Machine Learning Department
School of Computer Science
Carnegie Mellon University

PAC Learning

+

MLE/MAP

Matt Gormley & Henry Chai

Lecture 16

Mar. 12, 2025

Reminders

- **Homework 5: Neural Networks**
  - **Out: Wed, Feb-26**
  - **Due: Sun, Mar-16 at 11:59pm**

LEARNING THEORY

Questions

1. Given a classifier with **zero training error**, what can we say about **true error** (aka. generalization error)?
   (Sample Complexity, Realizable Case)

2. Given a classifier with **low training error**, what can we say about **true error** (aka. generalization error)?
   (Sample Complexity, Agnostic Case)

3. Is there a **theoretical justification for regularization** to avoid overfitting?
   (Structural Risk Minimization)

IMPORTANT NOTE

In our discussion of PAC Learning, we are only concerned with the problem of **binary** classification

There are other theoretical frameworks (including PAC) that handle other learning settings, but this provides us with a representative one.

PAC / SLT Model

We've also referred to this as the "Function Approximation View"

1. Generate instances from *unknown* distribution p^*

$$\mathbf{x}^{(i)} \sim p^*(\mathbf{x}),\ \forall i \tag{1}$$

2. Oracle labels each instance with *unknown* function c^*

$$y^{(i)} = c^*(\mathbf{x}^{(i)}),\ \forall i \tag{2}$$

3. Learning algorithm chooses hypothesis $h \in \mathcal{H}$ with low(est) training error, $\hat{R}(h)$

$$\hat{h} = \operatorname*{argmin}_h \hat{R}(h) \tag{3}$$

4. Goal: Choose an h with low generalization error $R(h)$

Three Hypotheses of Interest

The **true function** c^* is the one we are trying to learn and that labeled the training data:

$$y^{(i)} = c^*(\mathbf{x}^{(i)}),\ \forall i \tag{1}$$

The **expected risk minimizer** has lowest true error:

$$h^* = \underset{h \in \mathcal{H}}{\operatorname{argmin}}\, R(h) \tag{2}$$

The **empirical risk minimizer** has lowest training error:

$$\hat{h} = \underset{h \in \mathcal{H}}{\operatorname{argmin}}\, \hat{R}(h) \tag{3}$$

SAMPLE COMPLEXITY RESULTS

Sample Complexity Results

Definition 0.1. The **sample complexity** of a learning algorithm is the number of examples required to achieve arbitrarily small error (with respect to the optimal hypothesis) with high probability (i.e. close to 1).

Four Cases we care about…

	Realizable	Agnostic
Finite $\lvert\mathcal{H}\rvert$	**Thm. 1** $N \geq \frac{1}{\epsilon}\left[\log(\lvert\mathcal{H}\rvert) + \log(\frac{1}{\delta})\right]$ labeled examples are sufficient so that with probability $(1-\delta)$ all $h \in \mathcal{H}$ with $\hat{R}(h) = 0$ have $R(h) \leq \epsilon$.	
Infinite $\lvert\mathcal{H}\rvert$		

Sample Complexity Results

Definition 0.1. The **sample complexity** of a learning algorithm is the number of examples required to achieve arbitrarily small error (with respect to the optimal hypothesis) with high probability (i.e. close to 1).

Four Cases we care about…

$c^* \in \mathcal{H}$ $c^* \in \mathcal{H}$ or $c^* \notin \mathcal{H}$

	Realizable	Agnostic
Finite $\lvert\mathcal{H}\rvert$	**Thm. 1** $N \geq \frac{1}{\epsilon}\left[\log(\lvert\mathcal{H}\rvert) + \log(\frac{1}{\delta})\right]$ labeled examples are sufficient so that with probability $(1-\delta)$ all $h \in \mathcal{H}$ with $\hat{R}(h) = 0$ have $R(h) \leq \epsilon$.	**Thm. 2** $N \geq \frac{1}{2\epsilon^2}\left[\log(\lvert\mathcal{H}\rvert) + \log(\frac{2}{\delta})\right]$ labeled examples are sufficient so that with probability $(1-\delta)$ for all $h \in \mathcal{H}$ we have that $\lvert R(h) - \hat{R}(h)\rvert \leq \epsilon$.
Infinite $\lvert\mathcal{H}\rvert$		

1. Bound is **inversely linear in epsilon** (e.g. halving the error requires double the examples)
2. Bound is **only logarithmic in |H|** (e.g. quadrupling the hypothesis space only requires double the examples)

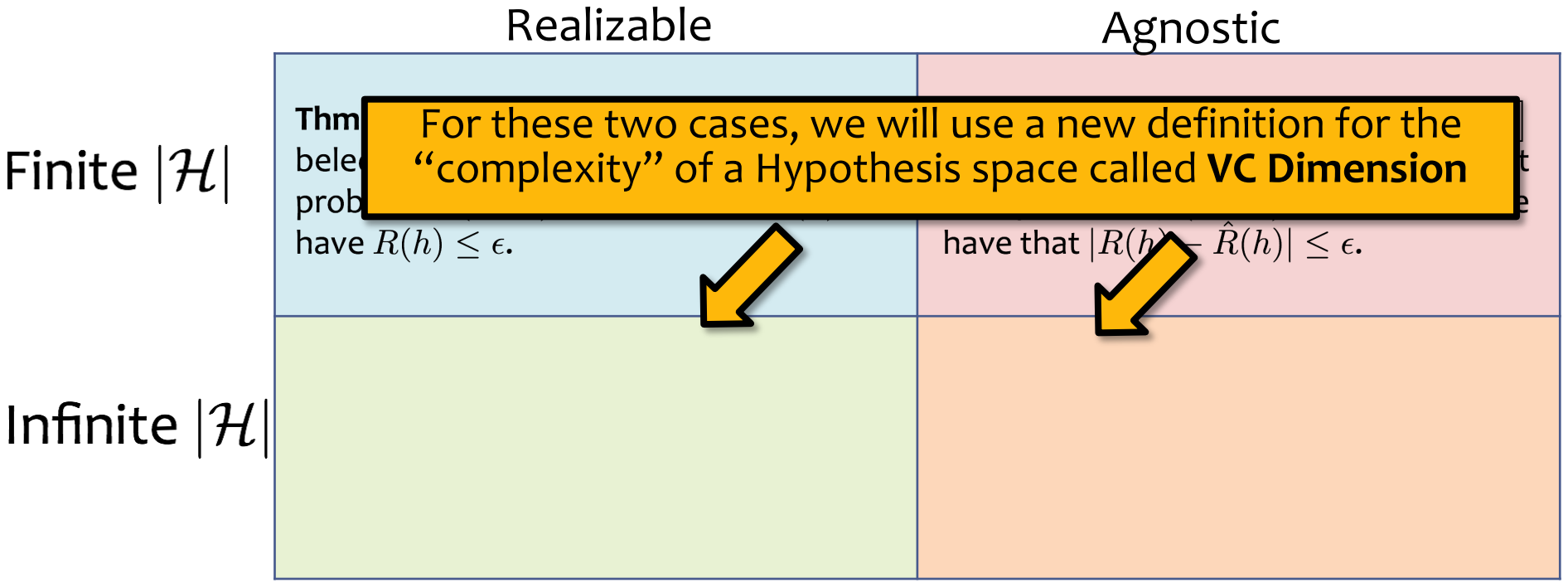
1. Bound is **inversely quadratic in epsilon** (e.g. halving the error requires 4x the examples)
2. Bound is **only logarithmic in |H|** (i.e. same as Realizable case)

