REVIEW:
PROBABILITY, LINEAR ALGEBRA, CALCULUS

THE LECTURE MAY HAVE AUDIO RECORDINGS FOR ACADEMIC ACCOMMODATION PURPOSES

PROBABILITY

Randomness and uncertainty in the world, our decisions, and the evaluation

- Black Jack: no guarantee on a hand (reward is random from the agent's perspective)
- No deterministic execution:
  - Decide to leave the highway but miss the exit
- Unmodelable dynamics of the world
  - Robot hits a golf ball — the wind can come and move the ball

PROBABILITY

RANDOM VARIABLE

Random variable X is a function from possible outcomes in a *sample space* to a *measurable space.* (Wikipedia)

Example: Flipping a coin

- Sample space: The set {heads, tails}
- Measurable space: The set {-1, 1}
- X: {heads, tails} -> {-1, 1}

PROBABILITY

RANDOM VARIABLE

Example: Rolling a 20 sided die twice and taking the sum

- Sample space: The set {(1,1), (1,2), …(2,1),…,(20,20)}
- Measurable space: The set {2,3,…,40}
- X: {(1,1),…,(20,20)} -> {2,3,…,40}

PROBABILITY

RANDOM VARIABLE

Example: D&D Sword (roll for damage)

- Sample space: The set {1,2,3,…,20}
- Measurable space: The set {1,2,4}
- X: {1,2,3,…,20} -> {1,2,4}

PROBABILITY

DEFINITION

$\Pr(A)$ is a measure of how likely is the event A

$$\Pr(A) = \frac{\text{\# occurances of } A}{\text{\# of all possible outcomes}} \in [0,1]$$

Dice roll X: {1,2,…,20} -> {1,2,…,20}

Event $A := X \in [15,20]$

$$\Pr(A) = \frac{6}{20}$$

PROBABILITY

NOTE

$\Pr(X)$ is undefined because X is not an event

PROBABILITY

PROPERTIES

Normalization $\sum_{e \in \mathcal{S}} \Pr(e) = 1$

Additivity $\Pr(A \cup B) = \Pr(A) + \Pr(B)$ if $A \cap B = \{\,\}$

PROBABILITY

COMMON MANIPULATIONS OF PROBABILITIES

Chain Rule: $\Pr(A \cap B) = \Pr(A, B) = \Pr(A \mid B)\Pr(B)$

Conditional Probability: $\Pr(A \mid B) = \dfrac{\Pr(A, B)}{\Pr(B)}$

Bayes Rule: $\Pr(A \mid B) = \dfrac{\Pr(A, B)}{\Pr(B)} = \dfrac{\Pr(B \mid A)\Pr(A)}{\Pr(B)}$

PROBABILITY

CHAIN RULE WITH MULTIPLE RANDOM VARIABLES

$\Pr(A, B, C, D) = \, ?$

Let $E = A, B, D$

$$\Pr(A, B, C, D) = \Pr(E, C) = \Pr(C \mid E) \Pr(E) = \Pr(C \mid A, B, D) \Pr(A, B, D)$$

Let $F = B, D$

$$\Pr(A, B, D) = \Pr(A, F) = \Pr(A \mid F) \Pr(F) = \Pr(A \mid B, D) \Pr(B, D)$$

$$\Pr(B, D) = \Pr(D \mid B) \Pr(B)$$

Plug it all in:

$$\Pr(A, B, C, D) = \Pr(C \mid A, B, D) \Pr(A \mid B, D) \Pr(D \mid B) \Pr(B)$$

PROBABILITY

INDEPENDENCE

A and B are independent if $\Pr(A \mid B) = \Pr(A)$

Conditional independence:

A and B are conditionally independent given C if

$$\Pr(A \mid B, C) = \Pr(A \mid C)$$

Conditional independence lets us remove variables (e.g., B) from an expression.

PROBABILITY

MARGINALIZATION

Two random variables $X \in \mathscr{X}$ and $Y \in \mathscr{Y}$

$$\Pr(X = x) = \sum_{y \in \mathscr{Y}} \Pr(X = x, Y = y)$$

$$\Pr(X = x) = \sum_{y \in \mathscr{Y}} \Pr(X = x \mid Y = y) \Pr(Y = y)$$

PROBABILITY

EXPECTATION

An expectation is the mean or average value of a random variable.

