

Math Foundations for ML

10-606

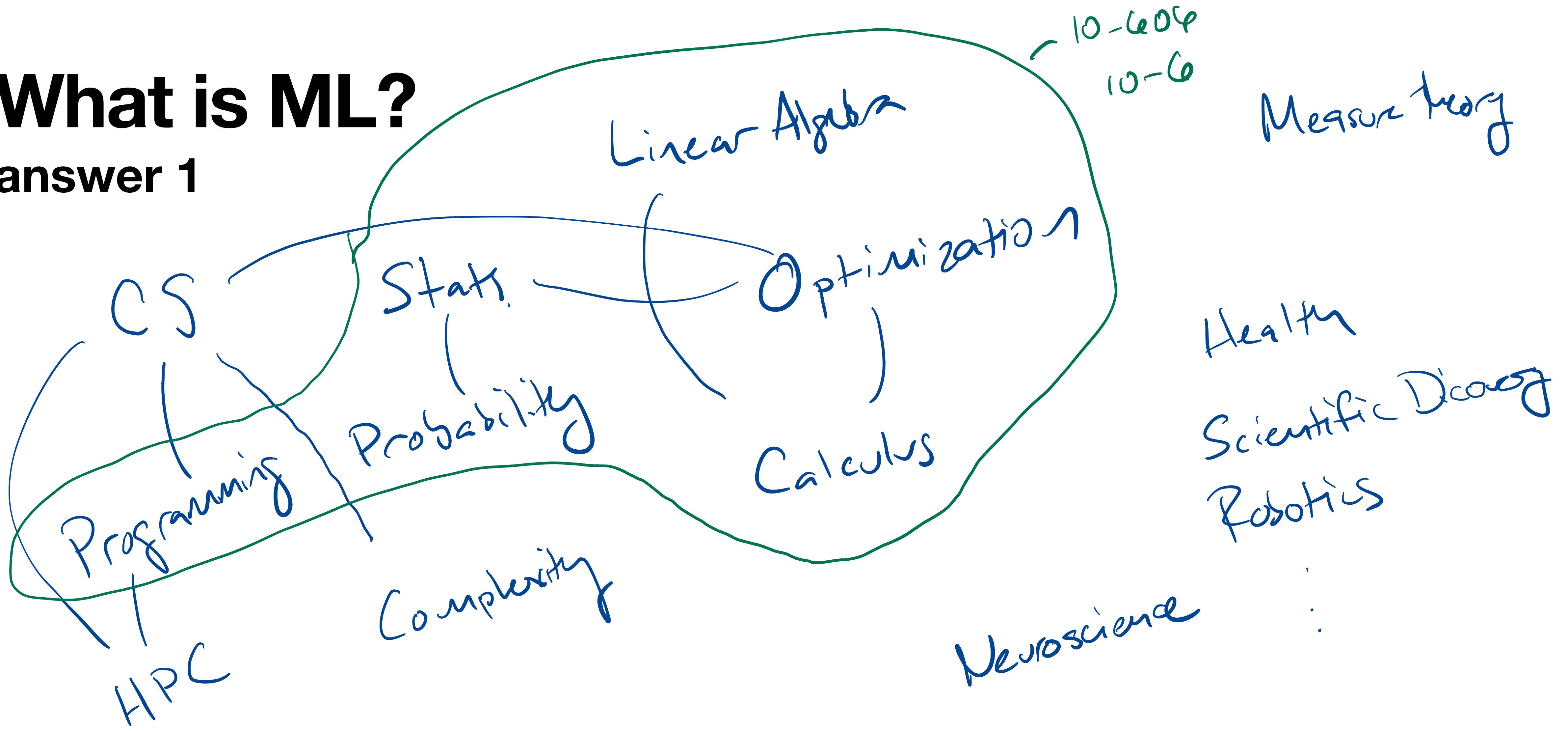
Geoff Gordon

About us

- Me: Geoff Gordon
- TAs:
 - Aditya Paul
 - Xiaoyu Xu
 - Aishwarya Jadhav

What is ML?

answer 1



What is ML?

answer 2

Speech Recognition

1. Learning to recognize spoken words

THEN	NOW
<p>"...the SPHINX system (e.g. Lee 1989) learns speaker-specific strategies for recognizing the primitive sounds (phonemes) and words from the observed speech signal...neural network methods...hidden Markov models..."</p>	