	Realizable	Agnostic
Finite $\lvert\mathcal{H}\rvert$	**Thm. 1** $N \geq \frac{1}{\epsilon}\left[\log(\lvert\mathcal{H}\rvert) + \log(\frac{1}{\delta})\right]$ labeled examples are sufficient so that with probability $(1-\delta)$ all $h \in \mathcal{H}$ with $\hat{R}(h) = 0$ have $R(h) \leq \epsilon$.	**Thm. 2** $N \geq \frac{1}{2\epsilon^2}\left[\log(\lvert\mathcal{H}\rvert) + \log(\frac{2}{\delta})\right]$ labeled examples are sufficient so that with probability $(1-\delta)$ for all $h \in \mathcal{H}$ we have that $\lvert R(h) - \hat{R}(h)\rvert \leq \epsilon$.
Infinite $\lvert\mathcal{H}\rvert$		

Sample Complexity Results

Definition 0.1. The **sample complexity** of a learning algorithm is the number of examples required to achieve arbitrarily small error (with respect to the optimal hypothesis) with high probability (i.e. close to 1).

Four Cases we care about…



Sample Complexity Results

Definition 0.1. The **sample complexity** of a learning algorithm is the number of examples required to achieve arbitrarily small error (with respect to the optimal hypothesis) with high probability (i.e. close to 1).

Four Cases we care about…

	Realizable	Agnostic
Finite $\lvert\mathcal{H}\rvert$	**Thm. 1** $N \geq \frac{1}{\epsilon}\left[\log(\lvert\mathcal{H}\rvert) + \log(\frac{1}{\delta})\right]$ labeled examples are sufficient so that with probability $(1-\delta)$ all $h \in \mathcal{H}$ with $\hat{R}(h) = 0$ have $R(h) \leq \epsilon$.	**Thm. 2** $N \geq \frac{1}{2\epsilon^2}\left[\log(\lvert\mathcal{H}\rvert) + \log(\frac{2}{\delta})\right]$ labeled examples are sufficient so that with probability $(1-\delta)$ for all $h \in \mathcal{H}$ we have that $\lvert R(h) - \hat{R}(h)\rvert \leq \epsilon$.
Infinite $\lvert\mathcal{H}\rvert$	**Thm. 3** $N = O(\frac{1}{\epsilon}\left[\text{VC}(\mathcal{H})\log(\frac{1}{\epsilon}) + \log(\frac{1}{\delta})\right])$ labeled examples are sufficient so that with probability $(1-\delta)$ all $h \in \mathcal{H}$ with $\hat{R}(h) = 0$ have $R(h) \leq \epsilon$.	**Thm. 4** $N = O(\frac{1}{\epsilon^2}\left[\text{VC}(\mathcal{H}) + \log(\frac{1}{\delta})\right])$ labeled examples are sufficient so that with probability $(1-\delta)$ for all $h \in \mathcal{H}$ we have that $\lvert R(h) - \hat{R}(h)\rvert \leq \epsilon$.

Sample Complexity Results

Four Cases we care about…

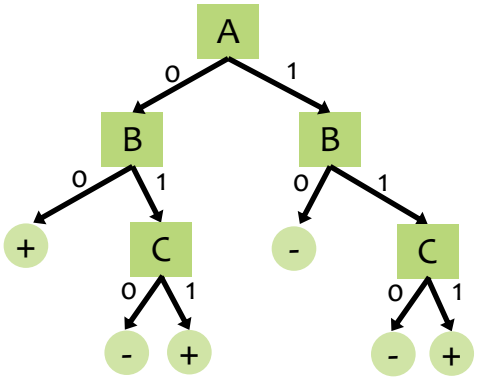
	Realizable	Agnostic
Finite $\lvert\mathcal{H}\rvert$	**Thm. 1** $N \geq \frac{1}{\epsilon}\left[\log(\lvert\mathcal{H}\rvert) + \log(\frac{1}{\delta})\right]$ labeled examples are sufficient so that with probability $(1-\delta)$ all $h \in \mathcal{H}$ with $\hat{R}(h) = 0$ have $R(h) \leq \epsilon$.	**Thm. 2** $N \geq \frac{1}{2\epsilon^2}\left[\log(\lvert\mathcal{H}\rvert) + \log(\frac{2}{\delta})\right]$ labeled examples are sufficient so that with probability $(1-\delta)$ for all $h \in \mathcal{H}$ we have that $\lvert R(h) - \hat{R}(h)\rvert \leq \epsilon$.
Infinite $\lvert\mathcal{H}\rvert$	**Thm. 3** $N=O(\frac{1}{\epsilon}\left[\text{VC}(\mathcal{H})\log(\frac{1}{\epsilon}) + \log(\frac{1}{\delta})\right])$ labeled examples are sufficient so that with probability $(1-\delta)$ all $h \in \mathcal{H}$ with $\hat{R}(h) = 0$ have $R(h) \leq \epsilon$.	**Thm. 4** $N = O(\frac{1}{\epsilon^2}\left[\text{VC}(\mathcal{H}) + \log(\frac{1}{\delta})\right])$ labeled examples are sufficient so that with probability $(1-\delta)$ for all $h \in \mathcal{H}$ we have that $\lvert R(h) - \hat{R}(h)\rvert \leq \epsilon$.

VC-DIMENSION

Finite vs. Infinite |H|

Finite |H|

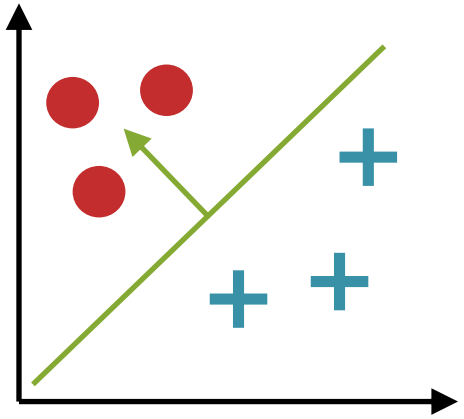
- *Example*: H = the set of all decision trees of depth D over binary feature vectors of length M



- *Example*: H = the set of all conjunctions over binary feature vectors of length M

Infinite |H|

- *Example*: H = the set of all linear decision boundaries in M dimensions



- *Example*: H = the set of all neural networks with 1-hidden layer with length M inputs

Labelings & Shattering

Def 1: A hypothesis h applied to some dataset S generates a **labeling** of S.

Example: labelings

Labelings & Shattering

Def 1: A hypothesis h applied to some dataset S generates a **labeling** of S.

Def 2: Let $\mathcal{H}[S]$ be the set of **all** (distinct) **labelings** of S generated by hypotheses $h \in \mathcal{H}$.