Let the random variable $X \in \mathcal{X}$

The mean of X is

$$\mathbb{E}[X] = \sum_{x \in \mathcal{X}} \Pr(X = x)x$$

For some function $f\colon \mathcal{X} \to \mathbb{R}$

$$\mathbb{E}[f(X)] = \sum_{x \in \mathcal{X}} \Pr(X = x)f(x)$$

PROBABILITY

EXPECTATION: PROPERTIES

Linearity of expectation

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$$

$$\mathbb{E}[\alpha X] = \alpha \mathbb{E}[X]$$

PROBABILITY

EXPECTATION

Conditional Expectation

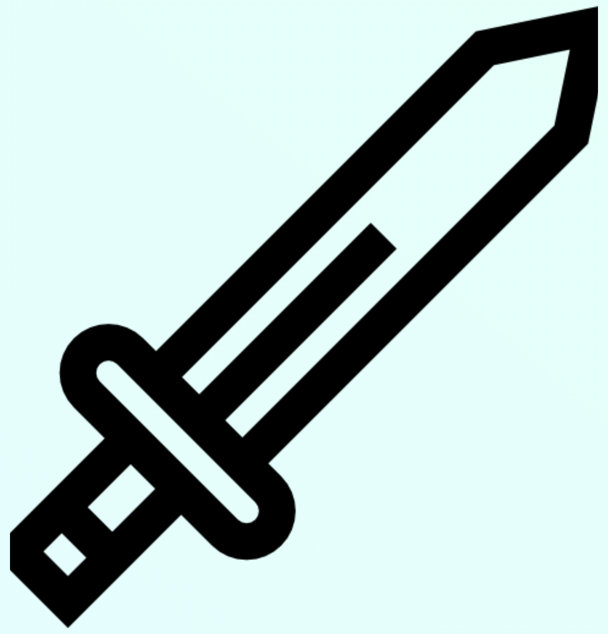
$$\mathbb{E}[X|A] = \sum_{x \in \mathcal{X}} \Pr(X = x|A)x$$

For some function $f \colon \mathcal{X} \to \mathbb{R}$

$$\mathbb{E}[f(X)] = \sum_{x \in \mathcal{X}} \Pr(X = x)f(x)$$

PROBABILITY

EXAMPLE: CHOOSING THE BEST WEAPON



D_S	$\Pr(D_S)$
1	0.2
2	0.5
4	0.3

D_H	$\Pr(D_H)$
0	0.4
4	0.6

PROBABILITY

EXAMPLE: CHOOSING THE BEST WEAPON

Choose the weapon that has the highest damage on average

$$w^* \in \arg\max_{w \in S,H} \mathbb{E}[D_w]$$

$$\mathbb{E}[D_S] = \sum_{d \in \{1,2,4\}} \Pr(D_S = d)d = 1(0.2) + 2(0.5) + 4(0.3) = 2.4$$

$$\mathbb{E}[D_H] = \sum_{d \in \{0,4\}} \Pr(D_H = d)d = 0(0.4) + 4(0.6) = 2.4$$

$$w^* \in \arg\max_{w \in S,H} \mathbb{E}[D_w] = \{S, H\}$$

PROBABILITY

EXAMPLE: OPTIMIZING FOR LAST HIT

Maximize the probability that a monster dies when it has x health left

The last hit occurs when $d \in \mathscr{L}_x = \{d \,|\, d \geq x\}$

$$w^* \in \operatorname{arg}\max_{w \in S,H} \Pr(D_w \in \mathscr{L}_x)$$

$$\Pr(D_S \in \mathscr{L}_x) = \sum_{d \in 1,2,4} \mathbf{1}_{d \in \mathscr{L}_x} \Pr(D_S = d)$$

$$\Pr(D_H \in \mathscr{L}_x) = \sum_{d \in 0,4} \mathbf{1}_{d \in \mathscr{L}_x} \Pr(D_H = d)$$

PROBABILITY

EXAMPLE: OPTIMIZING FOR LAST HIT

$x = 3$

$\Pr(D_S \in \mathscr{L}_x) = 0.3$

$\Pr(D_H \in \mathscr{L}_x) = 0.6$

$x = 1$

$\Pr(D_S \in \mathscr{L}_x) = 0.2 + 0.5 + 0.3 = 1$

$\Pr(D_H \in \mathscr{L}_x) = 0.6$

PROBABILITY

EXAMPLE: DIFFERENT MONSTERS

Orcs:

Hammer gets -1 damage on hit

Sword — no change

Goblins:

