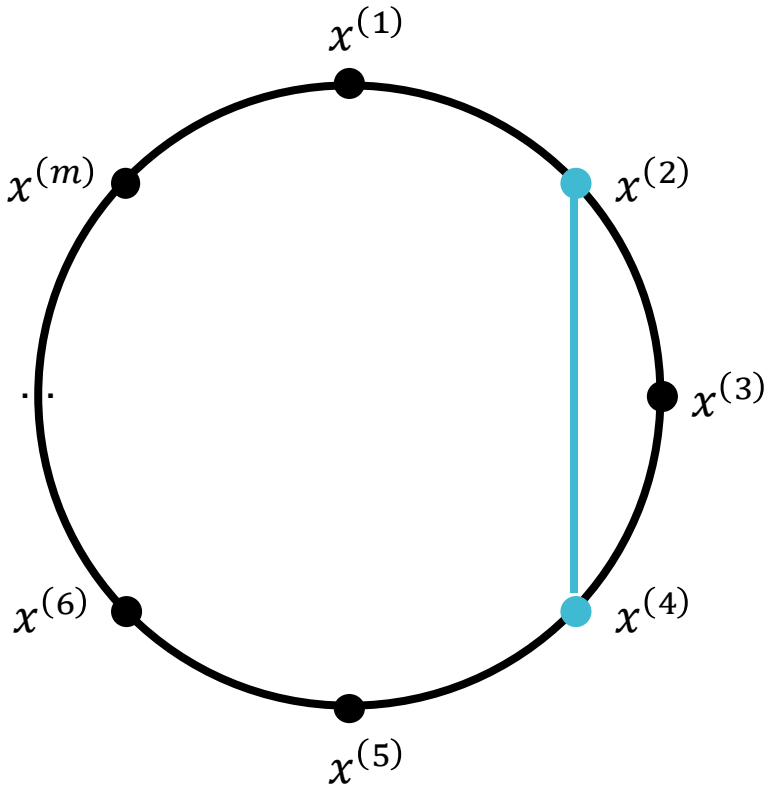
Growth Function: Example

- $x^{(m)} \in \mathbb{R}^2$ and $\mathcal{H} =$ all 2-dimensional positive convex sets
- What are $d_{VC}(\mathcal{H})$ and $g_{\mathcal{H}}(M)$?



Growth Function: Example

- $x^{(m)} \in \mathbb{R}^2$ and $\mathcal{H} =$ all 2-dimensional positive convex sets
- $d_{VC}(\mathcal{H}) = \infty$ and $g_{\mathcal{H}}(M) = 2^M$



Theorem 3: Vapnik-Chervonenkis (VC)-Bound

- Infinite, realizable case: for any hypothesis set $\mathcal{H}$ and distribution p^*, if the number of labelled training data points satisfies

$$M = O\left(\frac{1}{\epsilon}\left(d_{VC}(\mathcal{H})\log\left(\frac{1}{\epsilon}\right) + \log\left(\frac{1}{\delta}\right)\right)\right)$$

then with probability at least $1-\delta$, all $h \in \mathcal{H}$ with $\hat{R}(h) = 0$ have $R(h) \leq \epsilon$

Statistical Learning Theory Corollary

- Infinite, realizable case: for any hypothesis set $\mathcal{H}$ and distribution p^*, given a training data set S s.t. $|S| = M$, all $h \in \mathcal{H}$ with $\hat{R}(h) = 0$ have

$$R(h) \leq O\left(\frac{1}{M}\left(d_{VC}(\mathcal{H}) \log\left(\frac{M}{d_{VC}(\mathcal{H})}\right) + \log\left(\frac{1}{\delta}\right)\right)\right)$$

with probability at least $1 - \delta$.

Theorem 4: Vapnik-Chervonenkis (VC)-Bound

- Infinite, agnostic case: for any hypothesis set $\mathcal{H}$ and distribution p^*, if the number of labelled training data points satisfies

$$M = O\left(\frac{1}{\epsilon^2}\left(d_{VC}(\mathcal{H}) + \log\left(\frac{1}{\delta}\right)\right)\right)$$

then with probability at least $1 - \delta$, all $h \in \mathcal{H}$ have

$$\left|R(h) - \hat{R}(h)\right| \leq \epsilon$$

Statistical Learning Theory Corollary

- Infinite, agnostic case: for any hypothesis set $\mathcal{H}$ and distribution p^*, given a training data set S s.t. $|S| = M$, all $h \in \mathcal{H}$ have

$$R(h) \leq \hat{R}(h) + O\left(\sqrt{\frac{1}{M}\left(d_{VC}(\mathcal{H}) + \log\left(\frac{1}{\delta}\right)\right)}\right)$$

with probability at least $1 - \delta$.