$|\mathcal{H}[S]| = 2^4 - 2$

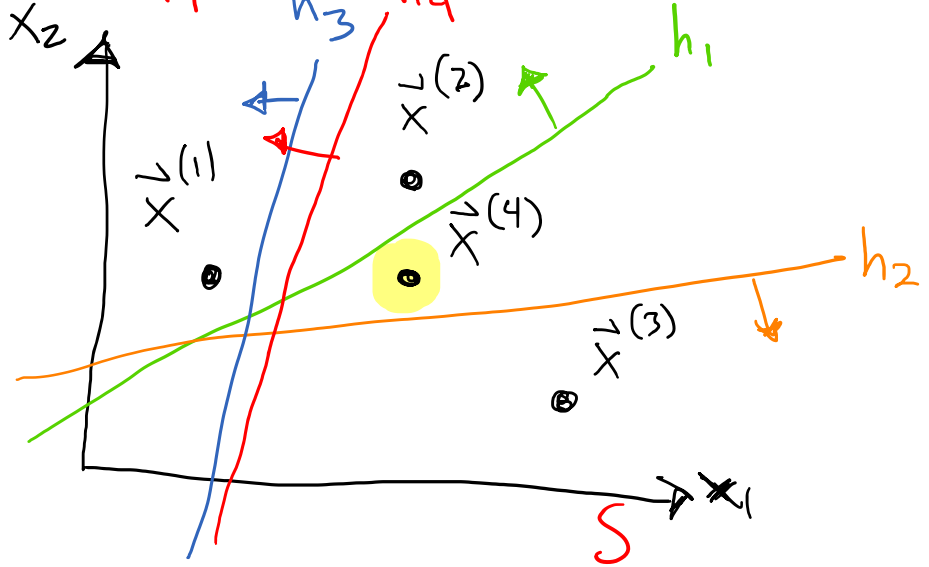
the two that we can't produce with $\mathcal{H}$

Example: labelings

$\mathcal{H}$ = all lin. dec. bound in 2D

Labelings

h_t	$\vec{x}^{(1)}$	$\vec{x}^{(2)}$	$\vec{x}^{(3)}$	$\vec{x}^{(4)}$
h_1	+	+	−	−
h_2	−	−	+	−
h_3	+	−	−	−
h_4	+	−	−	−



Labelings & Shattering

Def 1: A hypothesis h applied to some dataset S generates a **labeling** of S.

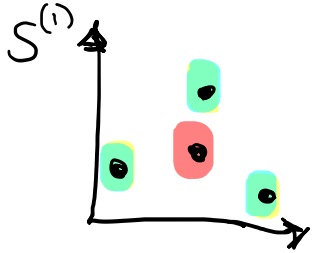
Def 2: Let $\mathcal{H}[S]$ be the set of **all** (distinct) **labelings** of S generated by hypotheses $h \in \mathcal{H}$.

Def 3: $\mathcal{H}$ **shatters** S if $|\mathcal{H}[S]| = 2^{|S|}$

or equivalently, the hypotheses in $\mathcal{H}$ can generate every possible labeling of S.

Example: shattering

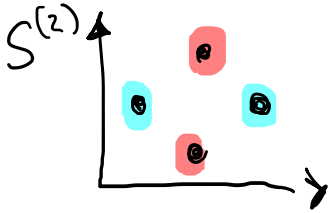
$\mathcal{H}$ = lin dec. bound in 2D



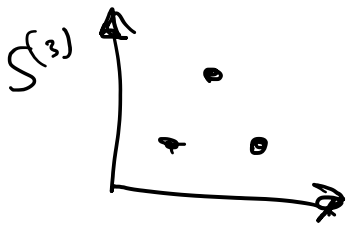
$|\mathcal{H}[S]| = 14$

$2^{|S|} = 16$

$\mathcal{H}$ does not shatter $S^{(1)}$



$\mathcal{H}$ does not shatter $S^{(2)}$



$\mathcal{H}$ does shatter $S^{(3)}$

VC-dimension

Def: The **VC-dimension** (or Vaporik-Chervonenkis dimension) of $\mathcal{H}$ is the cardinality of the largest set S such that $\mathcal{H}$ can shatter S.

Special Case: If $\mathcal{H}$ can shatter arbitrarily large finite sets, then the VC-dimension of $\mathcal{H}$ is infinity

Notation: We write $\mathrm{VC}(\mathcal{H}) = d$ to say the VC-Dimension of a hypothesis space $\mathcal{H}$ is d

VC-dimension Proof

Proof Technique: To **prove** that $\mathrm{VC}(\mathcal{H}) = d$ there are two steps:

1. show that there exists a set of d points that can be shattered by $\mathcal{H}$
   ➔ $\mathrm{VC}(\mathcal{H}) \geq d$
2. show that there does NOT exist a set of $d + 1$ points that can be shattered by $\mathcal{H}$
   ➔ $\mathrm{VC}(\mathcal{H}) < d + 1$

Claim: If $\mathcal{H}$ = {all linear separators in M dimensions} then $VC(\mathcal{H}) = M + 1$

VC-dimension

Proof Technique: To **prove** that $\mathrm{VC}(\mathcal{H}) = d$ there are two steps:

1. show that there exists a set of d points that can be shattered by $\mathcal{H}$
   ➔ $\mathrm{VC}(\mathcal{H}) \geq d$
2. show that there does NOT exist a set of $d + 1$ points that can be shattered by $\mathcal{H}$
   ➔ $\mathrm{VC}(\mathcal{H}) < d + 1$

VC-dimension Example: linear separators in 2D

To show (1):

Pick a dataset S (unlabeled) of size d

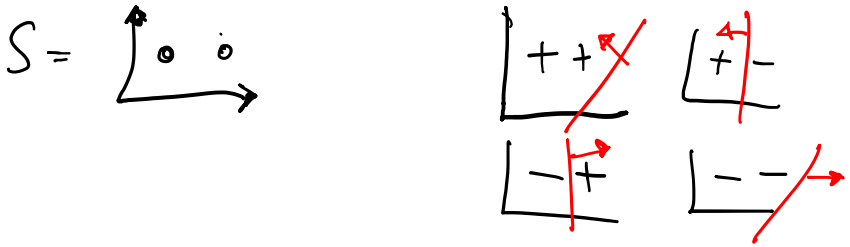
List all labelings of S

Show that $\mathcal{H}$ shatters S

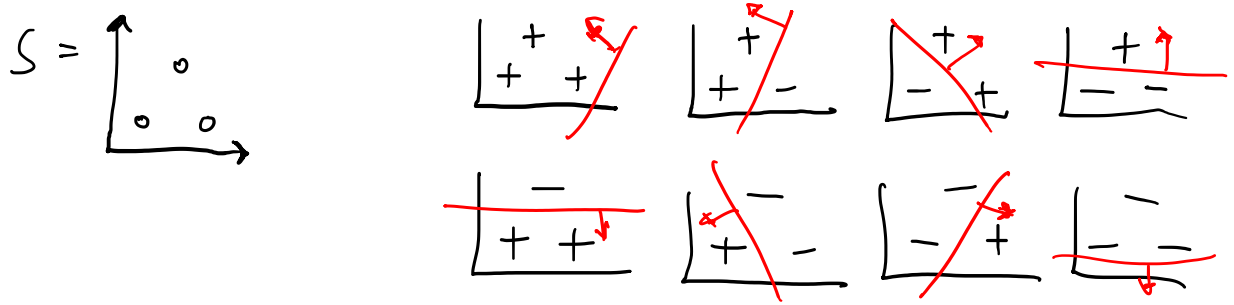
$\Rightarrow VC(\mathcal{H}) \geq 3$

$d = 1$

$d = 2$



$d = 3$



VC-dimension

Proof Technique: To **prove** that $\mathrm{VC}(\mathcal{H}) = d$ there are two steps:

1. show that there exists a set of d points that can be shattered by $\mathcal{H}$
   ➔ $\mathrm{VC}(\mathcal{H}) \geq d$
2. show that there does NOT exist a set of $d+1$ points that can be shattered by $\mathcal{H}$
   ➔ $\mathrm{VC}(\mathcal{H}) < d+1$

