10-301/601: Introduction to Machine Learning
Lecture 16 – Learning Theory (Infinite Case)

Henry Chai
10/26/22
Let $\mathcal{H}$ be the set of all conjunctions over $M$ Boolean variables, $x \in \{0,1\}^M$; examples of conjunctions are

- $h(x) = x_1(1 - x_2)x_4x_{10}$
- $h(x) = (1 - x_3)(1 - x_4)x_8$

Assuming $c^* \in \mathcal{H}$, if $M = 10$, $\varepsilon = 0.1$, and $\delta = 0.01$, at least how many labelled examples do we need to satisfy the PAC criterion using Theorem 1?

A. 1 (TOXIC)
B. $10(2 \ln 10 + \ln 100) \approx 92$
C. $10(3 \ln 10 + \ln 100) \approx 116$
D. $10(10 \ln 2 + \ln 100) \approx 116$
E. $10(10 \ln 3 + \ln 100) \approx 156$
F. $100(2 \ln 10 + \ln 10) \approx 691$
G. $100(3 \ln 10 + \ln 10) \approx 922$
H. $100(10 \ln 2 + \ln 10) \approx 924$
I. $100(10 \ln 3 + \ln 10) \approx 1329$

Q & A:
Why is the answer C?
Great question, it’s not! It’s E (my bad)
Q & A:
How does the statistical learning theory corollary follow from this theorem?

• For a finite hypothesis set $\mathcal{H}$ s.t. $c^* \in \mathcal{H}$ and arbitrary distribution $p^*$, if the number of labelled training data points satisfies

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

then with probability at least $1 - \delta$, all $h \in \mathcal{H}$ with \( \hat{R}(h) = 0 \) have $R(h) \leq \epsilon$
For a finite hypothesis set $\mathcal{H}$ s.t. $c^* \in \mathcal{H}$ and arbitrary distribution $p^*$, if the number of labelled training data points satisfies

$$M = \frac{1}{\epsilon} \left( \ln(|\mathcal{H}|) + \ln \left( \frac{1}{\delta} \right) \right)$$

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

• Solving for $\epsilon$ gives...
Q & A:

How does the statistical learning theory corollary follow from this theorem?

- For a finite hypothesis set $\mathcal{H}$ s.t. $c^* \in \mathcal{H}$ and arbitrary distribution $p^*$, given a training data set $S$ s.t. $|S| = M$, all $h \in \mathcal{H}$ with $\hat{R}(h) = 0$ have

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

with probability at least $1 - \delta$. 
· Announcements
  · HW5 released 10/13, due 10/27 (tomorrow) at 11:59 PM
  · HW6 released 10/27 (tomorrow), due 11/4 at 11:59 PM
    · Only two late days allowed on HW6
  · Exam 2 on 11/10, two weeks from tomorrow (more details to follow)
    · All topics between Lecture 8 and Lecture 17 (next Monday’s lecture) are in-scope
    · Exam 1 content may be referenced but will not be the primary focus of any question
  · Exam 3 scheduled
    · Thursday, December 15th from 9:30 AM to 11:30 AM
  · Sign up for peer tutoring! See Piazza for more details
Recall - Theorem 1: Finite, Realizable Case

- For a finite hypothesis set $\mathcal{H}$ s.t. $c^* \in \mathcal{H}$ and arbitrary distribution $p^*$, if the number of labelled training data points satisfies

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

then with probability at least $1 - \delta$, all $h \in \mathcal{H}$ with $\hat{R}(h) = 0$ have $R(h) \leq \epsilon$
Recall - Theorem 2: Finite, Agnostic Case

- For a finite hypothesis set $\mathcal{H}$ and arbitrary distribution $p^*$, if the number of labelled training data points satisfies
  \[
  M \geq \frac{1}{2\varepsilon^2} \left( \ln(|\mathcal{H}|) + \ln \left( \frac{2}{\delta} \right) \right)
  \]
  then with probability at least $1 - \delta$, all $h \in \mathcal{H}$ satisfy
  \[
  |R(h) - \hat{R}(h)| \leq \varepsilon
  \]
  - Bound is inversely quadratic in $\varepsilon$, e.g., halving $\varepsilon$ means we need four times as many labelled training data points
What happens when $|\mathcal{H}| = \infty$?