Approximation Generalization Tradeoff

How well does h generalize?

$$R(h) \leq \hat{R}(h) + O\left(\sqrt{\frac{1}{M}\left(d_{VC}(\mathcal{H}) + \log\left(\frac{1}{\delta}\right)\right)}\right)$$

How well does h approximate c^* on training data?

Approximation Generalization Tradeoff

Increases as $d_{VC}(\mathcal{H})$ increases

$$R(h) \leq \hat{R}(h) + O\left(\sqrt{\frac{1}{M}\left(d_{VC}(\mathcal{H}) + \log\left(\frac{1}{\delta}\right)\right)}\right)$$

Decreases as $d_{VC}(\mathcal{H})$ increases

VC dimension and size of hypothesis space

- To be able to shatter m points, how many hypothesis do we need?

  2^m labelings ⇨ $|H| \geq 2^m$

- Given $|H|$ hypothesis, number of points we can shatter $m <= \log_2|H|$

  $VC(H) \leq \log_2|H|$

- So VC bound is tighter.

Limitation of VC dimension

- Hard to compute for many hypothesis spaces

  VC(H) ≥ lower bound (easy)

  VC(H) = … (HARD!)

  For all placements of VC(H)+1 points, there exists a labeling that can't be shattered

- Too loose for many hypothesis spaces

  linear SVMs, VC dim = d+1 (d features)

  kernel SVMs, VC dim = ??

  = ∞ (Gaussian kernels)

Suggests Gaussian kernels are really BAD!!

Quick Game

Quick game: name the case

The lecture organizes learning-theory guarantees into four cases. Match each scenario to the right quadrant.

Scenarios

- **A** Decision tree of depth d; one hypothesis fits the data perfectly
- **B** Decision tree of depth d; labels have noise, so zero train error may be impossible
- **C** Infinite hypothesis class H; assume H contains the target concept
- **D** Infinite H; want bounds when training error may remain nonzero

Quick Game

Quick game: name the case

The lecture organizes learning-theory guarantees into four cases. Match each scenario to the right quadrant.

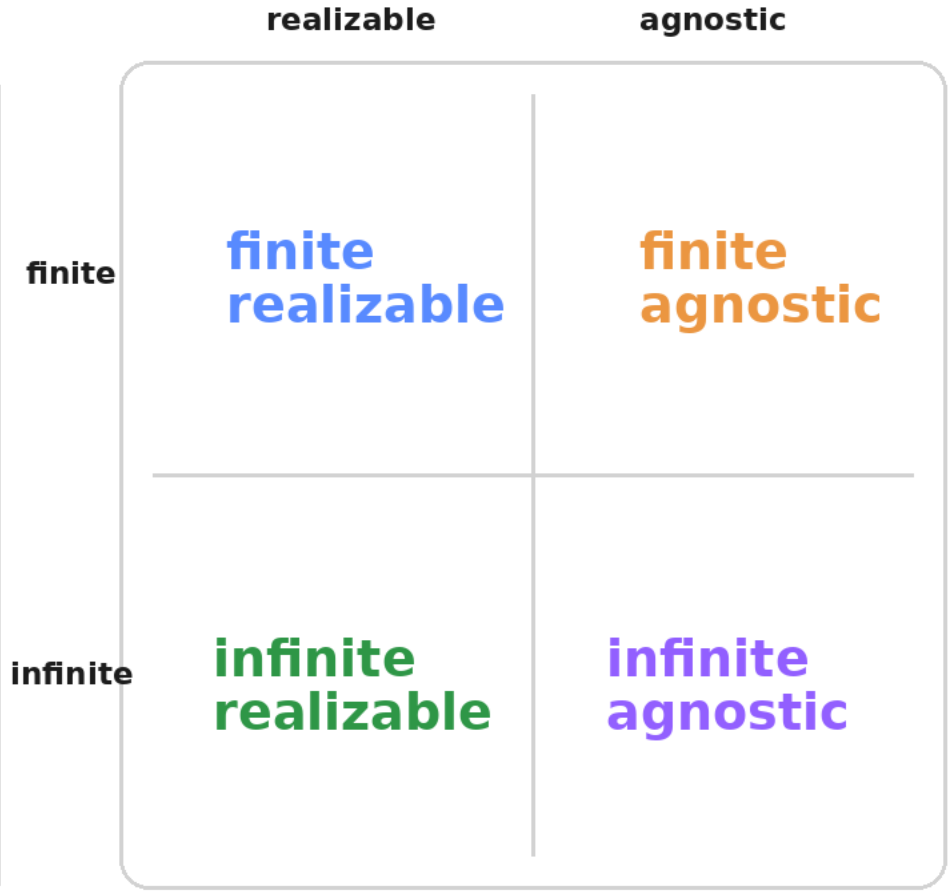
Scenarios

A Decision tree of depth d; one hypothesis fits the data perfectly

B Decision tree of depth d; labels have noise, so zero train error may be impossible

C Infinite hypothesis class H; assume H contains the target concept

D Infinite H; want bounds when training error may remain nonzero



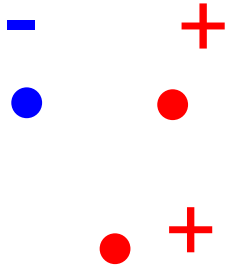
Reveal: A→finite realizable, B→finite agnostic, C→infinite realizable, D→infinite agnostic.