Hammer gets $+2$ damage on hit

Sword — no change

$M \in \mathscr{M} = \{\text{Orc}, \text{Goblin}\}$

PROBABILITY

EXAMPLE: DIFFERENT MONSTERS

Choose the best weapon on average despite the monster

$$w^* = \text{arg}max_{w\in\{S,H\}}\mathbb{E}[D_w]$$

$$\mathbb{E}[D_w] = \sum_d \sum_{m\in\mathcal{M}} \Pr(D_w = d, M = m)d$$

$$\mathbb{E}[D_w] = \sum_d \sum_{m\in\mathcal{M}} \Pr(D_w = d \,|\, M = m)\Pr(M = m)d$$

$$\mathbb{E}[D_w] = \sum_{m\in\mathcal{M}} \Pr(M = m)\sum_d \Pr(D_w = d \,|\, M = m)d$$

$$\mathbb{E}[D_w] = \sum_{m\in\mathcal{M}} \Pr(M = m)\mathbb{E}[D_w \,|\, M = m]$$

PROBABILITY

EXAMPLE: DIFFERENT MONSTERS

$$\mathbb{E}[D_S] = \sum_{m \in \mathcal{M}} \Pr(M = m)\mathbb{E}[D_S \mid M = m] = \sum_{m \in \mathcal{M}} \Pr(M = m) \underbrace{\mathbb{E}[D_S]}_{D_S \text{ is independent of } M}$$

$$= \mathbb{E}[D_S] \sum_{m \in \mathcal{M}} \Pr(M = m) = \mathbb{E}[D_S]$$

PROBABILITY

EXAMPLE: DIFFERENT MONSTERS

$$
\begin{aligned}
\mathbb{E}[D_H] &= \sum_{m \in \mathscr{M}} \Pr(M = m)\mathbb{E}[D_H | M = m] \\
&= \mathbb{E}[D_H | M = \text{Orc}] \Pr(M = \text{Orc}) + \mathbb{E}[D_H | M = \text{Goblin}] \Pr(M = \text{Goblin})
\end{aligned}
$$

$$\mathbb{E}[D_H | M = \text{Orc}] = 0(0.4) + (4 - 1)(0.6) = 1.8$$

$$\mathbb{E}[D_H | M = \text{Goblin}] = 0(0.4) + (4 + 2)(0.6) = 3.6$$

$$\mathbb{E}[D_H] = 1.8 \Pr(M = \text{Orc}) + 3.6 * \Pr(M = \text{Goblin})$$

PROBABILITY

EXAMPLE: DIFFERENT MONSTERS

Use a sword when

$$\begin{aligned} \mathbb{E}[D_S] &\geq \mathbb{E}[D_H] \\ 2.4 &\geq 1.8 \Pr(M = \text{Orc}) + 3.6 \Pr(M = \text{Goblin}) \end{aligned}$$

LINEAR ALGEBRA

DEFINITIONS

1. Linear algebra lets us summarize groups of numbers with convenient notation
2. Help us represent functions of many variables more compactly

$x \in \mathbb{R}^3$ is a column vector of elements, $A \in \mathbb{R}^{2,3}$ is the matrix

$$x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{13} \end{bmatrix}$$

LINEAR ALGEBRA

PROPERTIES

Linearity

$\alpha x = [\alpha x_1, \alpha x_2, \alpha x_3]^\top$

$$\alpha A = \begin{bmatrix} \alpha a_{11} & \alpha a_{12} & \alpha a_{13} \\ \alpha a_{21} & \alpha a_{22} & \alpha a_{13} \end{bmatrix}$$

$\top$ is the transpose operator that flips the rows and columns, e.g., $x^\top \in \mathbb{R}^{1,3}$ and $A^\top \in \mathbb{R}^{3,2}$

$x \in \mathbb{R}^3$ and $y \in \mathbb{R}^3$

$$x + y = \begin{bmatrix} x_1 + y_1 \\ x_2 + y_2 \\ x_3 + y_3 \end{bmatrix}$$

$A \in \mathbb{R}^{2,3}$ and $B \in \mathbb{R}^{2,3}$

$$A + B = \begin{bmatrix} a_{11} + b_{11} & a_{12} + b_{12} & \alpha a_{12} \\ a_{21} + b_{21} & a_{22} + b_{22} & \alpha a_{22} \end{bmatrix}$$