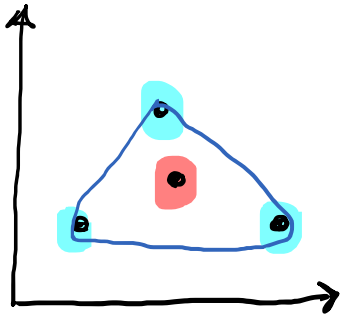
VC-dimension Example: linear separators in 2D

To show (2):

Divide all datasets of size 4 into ~~two~~ three categories

⇒ $\mathrm{VC}(\mathcal{H}) < 4$

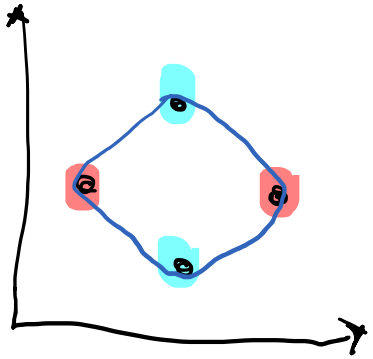
Case 1



some point is inside the convex hull

$\mathcal{H}$ does not shatter S

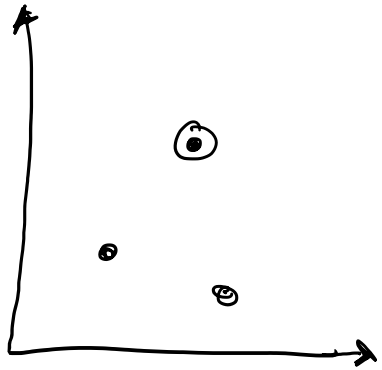
Case 2



no point is inside convex hull

$\mathcal{H}$ does not shatter S

Case 3



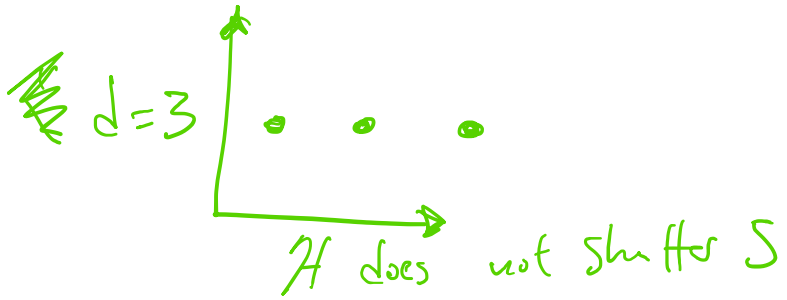
some points are overlapping

$\mathcal{H}$ does not shatter S

(1) + (2)

⇒ $\mathrm{VC}(\mathcal{H}) = 3$

∃ vs. ∀



VC-dimension

- Proving **VC-dimension** requires us to show that **there exists** (∃) a dataset of size d that can be shattered and that **there does not exist** (∄) a dataset of size d+1 that can be shattered

Shattering

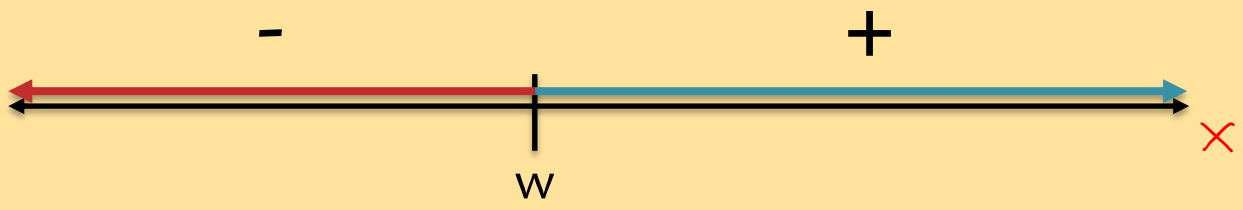
- Proving that a particular dataset can be **shattered** requires us to show that **for all** (∀) labelings of the dataset, our hypothesis class contains a hypothesis that can correctly classify it

VC-dimension Examples

- *Definition*: If VC(H) = d, then **there exists** (∃) a dataset of size d that can be shattered and that **there does not exist** (∄) a dataset of size d+1 that can be shattered

Question: 1

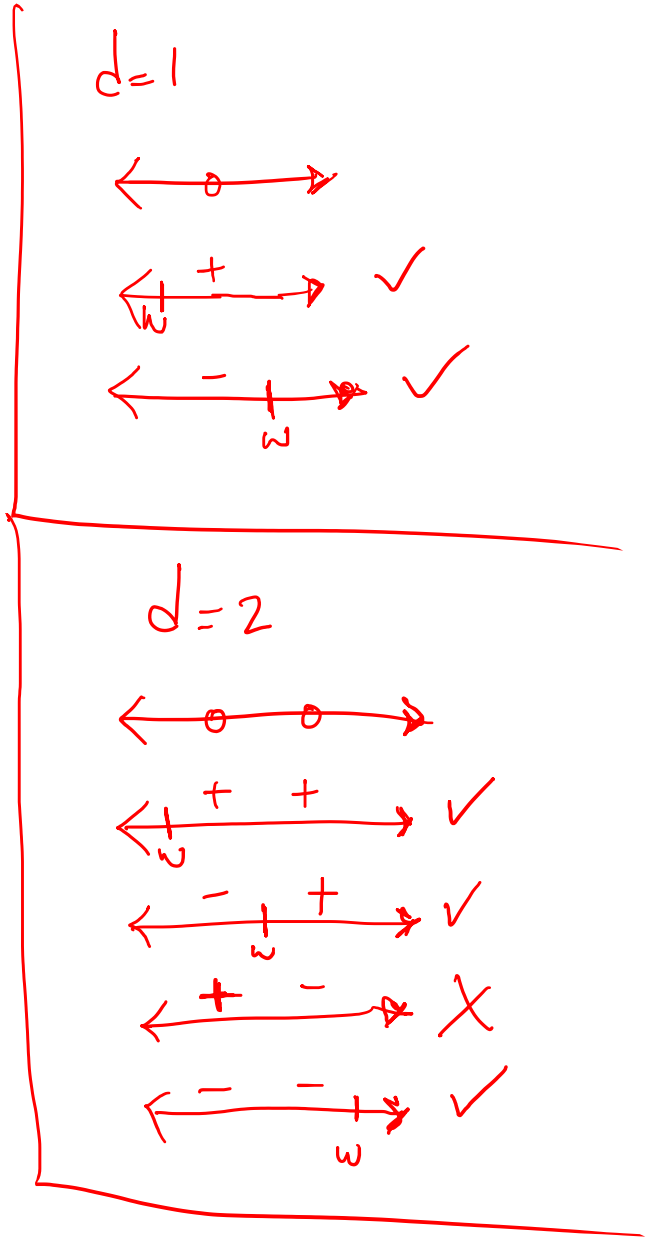
What is the VC-dimension of H = **1D positive rays**. That is for a threshold w, everything to the right of w is labeled as +1, everything else is labeled -1.