Rademacher Complexity

- Instead of all possible labelings, measure complexity by how accurately a hypothesis space can match a random labeling of the data.

For each data point i, draw random label

σ_i s.t. $P(\sigma_i = +1) = ½ = P(\sigma_i = -1)$



Then empirical Rademacher complexity of H is

$$\widehat{R}_m(H) = \mathbb{E}_\sigma \left[\sup_{h \in H} \left(\frac{1}{m} \sum_{i=1}^{m} \sigma_i h(X_i) \right) \right]$$

Max correlation possible with random labels

Rademacher Bounds

- With probability ≥ 1-δ,

$$\text{error}_{true}(h) \leq \text{error}_{train}(h) + \widehat{R}_m(H) + 3\sqrt{\frac{\log(2/\delta)}{m}}$$

where empirical Rademacher complexity of H

$$\widehat{R}_m(H) = \mathbb{E}_\sigma \left[\sup_{h \in H} \left(\frac{1}{m} \sum_{i=1}^{m} \sigma_i h(X_i) \right) \right]$$

is purely data-dependent.

Myth Busters

MythBusters #2

“A data-dependent bound is cheating.”

Myth: if the complexity measure depends on the observed sample, the guarantee is no longer meaningful.

Reality: data-dependent quantities can be better matched to the sample at hand. Empirical Rademacher complexity is useful because it adapts to the observed data, not because it memorizes labels.

Worst-case view

same complexity no matter what sample appears

Data-dependent view

complexity adapts to the observed sample geometry

The benefit is sharper bounds when the realized sample is simpler than the worst case.

Summary of PAC bounds

With probability ≥ 1-δ,

1) for all h ∈ H s.t. $error_{train}(h) = 0$,

$$error_{true}(h) \leq \varepsilon = \frac{\ln |H| + \ln \frac{1}{\delta}}{m}$$

2) for all h ∈ H,

$$|error_{true}(h) - error_{train}(h)| \leq \varepsilon = \sqrt{\frac{\ln |H| + \ln \frac{1}{\delta}}{2m}}$$

Finite hypothesis space

3) for all h ∈ H,

$$|error_{true}(h) - error_{train}(h)| \leq \varepsilon = 8\sqrt{\frac{VC(H)\left(\ln \frac{m}{VC(H)} + 1\right) + \ln \frac{8}{\delta}}{2m}}$$

4) For all h ∈ H,

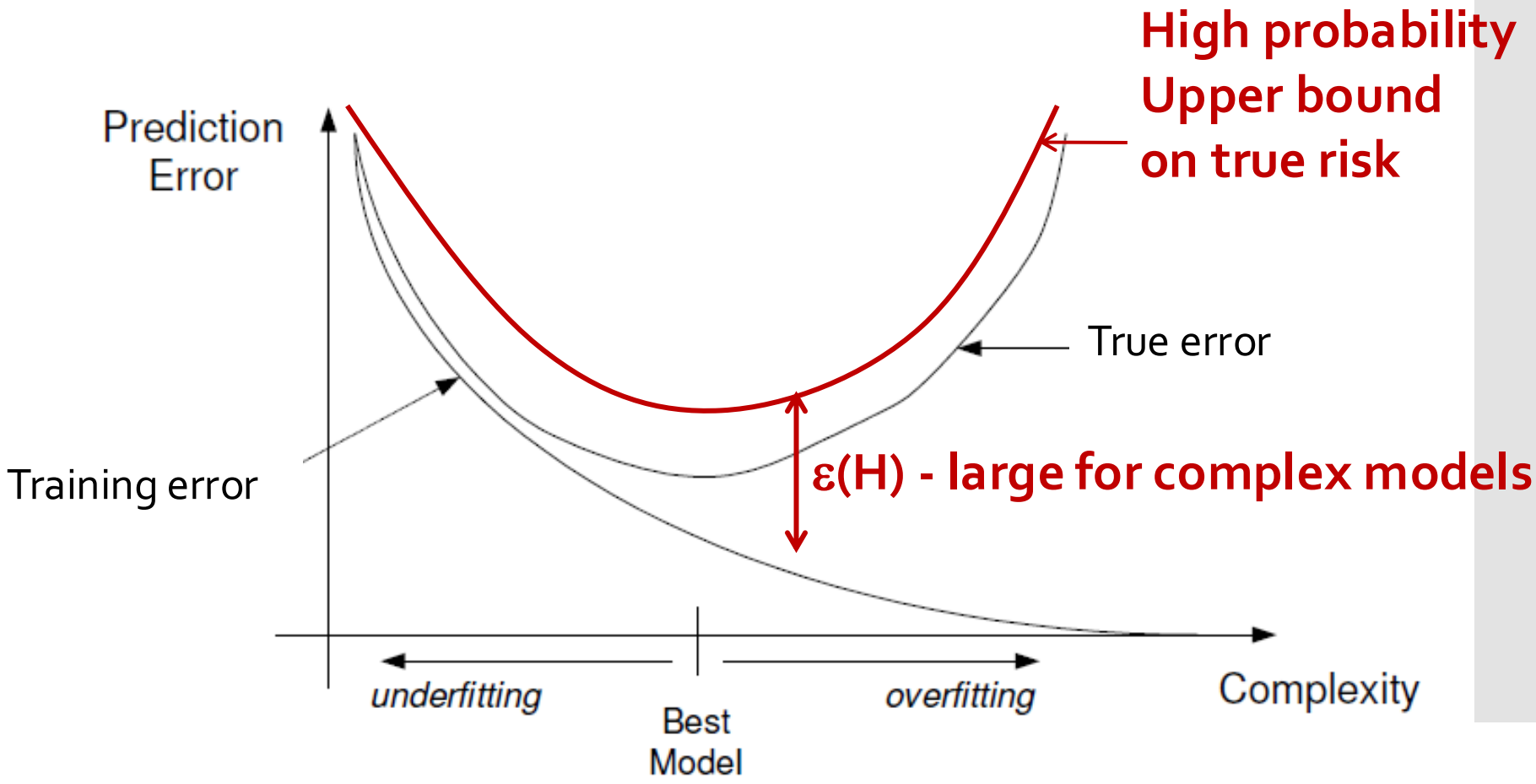
$$|error_{true}(h) - error_{train}(h)| \leq \varepsilon = \widehat{R}_m(H) + 3\sqrt{\frac{\log(2/\delta)}{m}}$$