(Mitchell, 1997)

Source: <https://www.stonetemple.com/great-knowledge-box-showdown/#VoiceStudyResults>

Robotics

2. Learning to drive an autonomous vehicle

THEN	NOW
<p>"...the ALVINN system (Pomerleau 1989) has used its learned strategies to drive unassisted at 70 miles per hour for 90 miles on public highways among other cars..."</p>	

(Mitchell, 1997)

waymo.com

Games / Reasoning

3. Learning to beat the masters at board games

THEN	NOW
<p>"...the world's top computer program for backgammon, TD-GAMMON (Tesauro, 1992, 1995), learned its strategy by playing over one million practice games against itself..."</p>	

(Mitchell, 1997)

Computer Vision

4. Learning to recognize images

THEN	NOW
<p>"...The recognizer is a convolution network that can be spatially replicated. From the network output, a hidden Markov model produces word scores. The entire system is globally trained to minimize word-level errors..."</p>	

(LeCun et al., 1995)

Images from <https://blog.openai.com/generative-models/>

Learning Theory

5. In what cases and how well can we learn?

Sample Complexity Results

Definition 0.4. 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...

Finite (H)	Realizable	Agnostic
$ H < \infty$ $H = \{h_1, \dots, h_N\}$ $R(h) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \mathbb{1}(h(x) \neq y) p(x, y)$	$R(h^*) = 0$ $R(h) > 0$ for all $h \in H$ with $R(h) > 0$	$R(h^*) > 0$ $R(h) > 0$ for all $h \in H$ with $R(h) > 0$

Handwritten notes:

Finite (H): $R(h) = P_{x,y} \sum_{y \in \mathcal{Y}} \mathbb{1}(h(x) \neq y) p(x, y)$ (only uniform)

Infinite (H): $R(h) = P_{x,y} \sum_{y \in \mathcal{Y}} \mathbb{1}(h(x) \neq y) p(x, y)$ (learn, complex)

PK Learning:

Q: Can we bound $R(h)$ in terms of $\hat{R}(h)$?
 A: Yes!

PK shows R (Probably Approximately Correct) \rightarrow PK shows yields hypothesis h , which is approximately correct with high probability $P(R(h) \leq \epsilon) \geq 1 - \delta$

Def: PK Criterion $P(R(h) - \hat{R}(h) \leq \epsilon) \geq 1 - \delta$

1. How many examples do we need to learn?
 2. How do we quantify our ability to generalize to unseen data?
 3. Which algorithms are better suited to specific learning settings?

Why this course?

- Dual problem (derivation):

$$L(\mathbf{w}, b, \alpha) = \frac{1}{2} \mathbf{w} \cdot \mathbf{w} - \sum_{j=1}^n \alpha_j [(\mathbf{w} \cdot \mathbf{x}_j + b) y_j - 1]$$

$$\alpha_j \geq 0, \forall j$$

α - weights on training pts (n-dim problem)

Dual SVM – linearly separable case

- Dual problem (derivation):

$$\max_{\alpha_j \geq 0} d(\alpha)$$

$$\max_{\alpha} \min_{\mathbf{w}, b} L(\mathbf{w}, b, \alpha) = \frac{1}{2} \mathbf{w} \cdot \mathbf{w} - \sum_j \alpha_j [(\mathbf{w} \cdot \mathbf{x}_j + b) y_j - 1]$$

$$\alpha_j \geq 0, \forall j$$

$$\frac{\partial L}{\partial \mathbf{w}} = \mathbf{w} - \sum_j \alpha_j \mathbf{x}_j y_j$$

$$\rightarrow \frac{\partial L}{\partial \mathbf{w}} = 0 \Rightarrow \mathbf{w} = \sum_j \alpha_j y_j \mathbf{x}_j$$

$$\rightarrow \frac{\partial L}{\partial b} = 0 \Rightarrow \sum_j \alpha_j y_j = 0$$

$$\frac{\partial L}{\partial b} = \sum_j \alpha_j y_j$$

Why this course?