Answer: 46% 46%

A = 0 B = 1 C = 1.5 D = 2 E = 3 F = 4

toxic

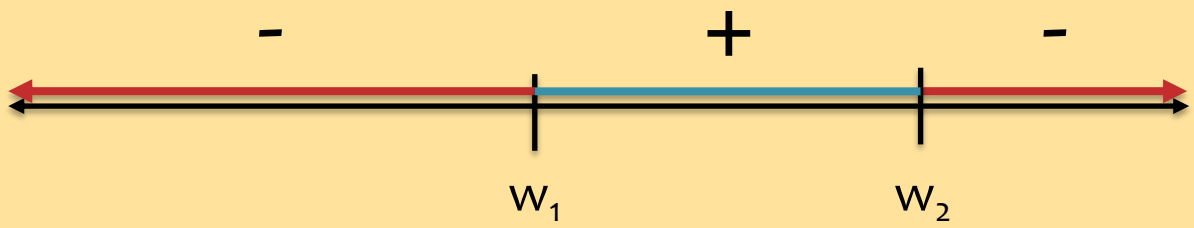


VC-dimension Examples

- *Definition*: If VC(H) = d, then **there exists** (∃) a dataset of size d that can be shattered and that **there does not exist** (∄) a dataset of size d+1 that can be shattered

Question:

What is the VC-dimension of H = **1D positive intervals**. That is for an interval (w_1, w_2), everything inside the interval is labeled as +1, everything else is labeled -1.



Answer:

A = 0 B = 1 C = 1.5 D = 2 E = 3 F = 4

Sample Complexity Results

Four Cases we care about…

	Realizable	Agnostic
Finite $\lvert\mathcal{H}\rvert$	**Thm. 1** $N \geq \frac{1}{\epsilon}\left[\log(\lvert\mathcal{H}\rvert) + \log(\frac{1}{\delta})\right]$ labeled examples are sufficient so that with probability $(1-\delta)$ all $h \in \mathcal{H}$ with $\hat{R}(h) = 0$ have $R(h) \leq \epsilon$.	**Thm. 2** $N \geq \frac{1}{2\epsilon^2}\left[\log(\lvert\mathcal{H}\rvert) + \log(\frac{2}{\delta})\right]$ labeled examples are sufficient so that with probability $(1-\delta)$ for all $h \in \mathcal{H}$ we have that $\lvert R(h) - \hat{R}(h)\rvert \leq \epsilon$.
Infinite $\lvert\mathcal{H}\rvert$	**Thm. 3** $N = O(\frac{1}{\epsilon}\left[\text{VC}(\mathcal{H})\log(\frac{1}{\epsilon}) + \log(\frac{1}{\delta})\right])$ labeled examples are sufficient so that with probability $(1-\delta)$ all $h \in \mathcal{H}$ with $\hat{R}(h) = 0$ have $R(h) \leq \epsilon$.	**Thm. 4** $N = O(\frac{1}{\epsilon^2}\left[\text{VC}(\mathcal{H}) + \log(\frac{1}{\delta})\right])$ labeled examples are sufficient so that with probability $(1-\delta)$ for all $h \in \mathcal{H}$ we have that $\lvert R(h) - \hat{R}(h)\rvert \leq \epsilon$.

SLT-STYLE COROLLARIES

SLT-style Corollaries

Thm. 1 $N \geq \frac{1}{\epsilon}\left[\log(|\mathcal{H}|) + \log(\frac{1}{\delta})\right]$ labeled examples are sufficient so that with probability $(1-\delta)$ all $h \in \mathcal{H}$ with $\hat{R}(h) = 0$ have $R(h) \leq \epsilon$.

Solve the inequality in Thm.1 for epsilon to obtain Corollary 1

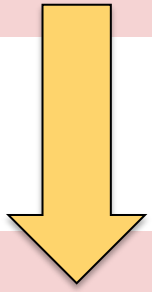
Corollary 1 (Realizable, Finite $|\mathcal{H}|$). For some $\delta > 0$, with probability at least $(1-\delta)$, for any h in $\mathcal{H}$ consistent with the training data (i.e. $\hat{R}(h) = 0$),

$$R(h) \leq \frac{1}{N}\left[\ln(|\mathcal{H}|) + \ln\left(\frac{1}{\delta}\right)\right]$$

We can obtain similar corollaries for each of the theorems…

SLT-style Corollaries

Thm. 2 $N \geq \frac{1}{2\epsilon^2}\left[\log(|\mathcal{H}|) + \log(\frac{2}{\delta})\right]$ labeled examples are sufficient so that with probability $(1-\delta)$ for all $h \in \mathcal{H}$ we have that $|R(h) - \hat{R}(h)| \leq \epsilon$.

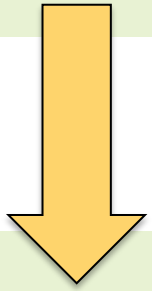


Corollary 2 (Agnostic, Finite $|\mathcal{H}|$). For some $\delta > 0$, with probability at least $(1-\delta)$, for all hypotheses h in $\mathcal{H}$,

$$R(h) \leq \hat{R}(h) + \sqrt{\frac{1}{2N}\left[\ln(|\mathcal{H}|) + \ln\left(\frac{2}{\delta}\right)\right]}$$

SLT-style Corollaries

Thm. 3 $N = O\left(\frac{1}{\epsilon}\left[\text{VC}(\mathcal{H})\log\left(\frac{1}{\epsilon}\right) + \log\left(\frac{1}{\delta}\right)\right]\right)$ labeled examples are sufficient so that with probability $(1-\delta)$ all $h \in \mathcal{H}$ with $\hat{R}(h) = 0$ have $R(h) \leq \epsilon$.



Corollary 3 (Realizable, Infinite $|\mathcal{H}|$). For some $\delta > 0$, with probability at least $(1-\delta)$, for any hypothesis h in $\mathcal{H}$ consistent with the data (i.e. with $\hat{R}(h) = 0$),

$$R(h) \leq O\left(\frac{1}{N}\left[\text{VC}(\mathcal{H})\ln\left(\frac{N}{\text{VC}(\mathcal{H})}\right) + \ln\left(\frac{1}{\delta}\right)\right]\right) \qquad (1)$$