Infinite hypothesis space

PAC Bounds

With probability $\geq 1-\delta$, for all $h \in H$,

$$|error_{true}(h) - error_{train}(h)| \leq \varepsilon(H)$$



Key Takeaways

- For infinite hypothesis sets, use the VC-dimension (or the growth function) as a measure of complexity
  - Computing $d_{VC}(\mathcal{H})$ and $g_{\mathcal{H}}(M)$
  - Connection between VC-dimension and the growth function (Sauer-Shelah lemma)
  - Sample complexity and statistical learning theory style bounds using $d_{VC}(\mathcal{H})$

Quick Game

Quick game: finite or infinite?

Classify each hypothesis class example using the lecture’s four-case language.

Examples

A Depth-d decision trees over p Boolean features

B All linear separators in $\mathbb{R}^2$

C Intervals on the real line

Vote / classify

1. **finite H**
2. **infinite H but finite VC-dim**
3. **infinite H and use richer capacity t**

Quick Game

Quick game: finite or infinite?

Classify each hypothesis class example using the lecture’s four-case language.

Examples

A Depth-d decision trees over p Boolean features

B All linear separators in $\mathbb{R}^2$

C Intervals on the real line

Vote / classify

1. **finite H**
2. **infinite H but finite VC-dim**
3. **infinite H and use richer capacity t**

Reveal: A→1, B→2, C→2.

Risk Measures

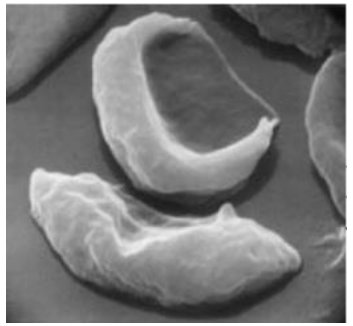
- The bounds so far have focused on binary classification and mis-classification error
  - Not all tasks are binary classification!
- The concepts of “training risk” and “true risk” are far more generally applicable
- But first, how do we define the risk?

Supervised Learning Prediction Task

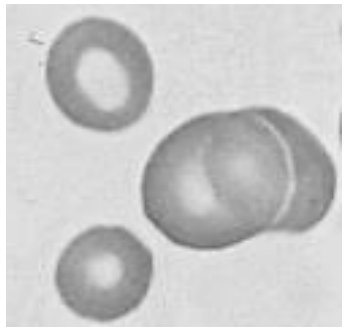
Task:

Given $X \in \mathcal{X}$, predict $Y \in \mathcal{Y}$.