LINEAR ALGEBRA

USEFUL OPERATIONS

Inner produt: $\sum_{i=1}^{3} x_i y_i = x^\top y$

Matrix-vector multiplication

$$Ax = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 \end{bmatrix}$$

LINEAR ALGEBRA

EXAMPLE

Represent parameters of a line $f(x) = mx + b$

Vector of parameters $\theta = [m, b]^\top$

Vector representation of inputs $\mathbf{x} = [x, 1]^\top$

Vector representation of f: $f \colon \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$

$$f(\mathbf{x}, \theta) = \theta^\top \mathbf{x} = \theta_1 \mathbf{x}_1 + \theta_2 \mathbf{x}_2 = mx + b(1)$$

LINEAR ALGEBRA

EXAMPLE

Predict a person's chance of heart attack, stroke, and death.

$$f_{\text{heart}}(\text{age}) = m_1\text{age} + b_1, \quad f_{\text{stroke}}(\text{age}) = m_2\text{age} + b_2, \quad f_{\text{death}}(\text{age}) = m_3\text{age} + b_3$$

$$\theta \in \mathbb{R}^{2,3} = \begin{bmatrix} m_1 & m_2 & m_3 \\ b_1 & b_2 & b_3 \end{bmatrix}, x = [\text{age}, 1]^\top$$

$$f_{\text{diseases}}(x, \theta) = \theta^\top x = \begin{bmatrix} m_1\text{age} + b_1 \\ m_2\text{age} + b_2 \\ m_3\text{age} + b_3 \end{bmatrix}$$