Dual SVM – linearly separable case

- Dual problem:

$$\sum_j \alpha_j b y_j = b \sum_j \alpha_j y_j = 0$$

$$\max_{\alpha} \min_{\mathbf{w}, b} L(\mathbf{w}, b, \alpha) = \frac{1}{2} \mathbf{w} \cdot \mathbf{w} - \sum_j \alpha_j [(\mathbf{w} \cdot \mathbf{x}_j + b) y_j - 1]$$

$$\alpha_j \geq 0, \forall j$$

$$\Rightarrow \mathbf{w} = \sum_j \alpha_j y_j \mathbf{x}_j \quad \Rightarrow \sum_j \alpha_j y_j = 0$$

$$\frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \mathbf{x}_i \cdot \mathbf{x}_j - \sum_j \alpha_j \left(\sum_i \alpha_i y_i \mathbf{x}_i \cdot \mathbf{x}_j \right) y_j + \sum_j \alpha_j$$

$$\sum_j \alpha_j - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \mathbf{x}_i \cdot \mathbf{x}_j =: d(\alpha)$$

Why this course?

4.2 Logistic Regression Implementation [30 points]

Implementation instructions.

1-Sentence Overview: You will be training a logistic regression model on the Homework 1 dataset by running gradient descent and then answering the written questions below.

Details: You will fill in the `logistic_regression.py` template and submit your complete file to Gradescope, where we will run your code against a suite of tests. Your grade will be automatically determined from the testing results. Since you get immediate feedback after ...