SLT-style Corollaries

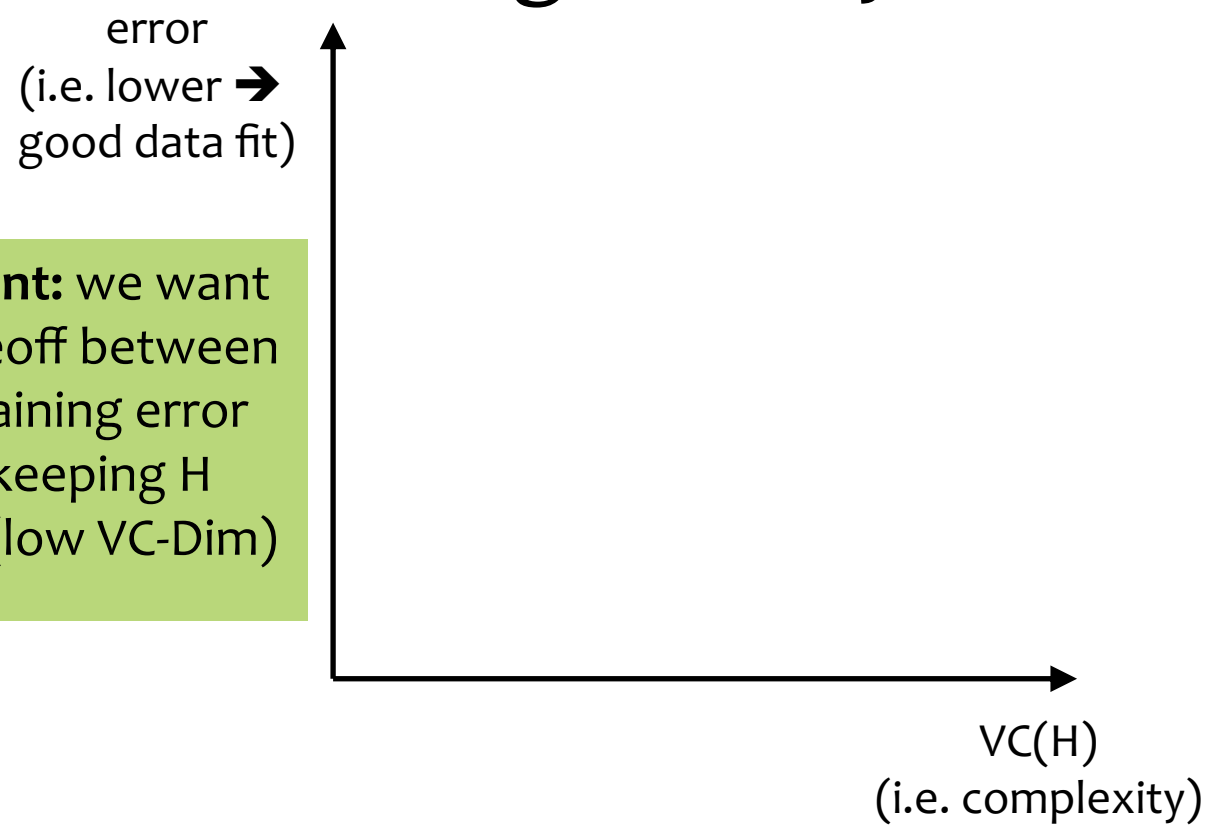
Thm. 4 $N = O(\frac{1}{\epsilon^2}\left[\text{VC}(\mathcal{H}) + \log(\frac{1}{\delta})\right])$ labeled examples are sufficient so that with probability $(1-\delta)$ for all $h \in \mathcal{H}$ we have that $|R(h) - \hat{R}(h)| \leq \epsilon$.

Corollary 4 (Agnostic, Infinite $|\mathcal{H}|$). For some $\delta > 0$, with probability at least $(1-\delta)$, for all hypotheses h in $\mathcal{H}$,

$$R(h) \leq \hat{R}(h) + O\left(\sqrt{\frac{1}{N}\left[\text{VC}(\mathcal{H}) + \ln\left(\frac{1}{\delta}\right)\right]}\right) \quad (2)$$

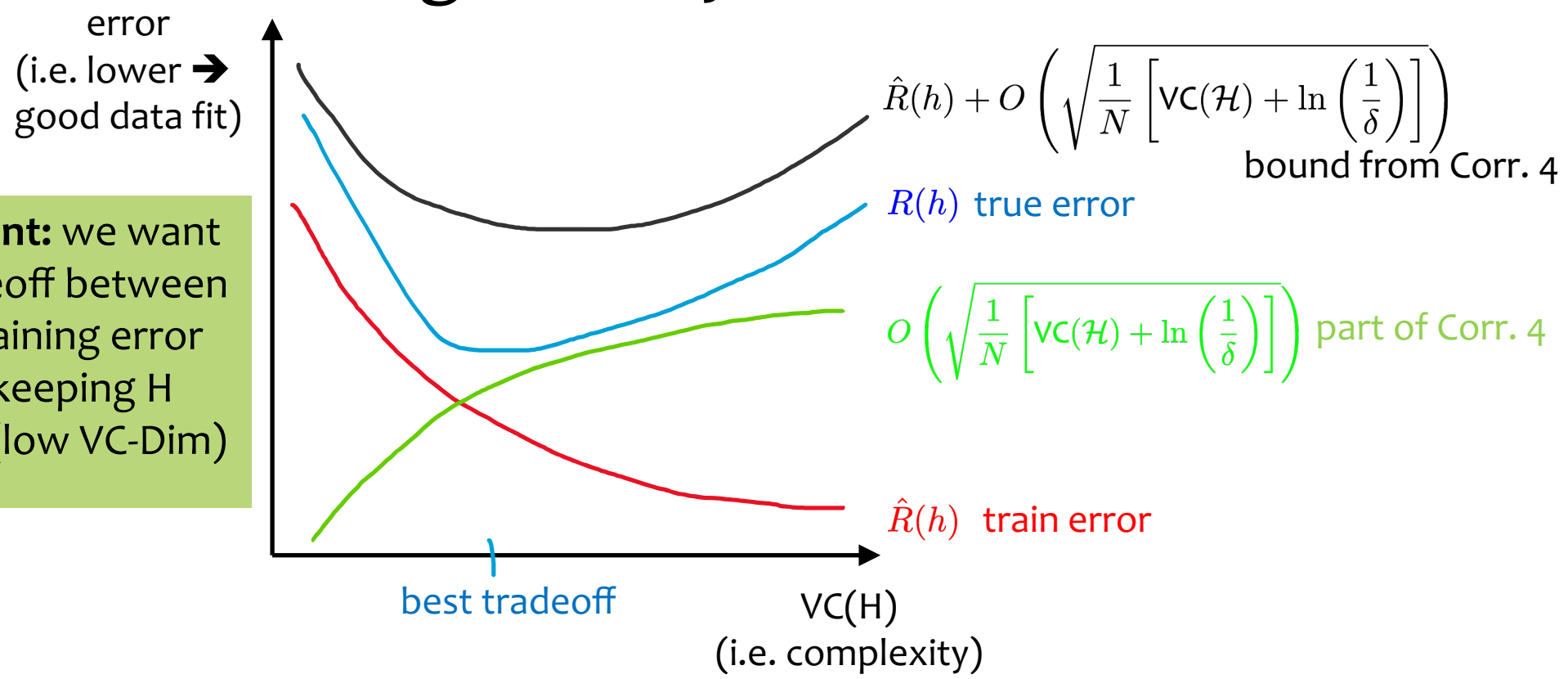
Should these corollaries inform how we do model selection?

Learning Theory & Model Selection



Key Point: we want to tradeoff between low training error and keeping H simple (low VC-Dim)

Learning Theory & Model Selection



Key Point: we want to tradeoff between low training error and keeping H simple (low VC-Dim)

Ex: H = Linear Separators in R^M

VC(H) = M+1

Q: In practice, how do we tradeoff between error and VC(H)?

A: Use a regularizer! That is, reducing the number of (effective) features reduces the VC dimension. More features usually leads to a better fit to the data.

Learning Theory Objectives

You should be able to…

- Identify the properties of a learning setting and assumptions required to ensure low generalization error
- Distinguish true error, train error, test error
- Define PAC and explain what it means to be approximately correct and what occurs with high probability
- Apply sample complexity bounds to real-world learning examples
- Distinguish between a large sample and a finite sample analysis
- Theoretically motivate regularization

PROBABILISTIC LEARNING

Probabilistic Learning

Function Approximation

Previously, we assumed that our output was generated using a **deterministic target function:**

$$\mathbf{x}^{(i)} \sim p^*(\cdot)$$

$$y^{(i)} = c^*(\mathbf{x}^{(i)})$$

Our goal was to learn a hypothesis h(**x**) that best approximates c^*(**x**)

Probabilistic Learning

Today, we assume that our output is **sampled** from a conditional **probability distribution:**

$$\mathbf{x}^{(i)} \sim p^*(\cdot)$$

$$y^{(i)} \sim p^*(\cdot|\mathbf{x}^{(i)})$$

Our goal is to learn a probability distribution p(y|**x**) that best approximates p^*(y|**x**)

MAXIMUM LIKELIHOOD ESTIMATION (MLE)

Likelihood Function

One R.V.