CALCULUS

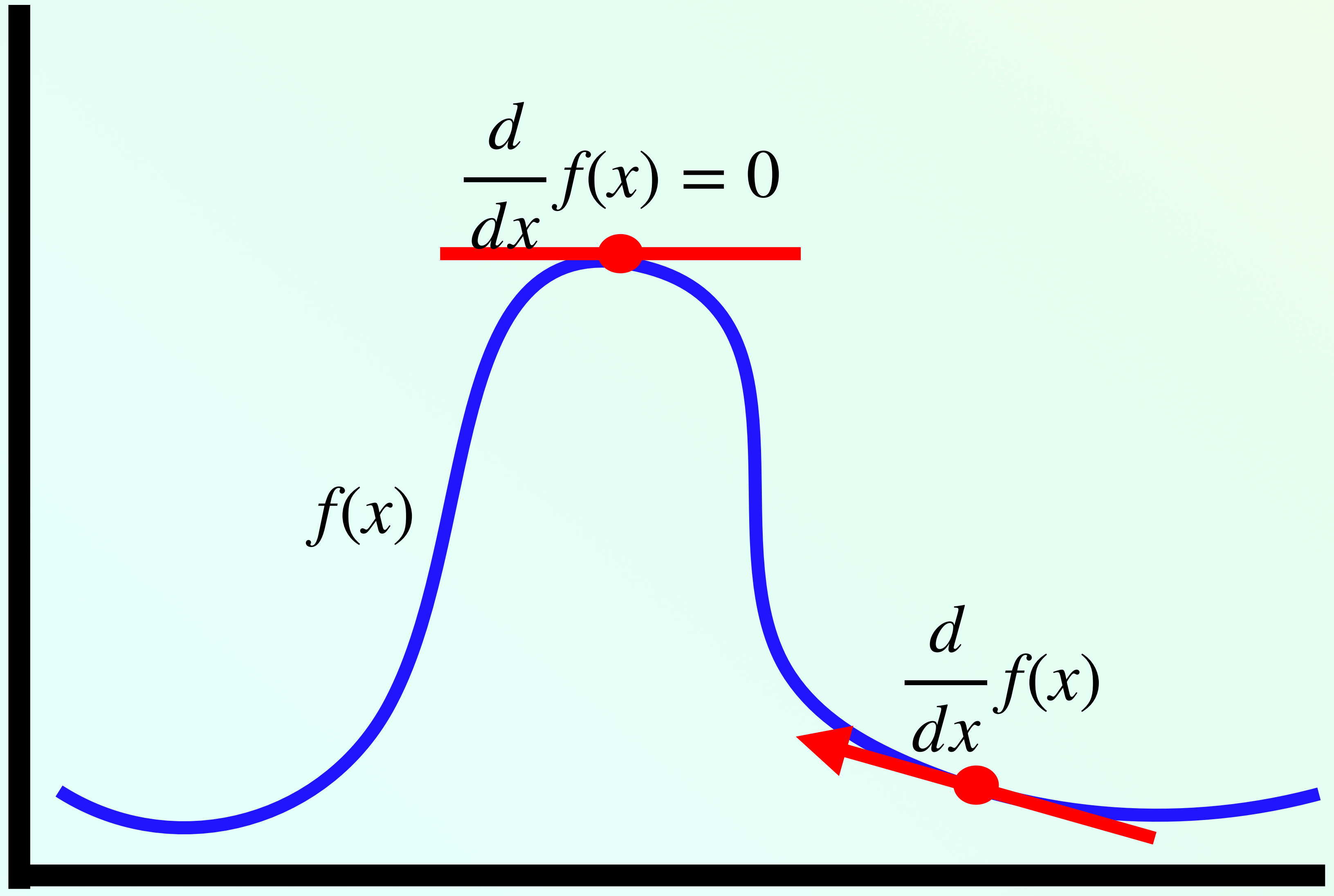
WHY

1. Take derivatives (how fast a function changes)
2. Helpful for finding the minimum and maximum of a function
3. Used in many iterative learning algorithms in the course