• For a finite hypothesis set $\mathcal{H}$ and arbitrary distribution $p^*$, if the number of labelled training data points satisfies

$$M \geq \frac{1}{2\epsilon^2} \left( \ln(|\mathcal{H}|) + \ln \left( \frac{2}{\delta} \right) \right)$$

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

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

• Insight: $|\mathcal{H}|$ measures how complex our hypothesis set is

• Idea: define a different measure of hypothesis set complexity
Labellings

- Given some finite set of data points $S = (x^{(1)}, ..., x^{(M)})$ and some hypothesis $h \in \mathcal{H}$, applying $h$ to each point in $S$ results in a **labelling**
  - $\left(h(x^{(1)}), ..., h(x^{(M)})\right)$ is a vector of $M$ +1’s and -1’s
- **Important note:** our discussion of PAC learning assumes binary classification

- Given $S = (x^{(1)}, ..., x^{(M)})$, 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
  $$\mathcal{H}(S) = \left\{ \left(h(x^{(1)}), ..., h(x^{(M)})\right) \mid h \in \mathcal{H} \right\}$$
Example: Labellings

\[ \mathcal{H} = \{ h_1, h_2, h_3 \} \]
Example: Labellings

\[ \mathcal{H} = \{ h_1, h_2, h_3 \} \]

\[
\left( h_1(x^{(1)}), h_1(x^{(2)}), h_1(x^{(3)}), h_1(x^{(4)}) \right)
= (-1, +1, -1, +1)
\]
Example: Labellings

\[ \mathcal{H} = \{h_1, h_2, h_3\} \]

\[ (h_2(x^{(1)}), h_2(x^{(2)}), h_2(x^{(3)}), h_2(x^{(4)})) = (-1, +1, -1, +1) \]
Example: Labellings

\[ \mathcal{H} = \{ h_1, h_2, h_3 \} \]

\[ (h_3(x^{(1)}), h_3(x^{(2)}), h_3(x^{(3)}), h_3(x^{(4)})) \]

\[ = (+1, +1, -1, -1) \]
Example: Labellings

\[ \mathcal{H} = \{ h_1, h_2, h_3 \} \]

\[ \mathcal{H}(S) = \{ (+1, +1, -1, -1), (-1, +1, -1, +1) \} \]

\[ |\mathcal{H}(S)| = 2 \]
Example: Labellings

\( \mathcal{H} = \{h_1, h_2, h_3\} \)

\( \mathcal{H}(S) = \{(+1, +1, -1, -1)\} \)

\( |\mathcal{H}(S)| = 1 \)
\( \mathcal{H}(S) \) is the set of all labellings induced by \( \mathcal{H} \) on \( S \)

- If \( |S| = M \), then \( |\mathcal{H}(S)| \leq 2^M \)
- \( \mathcal{H} \) shatters \( S \) if \( |\mathcal{H}(S)| = 2^M \)

The **VC-dimension** of \( \mathcal{H}, VC(\mathcal{H}) \), is the size of the largest set \( S \) that can be shattered by \( \mathcal{H} \).

- If \( \mathcal{H} \) can shatter arbitrarily large finite sets, then \( d_{VC}(\mathcal{H}) = \infty \)

To prove that \( VC(\mathcal{H}) = d \), you need to show

1. \( \exists \) some set of \( d \) data points that \( \mathcal{H} \) can shatter and
2. \( \not\exists \) a set of \( d + 1 \) data points that \( \mathcal{H} \) can shatter
• \( x \in \mathbb{R}^2 \) and \( \mathcal{H} = \) all 2-dimensional linear separators

• What is \( VC(\mathcal{H}) \)?
  • Can \( \mathcal{H} \) shatter some set of 1 point?
• $x \in \mathbb{R}^2$ and $\mathcal{H} =$ all 2-dimensional linear separators

• What is $VC(\mathcal{H})$?
  • Can $\mathcal{H}$ shatter some set of 1 point?
  • Can $\mathcal{H}$ shatter some set of 2 points?
• $x \in \mathbb{R}^2$ and $\mathcal{H} =$ all 2-dimensional linear separators

• What is $\text{VC}(\mathcal{H})$?
  • Can $\mathcal{H}$ shatter some set of 1 point?
  • Can $\mathcal{H}$ shatter some set of 2 points?
  • Can $\mathcal{H}$ shatter some set of 3 points?
• $x \in \mathbb{R}^2$ and $\mathcal{H} =$ all 2-dimensional linear separators