- Given N **independent, identically distributed (iid)** samples $D = \{x^{(1)}, x^{(2)}, \ldots, x^{(N)}\}$ from a **random variable** X …

- The **likelihood** function is
  - Case 1: X is **discrete** with probability mass function (*pmf*) $p(x|\theta)$
    $$L(\theta) = p(x^{(1)}|\theta)\, p(x^{(2)}|\theta) \ldots\ p(x^{(N)}|\theta)$$
  - Case 2: X is **continuous** with probability density function (pdf) $f(x|\theta)$
    $$L(\theta) = f(x^{(1)}|\theta)\, f(x^{(2)}|\theta) \ldots\ f(x^{(N)}|\theta)$$

The **likelihood** tells us how likely one sample is relative to another

- The **log-likelihood** function is
  - Case 1: X is **discrete** with probability mass function (*pmf*) $p(x|\theta)$
    $$\ell(\theta) = \log p(x^{(1)}|\theta) + \ldots + \log p(x^{(N)}|\theta)$$
  - Case 2: X is **continuous** with probability density function (pdf) $f(x|\theta)$
    $$\ell(\theta) = \log f(x^{(1)}|\theta) + \ldots + \log f(x^{(N)}|\theta)$$

Likelihood Function

Two R.V.s

- Given N **iid** samples $D = \{(x^{(1)}, y^{(1)}), \ldots, (x^{(N)}, y^{(N)})\}$ from a pair of **random variables** X, Y

- The **conditional likelihood** function:
  - Case 1: Y is **discrete** with *pmf* $p(y \mid x, \theta)$
    $$L(\theta) = p(y^{(1)} \mid x^{(1)}, \theta) \ldots p(y^{(N)} \mid x^{(N)}, \theta)$$
  - Case 2: Y is **continuous** with *pdf* $f(y \mid x, \theta)$
    $$L(\theta) = f(y^{(1)} \mid x^{(1)}, \theta) \ldots f(y^{(N)} \mid x^{(N)}, \theta)$$

- The **joint likelihood** function:
  - Case 1: X and Y are **discrete** with *pmf* $p(x,y|\theta)$
    $$L(\theta) = p(x^{(1)}, y^{(1)}|\theta) \ldots p(x^{(N)}, y^{(N)}|\theta)$$
  - Case 2: X and Y are **continuous** with *pdf* $f(x,y|\theta)$
    $$L(\theta) = f(x^{(1)}, y^{(1)}|\theta) \ldots f(x^{(N)}, y^{(N)}|\theta)$$

Likelihood Function

Two R.V.s

- Given N **iid** samples $D = \{(x^{(1)}, y^{(1)}), \ldots, (x^{(N)}, y^{(N)})\}$ from a pair of **random variables** X, Y

- The **joint likelihood** function:
  - Case 1: X and Y are **discrete** with *pmf* $p(x,y|\theta)$
    $$L(\theta) = p(x^{(1)}, y^{(1)}|\theta) \ldots p(x^{(N)}, y^{(N)}|\theta)$$
  - Case 2: X and Y are **continuous** with *pdf* $f(x,y|\theta)$
    $$L(\theta) = f(x^{(1)}, y^{(1)}|\theta) \ldots f(x^{(N)}, y^{(N)}|\theta)$$
  - Case 3: Y is **discrete** with *pmf* $p(y|\beta)$ and X is **continuous** with *pdf* $f(x|y,\alpha)$
    $$L(\alpha, \beta) = f(x^{(1)}|\, y^{(1)}, \alpha)\, p(y^{(1)}|\beta) \ldots f(x^{(N)}|\, y^{(N)}, \alpha)\, p(y^{(N)}|\beta)$$
  - Case 4: Y is **continuous** with *pdf* $f(y|\beta)$ and X is **discrete** with *pmf* $p(x|y,\alpha)$
    $$L(\alpha, \beta) = p(x^{(1)}|\, y^{(1)}, \alpha)\, f(y^{(1)}|\beta) \ldots p(x^{(N)}|\, y^{(N)}, \alpha)\, f(y^{(N)}|\beta)$$

Mixed discrete/ continuous!

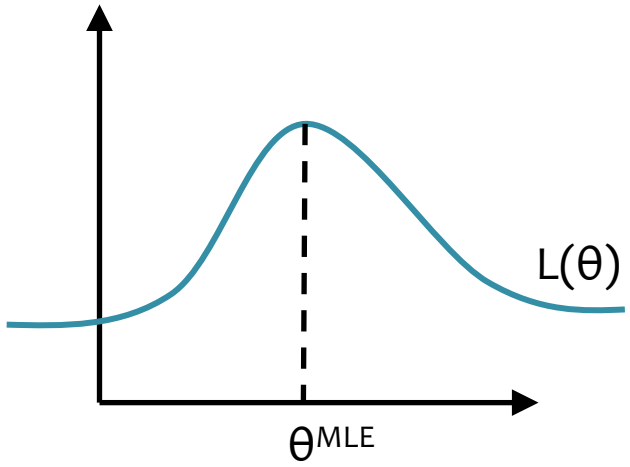
MLE

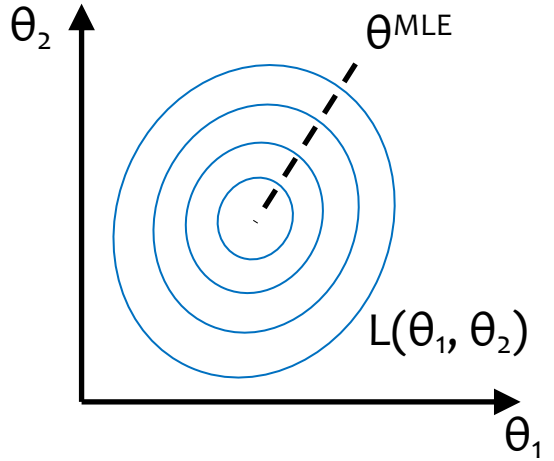
Suppose we have data $\mathcal{D} = \{x^{(i)}\}_{i=1}^{N}$

Principle of Maximum Likelihood Estimation:
Choose the parameters that maximize the likelihood of the data.

$$\boldsymbol{\theta}^{\text{MLE}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \prod_{i=1}^{N} p(\mathbf{x}^{(i)}|\boldsymbol{\theta})$$

Maximum Likelihood Estimate (MLE)





MLE

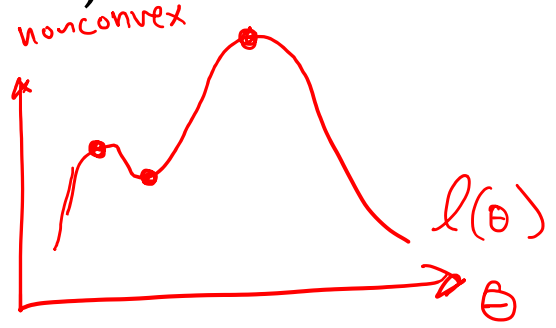
What does maximizing likelihood accomplish?