$\equiv$ Construct **prediction rule** $f : \mathcal{X} \to \mathcal{Y}$



"Lupus cell (0)"



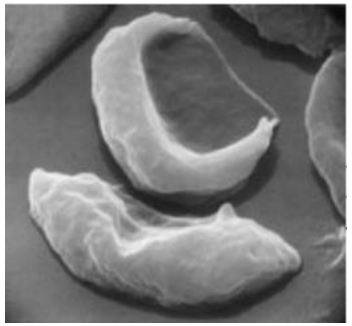
"Healthy cell (1)"

Supervised Learning Prediction Task

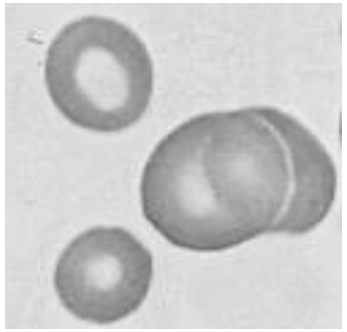
Task:

Given $X \in \mathcal{X}$, predict $Y \in \mathcal{Y}$.

$\equiv$ Construct **prediction rule** $f : \mathcal{X} \to \mathcal{Y}$



"Lupus cell (0)"



"Healthy cell (1)"

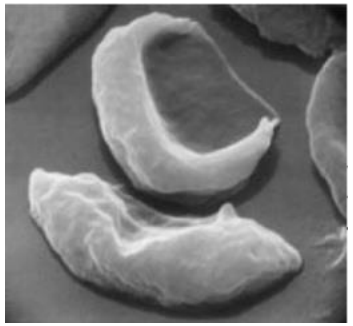
But I can always come up with a prediction rule: always say it's not LUPUS!

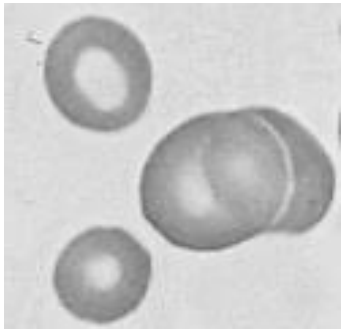
Example: Supervised Learning Prediction Task

Task:

Given $X \in \mathcal{X}$, predict $Y \in \mathcal{Y}$.

$\equiv$ Construct **prediction rule** $f : \mathcal{X} \to \mathcal{Y}$





"Lupus (0)"

"Healthy (1)"

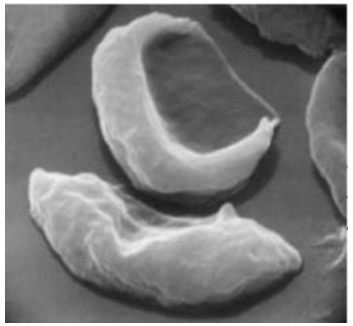
To complete the specification of the task, we need something more!!!

Characterize Task using Performance Measures

Performance Measure:

What is the "loss" I suffer when I take decision **f**?

$\text{loss}(Y, f(X))$ - Measure of closeness between true label *Y* and prediction *f*(*X*)

X	*Y*	*f*(*X*)	$\text{loss}(Y, f(X))$
	"Lupus"	"Lupus"	0
		"Healthy"	1

$$\text{loss}(Y, f(X)) = 1_{\{f(X) \neq Y\}}$$ **0/1 loss**

Characterize Task using Performance Measures

Performance Measure:

What is the "loss" I suffer when I take decision **f**?

$\text{loss}(Y, f(X))$ - Measure of closeness between true label *Y* and prediction *f*(*X*)

X	Share price, *Y*	*f*(*X*)	$\text{loss}(Y, f(X))$
Past performance, trade volume etc. as of Sept 8, 2010	"$24.50"	"$24.50"	0
		"$26.00"	1?
		"$26.10"	2?

$$\text{loss}(Y, f(X)) = (f(X) - Y)^2$$ **square loss**

Performance Measures

Performance Measure:

$\text{loss}(Y, f(X))$ - Measure of closeness between true label *Y* and prediction *f*(*X*)

We don't just want to correctly label one test sample (in this case, cell image), but most $X \in$ "$\mathcal{X}$" images

Given a cell image drawn randomly from the collection of all cell images, how well does the predictor perform on average?

$$\text{Risk } R(f) \equiv \mathbb{E}_{XY}\left[\text{loss}(Y, f(X))\right]$$

Performance Measures

Performance Measure:

$\text{loss}(Y, f(X))$ - Measure of closeness between true label *Y* and prediction *f*(*X*)

We don't just want to correctly label one test sample (in this case, cell image), but most cell images $X \in \mathcal{X}$

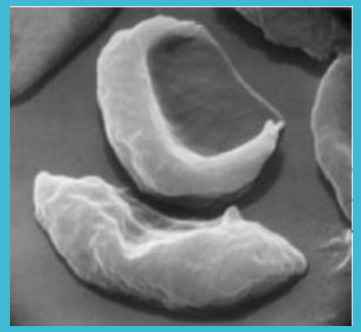
Given a cell image drawn randomly from the collection of all cell images, how well does the predictor perform on average?

What is the "risk" of taking decision **f**?