• What is $\text{VC}(\mathcal{H})$?
  • Can $\mathcal{H}$ shatter some set of 1 point?
  • Can $\mathcal{H}$ shatter some set of 2 points?
  • Can $\mathcal{H}$ shatter some set of 3 points?
• $\mathbf{x} \in \mathbb{R}^2$ and $\mathcal{H} =$ all 2-dimensional linear separators

• What is $VC(\mathcal{H})$?
  • Can $\mathcal{H}$ shatter some set of 1 point?
  • Can $\mathcal{H}$ shatter some set of 2 points?
  • Can $\mathcal{H}$ shatter some set of 3 points?

VC-Dimension: Example

$S$
• $x \in \mathbb{R}^2$ and $\mathcal{H} =$ all 2-dimensional linear separators

• What is $VC(\mathcal{H})$?
  • Can $\mathcal{H}$ shatter some set of 1 point?
  • Can $\mathcal{H}$ shatter some set of 2 points?
  • Can $\mathcal{H}$ shatter *some* set of 3 points?
• $x \in \mathbb{R}^2$ and $\mathcal{H} =$ all 2-dimensional linear separators

• What is $\text{VC}(\mathcal{H})$?
  • Can $\mathcal{H}$ shatter some set of 1 point?
  • Can $\mathcal{H}$ shatter some set of 2 points?
  • Can $\mathcal{H}$ shatter some set of 3 points?

$S_1$

$S_2$
• $x \in \mathbb{R}^2$ and $\mathcal{H}$ = all 2-dimensional linear separators

• What is $VC(\mathcal{H})$?
  • Can $\mathcal{H}$ shatter some set of 1 point?
  • Can $\mathcal{H}$ shatter some set of 2 points?
  • Can $\mathcal{H}$ shatter some set of 3 points?

$|\mathcal{H}(S_1)| = 6$

$|\mathcal{H}(S_2)| = 8$
\( x \in \mathbb{R}^2 \) and \( \mathcal{H} = \) all 2-dimensional linear separators

• What is \( VC(\mathcal{H}) \)?
  • Can \( \mathcal{H} \) shatter some set of 1 point?
  • Can \( \mathcal{H} \) shatter some set of 2 points?
  • Can \( \mathcal{H} \) shatter some set of 3 points?
  • Can \( \mathcal{H} \) shatter some set of 4 points?

\( S_1 \)

All points on the convex hull

\( S_2 \)

At least one point inside the convex hull
- $x \in \mathbb{R}^2$ and $\mathcal{H} =$ all 2-dimensional linear separators

- What is $\text{VC}(\mathcal{H})$?
  - Can $\mathcal{H}$ shatter some set of 1 point?
  - Can $\mathcal{H}$ shatter some set of 2 points?
  - Can $\mathcal{H}$ shatter some set of 3 points?
  - Can $\mathcal{H}$ shatter some set of 4 points?

VC-Dimension: Example

$S_1$: All points on the convex hull

$S_2$: At least one point inside the convex hull
- $x \in \mathbb{R}^2$ and $\mathcal{H} =$ all 2-dimensional linear separators

- What is $\text{VC}(\mathcal{H})$?
  - Can $\mathcal{H}$ shatter some set of 1 point?
  - Can $\mathcal{H}$ shatter some set of 2 points?
  - Can $\mathcal{H}$ shatter some set of 3 points?
  - Can $\mathcal{H}$ shatter some set of 4 points?

All points on the convex hull

At least one point inside the convex hull
• $x \in \mathbb{R}^2$ and $\mathcal{H} =$ all 2-dimensional linear separators

• What is $VC(\mathcal{H})$?
  • Can $\mathcal{H}$ shatter some set of 1 point?
  • Can $\mathcal{H}$ shatter some set of 2 points?
  • Can $\mathcal{H}$ shatter some set of 3 points?
  • Can $\mathcal{H}$ shatter some set of 4 points?