CALCULUS

DEFINITIONS

A derivative of a function gives the slope (tangent) of that function

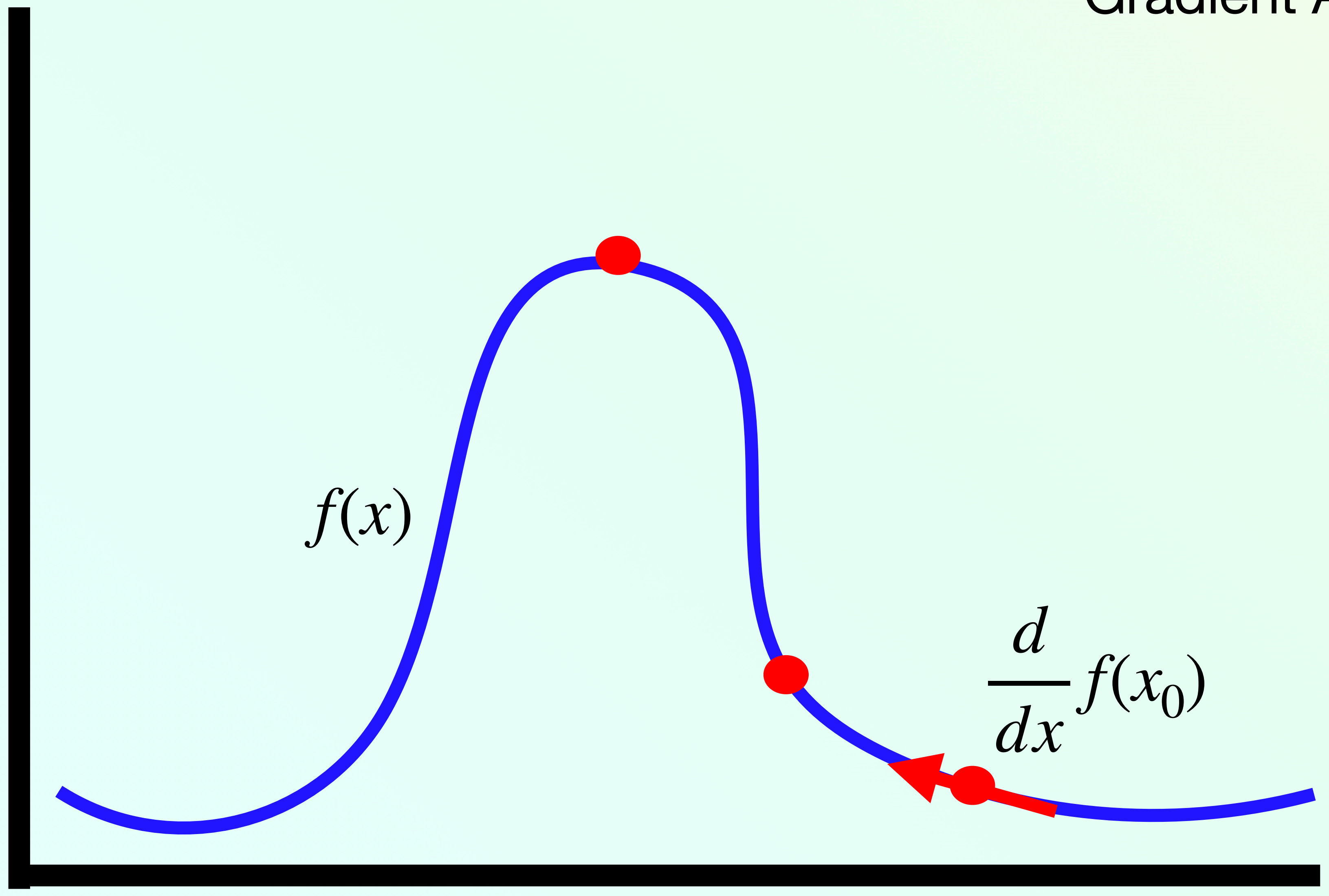


CALCULUS

OPTIMIZATION

Derivative says how to change the input to the function to increase that function

Gradient Ascent: $x_{i+1} \leftarrow x_i + \eta \frac{d}{dx} f(x_i)$

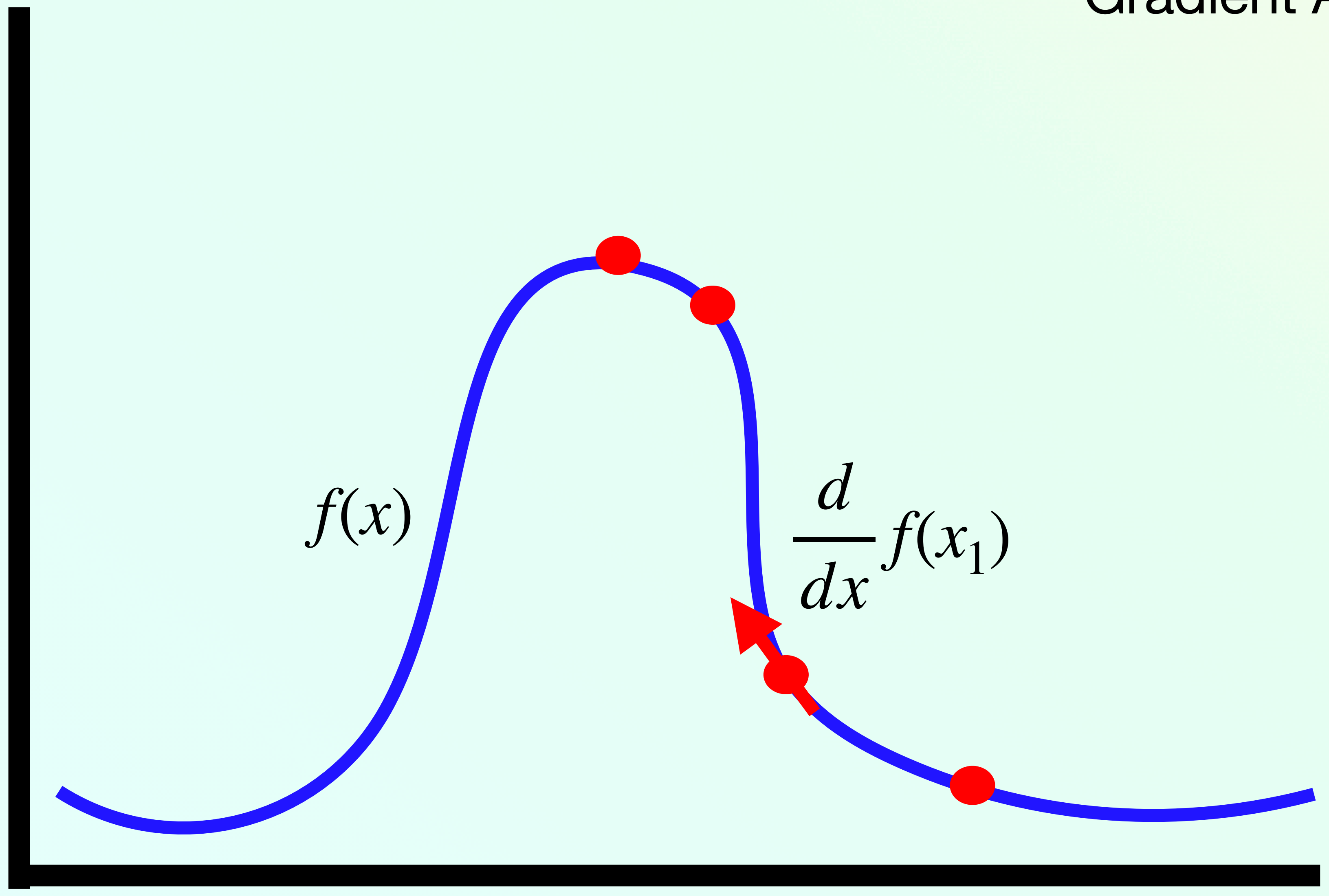


CALCULUS

OPTIMIZATION

Derivative says how to change the input to the function to increase that function