Course page

- <https://www.cs.cmu.edu/~ggordon/10606s22/syllabus-and-lecture-outline.html>

Help one another learn

- <https://xkcd.com/1053/>

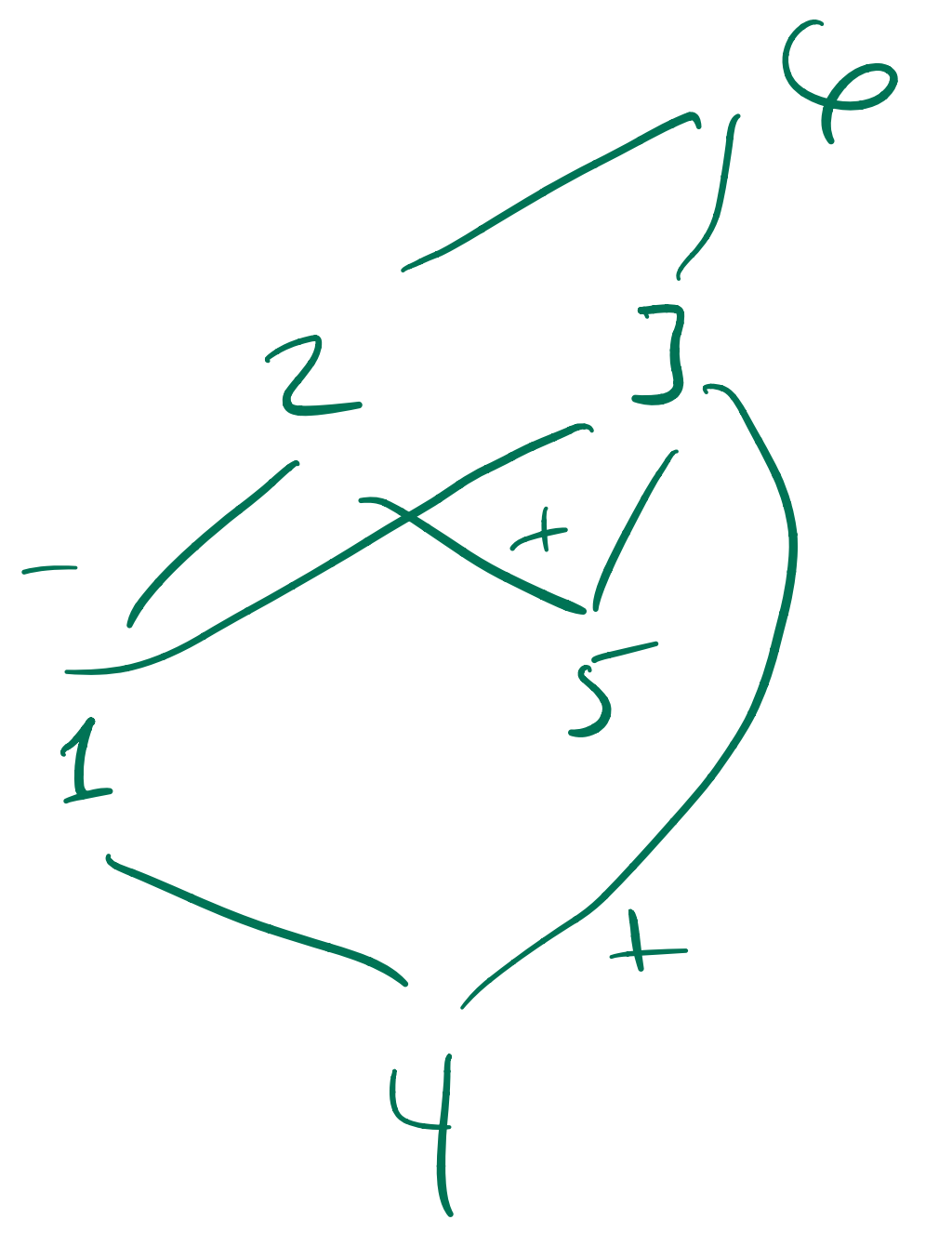
Formal systems

Partial derivatives

Integrals

0, 1, 2, 3, 4, 5, 6

+, -



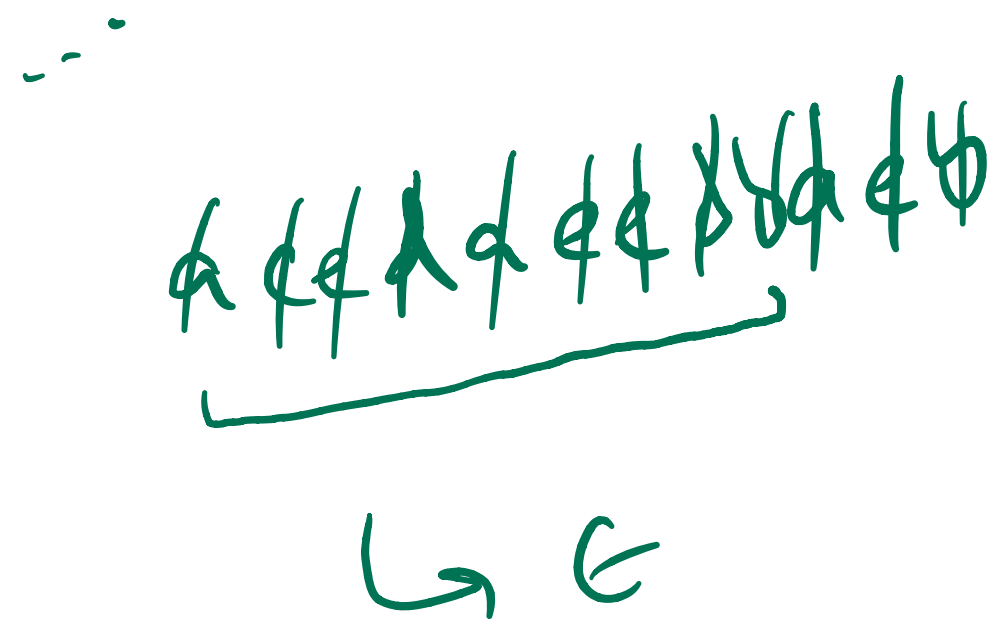
Car simulator

Gaussian elim

PM

propositional logic

classical intuitionist



expressions = [abcde]*

rules:

ε	a	[
ε	b]
ε	c	(
ε	d)
ε	e	anything else

Set builder notation

$$\{A, B, C\}$$

$$\{\} \rightarrow \emptyset$$

$$\{2x \mid x \in \mathbb{Z}\}$$

↑
expr

↑
property

$$\{2x \mid \underbrace{x \in \mathbb{Z}}_{\text{AND}} \underbrace{x > 0}\}$$

$$\{\underbrace{x \in \mathbb{Z}}_{\text{one simple property}} \mid x \% 2 = 0\}$$

$$\{x \times 2 \text{ for } x \text{ in } \text{range}(4)\}$$

$$\{x^2 \mid x \in \{0, 1, 2, 3\}\}$$