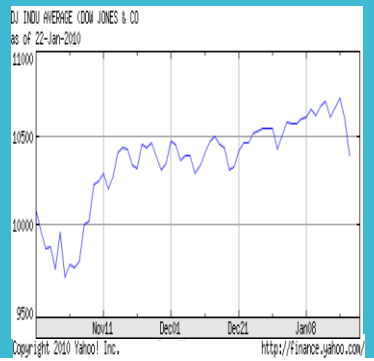
$$\text{Risk } R(f) \equiv \mathbb{E}_{XY}\left[\text{loss}(Y, f(X))\right]$$

Performance Measure:

$$\text{Risk } R(f) \equiv \mathbb{E}_{XY}\left[\text{loss}(Y, f(X))\right]$$



"Anemic cell"



Share Price
"$ 24.50"

$\text{loss}(Y, f(X))$	Risk $R(f)$
$1_{\{f(X) \neq Y\}}$	$P(f(X) \neq Y)$
0/1 loss	**Probability of Error**
$(f(X) - Y)^2$	$\mathbb{E}[(f(X) - Y)^2]$
square loss	**Mean Square Error**

Bayes Optimal Rule

Construct **prediction rule** $f^* : \mathcal{X} \to \mathcal{Y}$

$$f^*(P) = \arg\min_f \mathbb{E}_{(X,Y)\sim P}[\text{loss}(Y, f(X))]$$

Bayes optimal rule

Optimal rule is not computable
Depends on unknown distribution P over (X,Y) !

Empirical Risk Minimization

Empirical Risk Minimizer

Construct **prediction rule** $f^* : \mathcal{X} \to \mathcal{Y}$

$f^*(P) = \arg\min_f \mathbb{E}_{(X,Y)\sim P}[\text{loss}(Y, f(X))]$ Bayes optimal rule

Given $\{X_i, Y_i\}_{i=1}^n$, **learn** prediction rule $\widehat{f}_n : \mathcal{X} \to \mathcal{Y}$

$$\widehat{f}_n = \arg\min_f \frac{1}{n}\sum_{i=1}^{n} [\text{loss}(Y_i, f(X_i))]$$

$$\frac{1}{n}\sum_{i=1}^{n} [\text{loss}(Y_i, f(X_i))] \xrightarrow[\text{Numbers}]{\text{Law of Large}} \mathbb{E}_{XY}[\text{loss}(Y, f(X))]$$

Empirical Risk Minimization

- Given a loss function, and data, estimate decision function by minimizing “empirical risk”
- Restrict decision to lie within some restricted set of hypotheses

$$\widehat{f} = \arg \inf_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \text{loss}(Y_i, f(X_i)) \right\}$$

Empirical Risk Minimization: Considerations

$$\widehat{f} = \arg \inf_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \text{loss}(Y_i, f(X_i)) \right\}$$

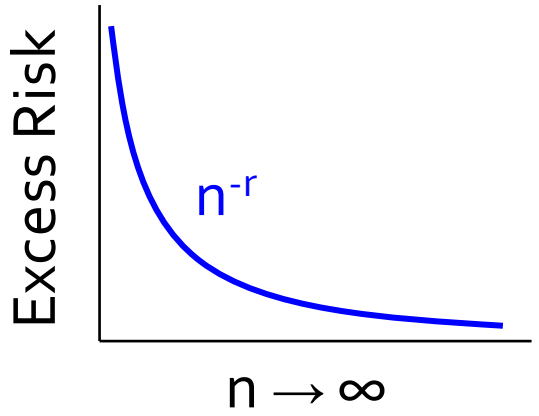
- **Computational Considerations**: How do we solve the above optimization problem in a computationally tractable manner?
  - Covered in the optimization lectures
- **Statistical Considerations**: What guarantees do I have for the empirical risk minimizer (ERM) estimator?
  - Covered in these learning theory lectures

Statistical Considerations

- How does the performance of the algorithm compare with ideal performance?

Excess Risk $\mathbb{E}_{D_n}\left[R(\hat{f}_n)\right] - R(f^*)$

- **Consistent** algorithm if Excess Risk → 0 as n → ∞
- **Rate of Convergence**



Computational Considerations

$$\widehat{f} = \arg \inf_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \text{loss}(Y_i, f(X_i)) \right\}$$

- Even when class of functions is simple (e.g. class of linear functions), the above optimization need not be **convex**
- This non-convexity, and consequently, computational intractability holds for 0-1 loss classification

Surrogate Losses for Classification

Recall the zero-one loss:

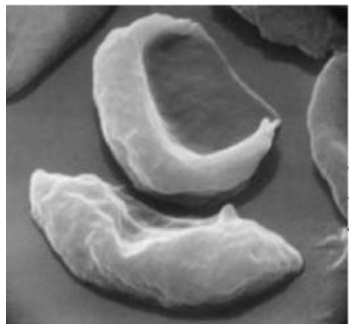
$$\ell_{0/1}(Y, f(X)) = \mathbb{I}(Y \neq f(X))$$