$|\mathcal{H}(S_1)| = 14$
All points on the convex hull

$S_2$
At least one point inside the convex hull
• \( x \in \mathbb{R}^2 \) and \( \mathcal{H} = \) all 2-dimensional linear separators

• What is \( VC(\mathcal{H}) \)?
  • Can \( \mathcal{H} \) shatter some set of 1 point?
  • Can \( \mathcal{H} \) shatter some set of 2 points?
  • Can \( \mathcal{H} \) shatter some set of 3 points?
  • Can \( \mathcal{H} \) shatter some set of 4 points?

\[ |\mathcal{H}(S_1)| = 14 \]

All points on the convex hull

\[ S_2 \]

At least one point inside the convex hull
• \( x \in \mathbb{R}^2 \) and \( \mathcal{H} = \) all 2-dimensional linear separators

• What is \( VC(\mathcal{H}) \)?
  • Can \( \mathcal{H} \) shatter some set of 1 point?
  • Can \( \mathcal{H} \) shatter some set of 2 points?
  • Can \( \mathcal{H} \) shatter some set of 3 points?
  • Can \( \mathcal{H} \) shatter some set of 4 points?

\[
\left| \mathcal{H}(S_1) \right| = 14
\]

All points on the convex hull

\[ S_2 \]
At least one point inside the convex hull
• \( x \in \mathbb{R}^2 \) and \( \mathcal{H} = \) all 2-dimensional linear separators

• What is \( VC(\mathcal{H}) \)?
  • Can \( \mathcal{H} \) shatter some set of 1 point?
  • Can \( \mathcal{H} \) shatter some set of 2 points?
  • Can \( \mathcal{H} \) shatter some set of 3 points?
  • Can \( \mathcal{H} \) shatter some set of 4 points?

\( |\mathcal{H}(S_1)| = 14 \)
All points on the convex hull

\( |\mathcal{H}(S_2)| = 14 \)
At least one point inside the convex hull
• $x \in \mathbb{R}^2$ and $\mathcal{H} = \text{all 2-dimensional linear separators}$

• $VC(\mathcal{H}) = 3$
  • Can $\mathcal{H}$ shatter some set of 1 point?
  • Can $\mathcal{H}$ shatter some set of 2 points?
  • Can $\mathcal{H}$ shatter some set of 3 points?
  • Can $\mathcal{H}$ shatter some set of 4 points?

$|\mathcal{H}(S_1)| = 14$
All points on the convex hull

$|\mathcal{H}(S_2)| = 14$
At least one point inside the convex hull
• $x \in \mathbb{R}^2$ and $\mathcal{H} =$ all $d$-dimensional linear separators

• $VC(\mathcal{H}) = d + 1$
\begin{itemize}
\item $x \in \mathbb{R}$ and $\mathcal{H} = \text{all 1-dimensional positive rays, i.e., all hypotheses of the form } h(x; a) = \text{sign}(x - a)$
\end{itemize}
Poll Question 1:

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

A. -1 (TOXIC)
B. 0
C. 1
D. 2
E. 3

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

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

• $VC(\mathcal{H}) = 1$
\[ x \in \mathbb{R} \text{ and } \mathcal{H} = \text{all 1-dimensional positive intervals} \]
Poll Question 2:
What is $VC(\mathcal{H})$?

A. 0
B. 1
C. 1.5 (TOXIC)
D. 2
E. 3

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

\[ VC(\mathcal{H}) = 2 \]
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

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

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

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

with probability at least $1 - \delta$. 

10/26/22
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

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

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

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

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

with probability at least $1 - \delta$. 
Agnostic case: for any hypothesis class $\mathcal{H}$ and distribution $D$, given a training data set $S$, all $h \in \mathcal{H}$ have $R_h \leq \hat{R}(h) + O \left( \frac{1}{M} \left( VC(\mathcal{H}) + \log \left( \frac{1}{\delta} \right) \right) \right)$ with probability at least $1 - \delta$.

Approximation Generalization Tradeoff

- How well does $h$ approximate $c^*$?
- How well does $h$ generalize?
Agnostic case: for any hypothesis class $\mathcal{H}$ and distribution $D$, given a training data set $S$, all $h \in \mathcal{H}$ have $R(h) \leq \hat{R}(h) + O\left(\frac{1}{M} \left( VC(\mathcal{H}) + \log \left(\frac{1}{\delta}\right) \right) \right)$ with probability at least $1 - \delta$.