Gradient Ascent: $x_{i+1} \leftarrow x_i + \eta \frac{d}{dx} f(x_i)$

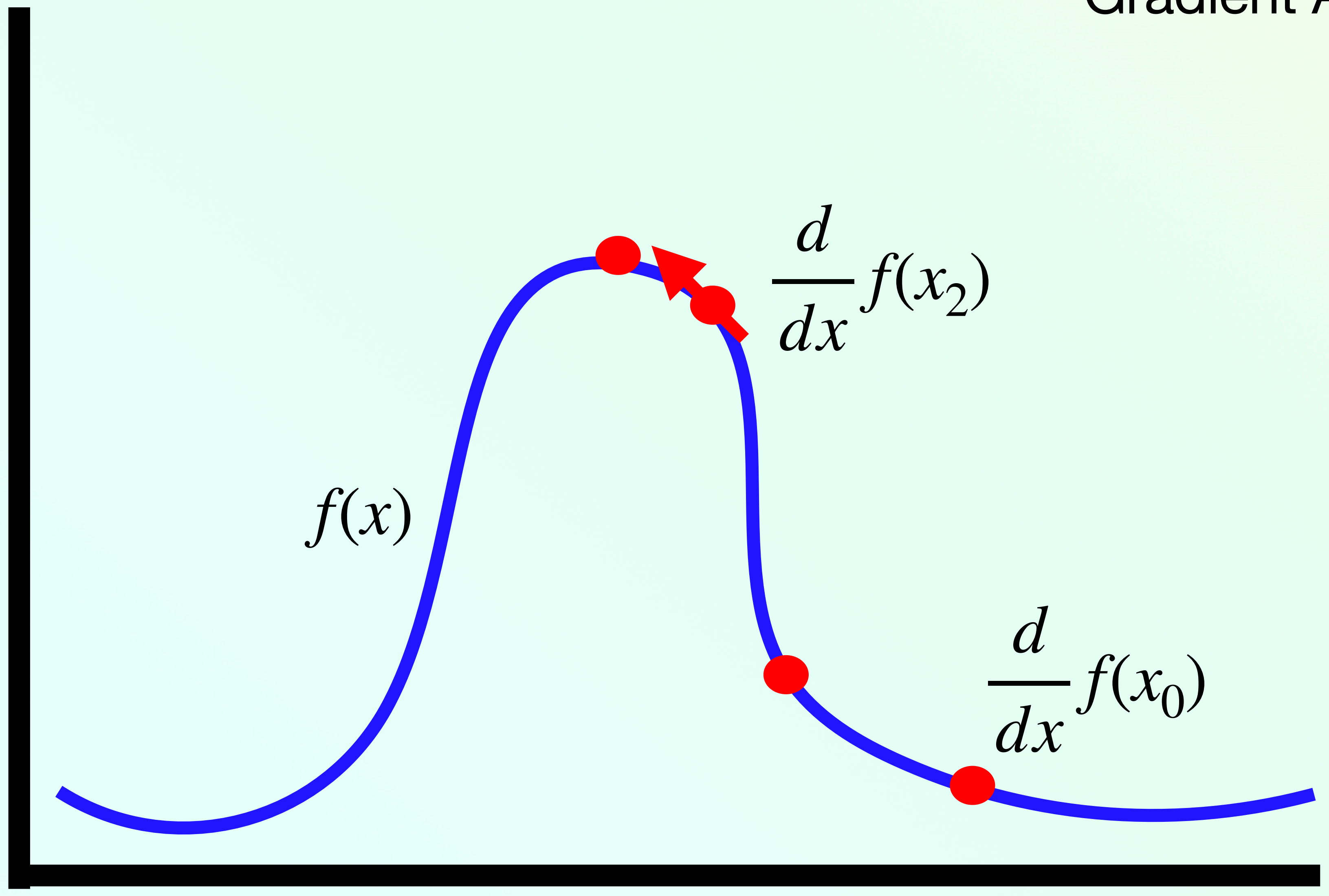


CALCULUS

OPTIMIZATION

Derivative says how to change the input to the function to increase that function