The loss is either zero or one (hence its name)

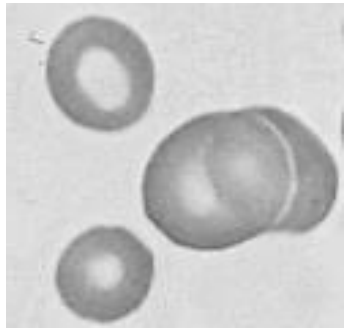
Loss is zero if the classifier outputs the label exactly

Loss is one if not

Binary Classification



"Lupus (-1)"



"Healthy (1)"

Binary Classification and 0-1 Loss

$$\ell_{0/1}(Y, f(X)) := \mathbb{I}(Y \neq \text{sign}(f(X)))$$
$$= \mathbb{I}(Y\, f(X) < 0)$$

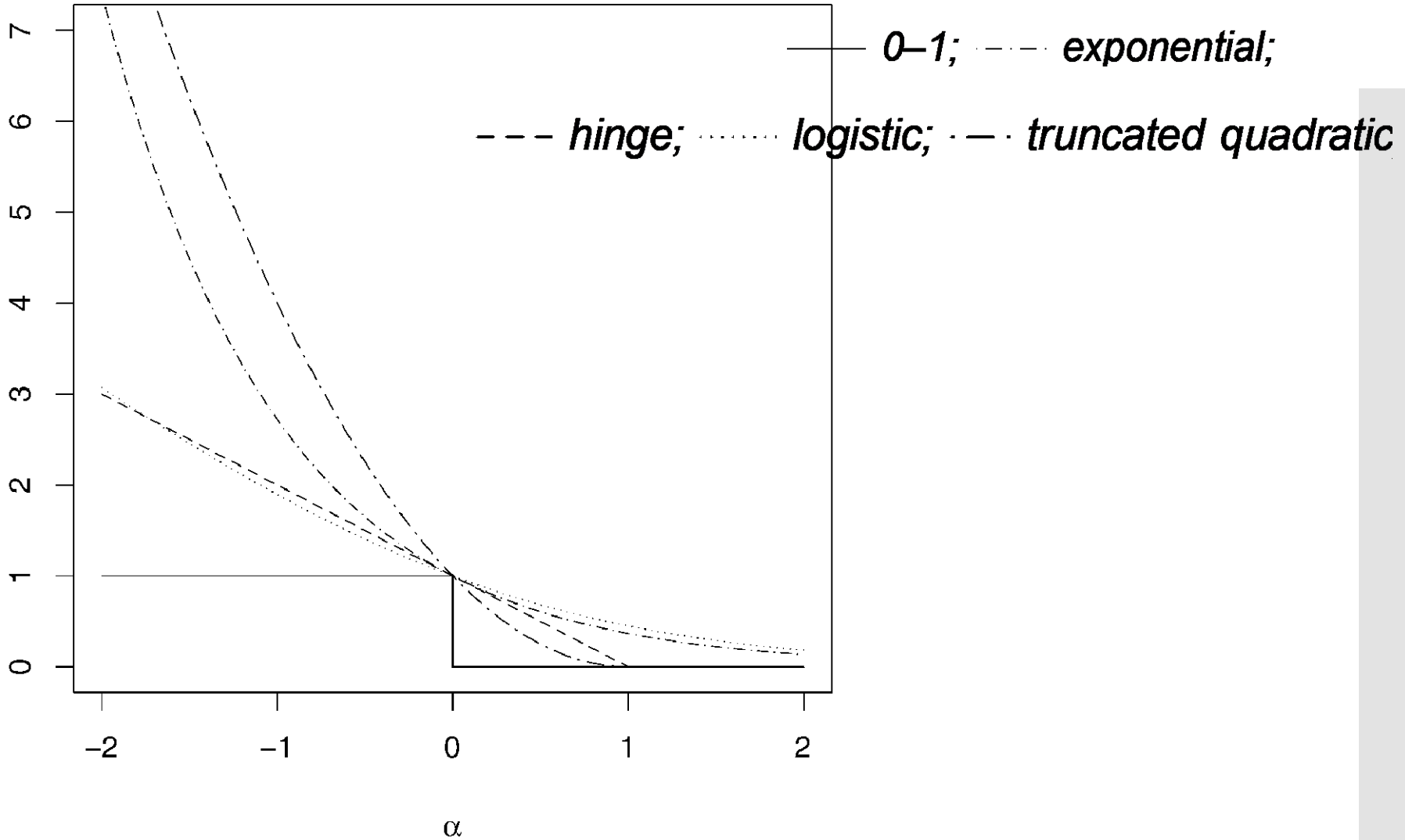
Binary Classification and 0-1 Loss

$$\ell_{0/1}(Y, f(X)) := \mathbb{I}(Y \neq \text{sign}(f(X)))$$
$$= \mathbb{I}(Y\, f(X) < 0)$$

$$\text{Can write it as } \ell_{0/1}(Y, f(X)) = \ell(Y\, f(X))$$
$$\text{where } \ell(\alpha) = \mathbb{I}(\alpha < 0)$$