Approximation Generalization Tradeoff

- Increases as $VC(\mathcal{H})$ increases
- Decreases as $VC(\mathcal{H})$ increases
Can we use this corollary to guide model selection?

• 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

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

with probability at least $1 - \delta$. 
How can we find this “best tradeoff” for linear separators?

Use a regularizer! By (effectively) reducing the number of features our model considers, we reduce its VC-dimension.
How can we find this “best tradeoff” for linear separators?

- Use a regularizer! By (effectively) reducing the number of features our model considers, we reduce its VC-dimension.
Learning Theory

Learning 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 machine learning examples
- Theoretically motivate regularization
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 machine learning examples
- Theoretically motivate regularization
Recall: Probabilistic Learning

- Previously:
  - (Unknown) Target function, $c^*: \mathcal{X} \rightarrow \mathcal{Y}$
  - Classifier, $h: \mathcal{X} \rightarrow \mathcal{Y}$
  - Goal: find a classifier, $h$, that best approximates $c^*$

- Now:
  - (Unknown) Target distribution, $y \sim p^*(Y|x)$
  - Distribution, $p(Y|x)$
  - Goal: find a distribution, $p$, that best approximates $p^*$
Recall: Maximum Likelihood Estimation (MLE)

• Insight: every valid probability distribution has a finite amount of probability mass as it must sum/integrate to 1

• Idea: set the parameter(s) so that the likelihood of the samples is maximized

• Intuition: assign as much of the (finite) probability mass to the observed data at the expense of unobserved data

• Example: the exponential distribution

Source: https://en.wikipedia.org/wiki/Exponential_distribution#/media/File:Exponential_probability_density.svg
• A Bernoulli random variable takes value 1 with probability $\phi$ and value 0 with probability $1 - \phi$

• The pmf of the Bernoulli distribution is

$$p(x|\phi) = \phi^x (1 - \phi)^{1-x}$$
A Bernoulli random variable takes value 1 (or heads) with probability $\phi$ and value 0 (or tails) with probability $1 - \phi$.

The pmf of the Bernoulli distribution is

$$p(x|\phi) = \phi^x (1 - \phi)^{1-x}$$

Given $N$ iid samples $\{x^{(1)}, \ldots, x^{(N)}\}$, the log-likelihood is

$$\ell(\phi) = \sum_{i=1}^{N} \log p(x^{(i)}|\phi) = \sum_{i=1}^{N} \log \phi^{x^{(i)}} (1 - \phi)^{1-x^{(i)}}$$

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

$$= N_1 \log \phi + N_0 \log(1 - \phi)$$

where $N_1$ is the number of 1’s in $\{x^{(1)}, \ldots, x^{(N)}\}$ and $N_0$ is the number of 0’s.
• A Bernoulli random variable takes value 1 (or heads) with probability $\phi$ and value 0 (or tails) with probability $1 - \phi$

• The pmf of the Bernoulli distribution is
  $$p(x|\phi) = \phi^x (1 - \phi)^{1-x}$$

• The partial derivative of the log-likelihood is
  $$\frac{\partial \ell}{\partial \phi} = \frac{N_1}{\phi} - \frac{N_0}{1 - \phi}$$

• where $N_1$ is the number of 1's in $\{x^{(1)}, ..., x^{(N)}\}$ and $N_0$ is the number of 0's
A Bernoulli random variable takes value 1 (or heads) with probability $\phi$ and value 0 (or tails) with probability $1 - \phi$.

The pmf of the Bernoulli distribution is

$$p(x|\phi) = \phi^x (1 - \phi)^{1-x}$$

The partial derivative of the log-likelihood is

$$\frac{\partial \ell}{\partial \phi} = N \left( \frac{\hat{\phi}}{1 - \hat{\phi}} - \frac{N_0}{1 - \hat{\phi}} \right) = 0 \rightarrow \frac{N_1}{\hat{\phi}} = \frac{N_0}{1 - \hat{\phi}}$$

$$\rightarrow N_1 (1 - \hat{\phi}) = N_0 \hat{\phi} \rightarrow N_1 = \hat{\phi} (N_0 + N_1)$$

$$\rightarrow \hat{\phi} = \frac{N_1}{N_0 + N_1}$$

where $N_1$ is the number of 1’s in $\{x^{(1)}, ..., x^{(N)}\}$ and $N_0$ is the number of 0’s.