Gradient Ascent: $x_{i+1} \leftarrow x_i + \eta \frac{d}{dx} f(x_i)$



CALCULUS

DEFINITIONS: MULTIVARIATE

Partial derivatives of a multivariate function $f(x, y) = x^2 + y^2$ are

$$\frac{\partial}{\partial x} f(x, y) = \frac{\partial}{\partial x} x^2 + \frac{\partial}{\partial x} y^2 = \frac{\partial}{\partial x} x^2$$

The gradient of a function is all partial derivatives of that function's inputs, i.e.,

$$\nabla f(x, y) = \left[\frac{\partial}{\partial x} f(x, y), \frac{\partial}{\partial y} f(x, y)\right]^\top$$

CALCULUS

WITH LINEAR ALGEBRA

$$w \in \mathbb{R}^3, x \in \mathbb{R}^3 \quad f(w,x) = w^\top x = \sum_{i=1}^{3} w_i x_i$$

$$\frac{\partial}{\partial w} f(w,x) = \begin{bmatrix} \frac{\partial f(w,x)}{\partial w_1} \\ \frac{\partial f(w,x)}{\partial w_2} \\ \frac{\partial f(w,x)}{\partial w_3} \end{bmatrix}$$

$$\frac{\partial f(w,x)}{\partial w_1} = \frac{\partial}{\partial w_1} \sum_{i=1}^{3} w_i x_i = \frac{\partial}{\partial w_1} w_1 x_1 + \frac{\partial}{\partial w_1} w_2 x_2 + \frac{\partial}{\partial w_1} w_3 x_3 = \frac{\partial}{\partial w_1} w_1 x_1 = x_1$$

$$\frac{\partial}{\partial w} f(w,x) = \begin{bmatrix} \frac{\partial f(w,x)}{\partial w_1} \\ \frac{\partial f(w,x)}{\partial w_2} \\ \frac{\partial f(w,x)}{\partial w_3} \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = x$$

CALCULUS

WITH LINEAR ALGEBRA

$$w \in \mathbb{R}^{3,2}, x \in \mathbb{R}^3 \quad f(w,x) = w^\top x = \left[\sum_{i=1}^{3} w_{i1}x_i, \sum_{i=1}^{3} w_{i2}x_i\right]^\top = \left[w_{\cdot 1}^\top x, w_{\cdot 2}^\top x\right]^\top \qquad g(w,x) = \sum_{k=1}^{2} f(w,x)_k$$

$$\frac{\partial}{\partial w} g(w,x) = \begin{bmatrix} \frac{\partial g(w,x)}{\partial w_{11}} & \frac{\partial g(w,x)}{\partial w_{12}} \\ \frac{\partial g(w,x)}{\partial w_{21}} & \frac{\partial g(w,x)}{\partial w_{22}} \\ \frac{\partial g(w,x)}{\partial w_{31}} & \frac{\partial g(w,x)}{\partial w_{32}} \end{bmatrix} = \begin{bmatrix} \frac{\partial g(w,x)}{\partial w_{\cdot 1}} & \frac{\partial g(w,x)}{\partial w_{\cdot 2}} \end{bmatrix}$$

$$\frac{\partial}{\partial w_{\cdot 1}} g(w,x) = \frac{\partial}{\partial w_{\cdot 1}} \left(w_{\cdot 1}^\top x + w_{\cdot 2}^\top x\right) = \frac{\partial}{\partial w_{\cdot 1}} w_{\cdot 1}^\top x + \frac{\partial}{\partial w_{\cdot 1}} w_{\cdot 2}^\top x = \frac{\partial}{\partial w_{\cdot 1}} w_{\cdot 1}^\top x = x$$

$$\frac{\partial}{\partial w} g(w,x) = \begin{bmatrix} \frac{\partial g(w,x)}{\partial w_{11}} & \frac{\partial g(w,x)}{\partial w_{12}} \\ \frac{\partial g(w,x)}{\partial w_{21}} & \frac{\partial g(w,x)}{\partial w_{22}} \\ \frac{\partial g(w,x)}{\partial w_{31}} & \frac{\partial g(w,x)}{\partial w_{32}} \end{bmatrix} = \begin{bmatrix} x & x \end{bmatrix}$$

NEXT CLASS

WHAT YOU SHOULD DO

1. Make sure you are registered on Coursera
2. Watch the Module 1 videos
3. The quiz is due by 11:59 pm after Friday’s class

Friday: Extra context and details on Module 1