Empirical Risk Minimizer with respect to 0-1 loss is computationally intractable
because 0-1 loss above is **non-convex**

Binary Classification: Convex Surrogates



Different loss functions $\ell(\alpha)$
where use in classification would be as:
$\ell(Y, f(X)) = \ell(Y\ f(X))$

Recall: Logistic Regression

$$f(x) = w_0 + \sum_j w_j x_j$$

Logistic regression assumes:

$$P(Y = 1|X) = \frac{1}{1 + \exp(f(x))}$$

And tries to maximize data likelihood:

$$P(\mathcal{D}|f) = \prod_{i=1}^{m} \frac{1}{1 + \exp(-y_i f(x_i))}$$

Equivalent to minimizing log loss

$$-\log P(\mathcal{D}|f) = \sum_{i=1}^{m} \ln(1 + \exp(-y_i f(x_i)))$$

Boosting (in upcoming lecture)

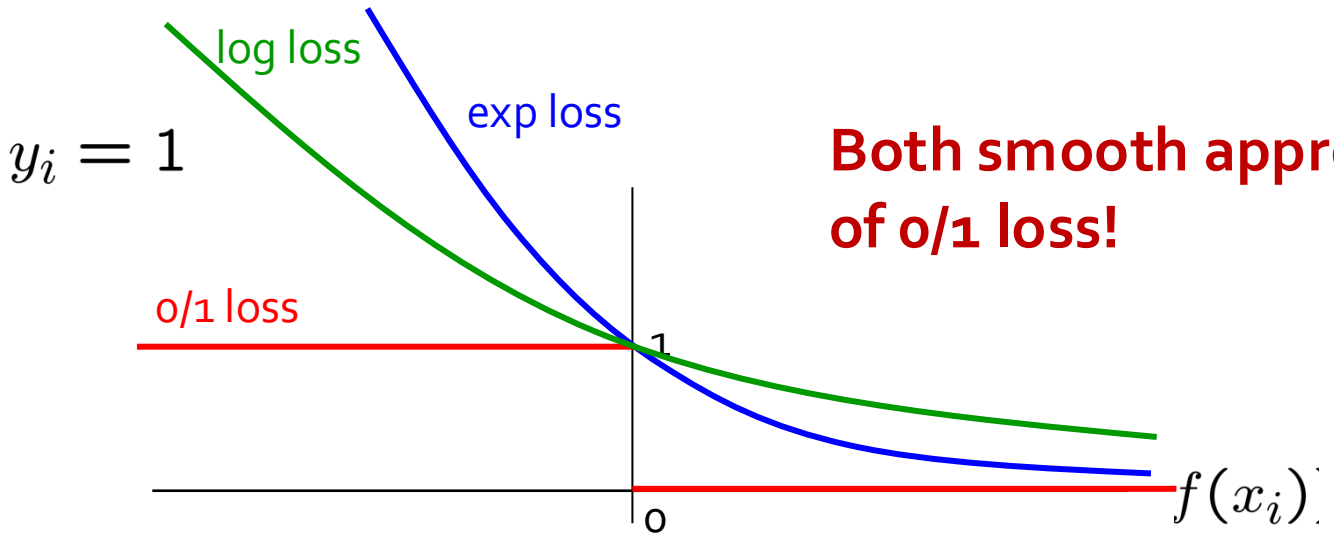
Boosting minimizes similar loss function!!

$$\frac{1}{m}\sum_{i=1}^{m}\exp(-y_i f(x_i))$$

$$f(x) = \sum_t \alpha_t h_t(x)$$

Weighted average of weak learners

Logistic regression equivalent to minimizing log loss



Both smooth approximations of 0/1 loss!

Takeaway: Binary Classification

$$\widehat{f}_{0/1} = \arg \inf_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell_{0/1}(Y_i \, f(X_i)) \right\}$$

ERM with respect to 0-1 loss: computationally intractable

$$\widehat{f}_{\phi} = \arg \inf_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \phi(Y_i \, f(X_i)) \right\}$$

ERM with respect to convex surrogate loss: computationally tractable!
Basis of all modern classifiers: boosting, support vector machines, logistic regression, etc.

Final Poll

Poll:

Lecture 21

Final poll (pick one)

Which statement best matches this lecture’s main message?

A. Once VC dimension is finite, there is little reason to consider any other complexity notion.

B. Empirical Rademacher complexity can give a sharper, data-dependent view of how rich a hypothesis class is on the observed sample.

C. Data-dependent bounds are invalid because they use the training sample.

D. Infinite hypothesis classes automatically imply that no meaningful generalization guarantees are possible.

E. Union-bound style counting is always as sharp as any more refined capacity analysis.