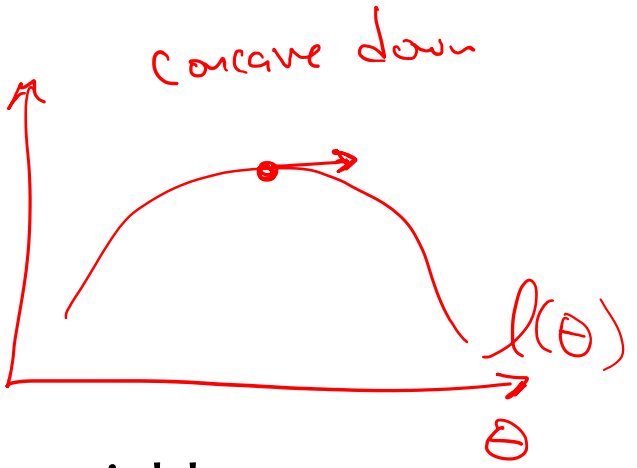
- There is only a finite amount of probability mass (i.e. sum-to-one constraint)
- MLE tries to allocate **as much** probability mass **as possible** to the things we have observed…

  …**at the expense** of the things we have **not** observed

Recipe for Closed-form MLE

1. Assume data was generated iid from some model, i.e., write the *generative story*
   $x^{(i)} \sim p(x|\boldsymbol{\theta})$
2. Write the log-likelihood
   $\ell(\boldsymbol{\theta}) = \log p(x^{(1)}|\boldsymbol{\theta}) + \ldots + \log p(x^{(N)}|\boldsymbol{\theta})$
3. Compute partial derivatives, i.e., the gradient
   $\partial\ell(\boldsymbol{\theta})/\partial\theta_1 = \ldots$
   $\ldots$
   $\partial\ell(\boldsymbol{\theta})/\partial\theta_M = \ldots$
4. Set derivatives equal to zero and solve for $\boldsymbol{\theta}$
   $\partial\ell(\boldsymbol{\theta})/\partial\theta_m = 0$ for all $m \in \{1, \ldots, M\}$
   $\boldsymbol{\theta}^{MLE}$ = solution to system of M equations and M variables
5. Compute the second derivative and check that $\ell(\boldsymbol{\theta})$ is concave down ~~at~~ everywhere $\boldsymbol{\theta}^{MLE}$





MLE EXAMPLES

MLE of Exponential Distribution

Goal:

- pdf of Exponential(λ): $f(x) = \lambda e^{-\lambda x}$
- Suppose $X_i \sim$ Exponential(λ) for $1 \leq i \leq N$.
- Find MLE for data $\mathcal{D} = \{x^{(i)}\}_{i=1}^{N}$

Steps:

- First write down log-likelihood of sample.
- Compute first derivative, set to zero, solve for λ.
- Compute second derivative and check that it is concave down at λ^{MLE}.

MLE of Exponential Distribution

- pdf of Exponential(λ): $f(x) = \lambda e^{-\lambda x}$
- Suppose $X_i \sim$ Exponential(λ) for $1 \leq i \leq N$.
- Find MLE for data $\mathcal{D} = \{x^{(i)}\}_{i=1}^{N}$

- First write down log-likelihood of sample.

$$\ell(\lambda) = \sum_{i=1}^{N} \log f(x^{(i)}) \quad (1)$$

$$= \sum_{i=1}^{N} \log(\lambda \exp(-\lambda x^{(i)})) \quad (2)$$

$\log(ab) = \log(a) + \log(b)$

$$= \sum_{i=1}^{N} \log(\lambda) + (-\lambda x^{(i)}) \quad (3)$$

$$= N \log(\lambda) - \lambda \sum_{i=1}^{N} x^{(i)} \quad (4)$$

MLE of Exponential Distribution

- pdf of Exponential(λ): $f(x) = \lambda e^{-\lambda x}$
- Suppose $X_i \sim$ Exponential(λ) for $1 \leq i \leq N$.
- Find MLE for data $\mathcal{D} = \{x^{(i)}\}_{i=1}^{N}$

- Compute first derivative, set to zero, solve for λ.

$\ell(\lambda)$

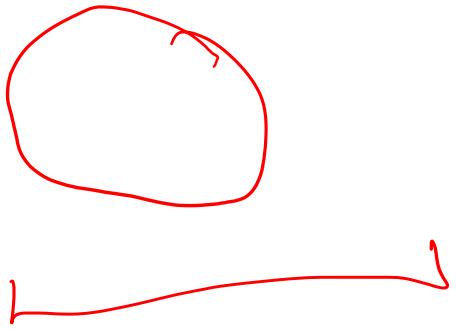
$$\frac{d\ell(\lambda)}{d\lambda} = \frac{d}{d\lambda} N \log(\lambda) - \lambda \sum_{i=1}^{N} x^{(i)} \quad (1)$$

$$= \frac{N}{\lambda} - \sum_{i=1}^{N} x^{(i)} = 0 \quad (2)$$

$$\Rightarrow \lambda^{\text{MLE}} = \frac{N}{\sum_{i=1}^{N} x^{(i)}} \quad (3)$$

MLE of Bernoulli Model

1. Model: $\mathbf{x}^{(i)} \sim \text{Bernoulli}(\phi)$ for $i = 1, \ldots, N$



likelihood

2. Log-~~posterior:~~

3. Derivative:

$$\ell_0 \frac{1}{1-\phi} \frac{\delta}{\delta\phi}(1-\phi)$$

-1

4. Set to zero and solve:

MLE of Bernoulli Model

1. Model: $\mathbf{x}^{(i)} \sim \text{Bernoulli}(\phi)$ for $i = 1, \ldots, N$

$$p(x^{(i)}|\phi) = \begin{cases} \phi & \text{if } x^{(i)} = 1 \\ (1-\phi) & \text{if } x^{(i)} = 0 \end{cases}$$
$$= \phi^{x^{(i)}}(1-\phi)^{1-x^{(i)}}$$

2. Log-posterior:

$$\ell_{\text{MLE}}(\phi) = \log p(\mathcal{D} \mid \phi)$$
$$= \log \prod_{i=1}^{N} p(x^{(i)} \mid \phi)$$
$$= \log \prod_{i=1}^{N} \phi^{x^{(i)}}(1-\phi)^{1-x^{(i)}}$$
$$= \log \left(\phi^{N_1}(1-\phi)^{N_0}\right)$$
$$= N_1 \log(\phi) + N_0 \log(1-\phi)$$

$$N_1 = \#(x^{(i)} = 1)$$
$$N_0 = \#(x^{(i)} = 0)$$

3. Derivative:

$$\frac{\partial \ell_{\text{MLE}}(\phi)}{\partial \phi} = \frac{\partial}{\partial \phi}\left(N_1 \log(\phi) + N_0 \log(1-\phi)\right)$$
$$= \frac{N_1}{\phi} - \frac{N_0}{1-\phi}$$

4. Set to zero and solve:

$$\frac{N_1}{\phi} - \frac{N_0}{1-\phi} = 0$$
$$\Rightarrow \phi_{\text{MLE}} = \frac{N_1}{N_1 + N_0} = \frac{N_1}{N}$$

Estimator vs. Estimate

often ignored

Maximum likelihood ESTIMATOR

- the formula used to compute the estimate with some data

$$\phi_{\text{MLE}} = \frac{N_1}{N_1 + N_0} = \frac{N_1}{N}$$

Maximum likelihood ESTIMATE

- plugs data into the estimator to compute an actual number

$x^{(1)} = 1$

$x^{(2)} = 0$

$x^{(3)} = 1$

$x^{(4)} = 1$

$\phi_{MLE} = \frac{3}{4}$