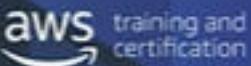


Courses



- Vectors and Matrices
- Linear Algebra Operations
- Probability
- Univariate Derivative Calculus
- Multivariate Derivatives



Machine Learning:

Field of study that gives computers the ability to learn without being explicitly programmed.¹



Arthur Samuel, 1959



What is Machine Learning?



-4:43



What is Machine Learning?

A Well-posed Learning Problem



Well-posed Learning Problem:

A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E.²



Tom Mitchell, 1998

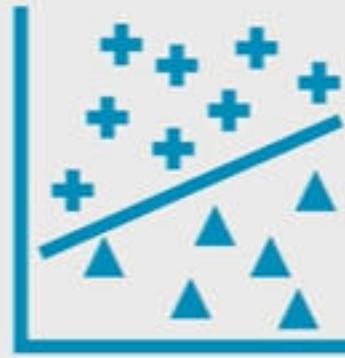
Applied Machine Learning



In practice, machine learning is:



A collection of
methods



that allow the
extraction of **rules** or
patterns from data

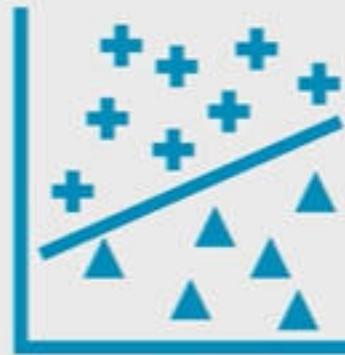
Applied Machine Learning



In practice, machine learning is:



A collection of
methods



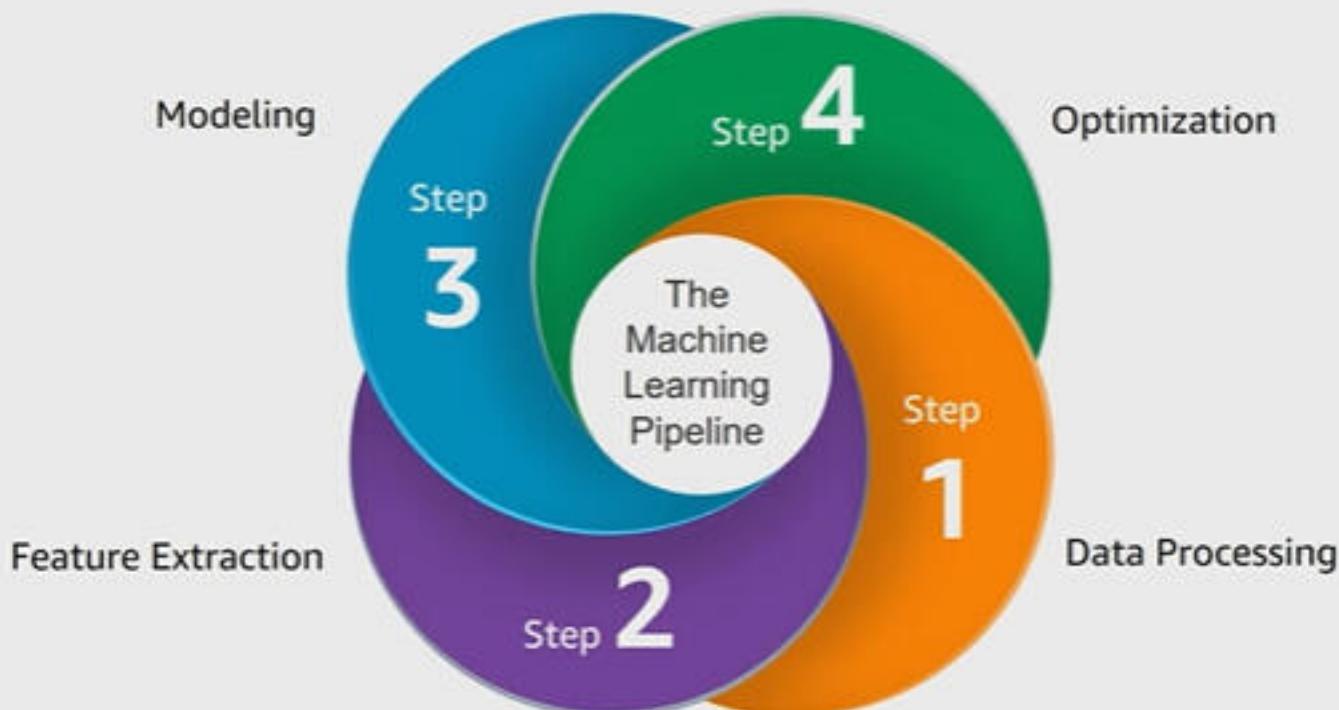
that allow the
extraction of **rules** or
patterns from data



rather than explicit
construction from a
programmer

The Machine Learning Pipeline

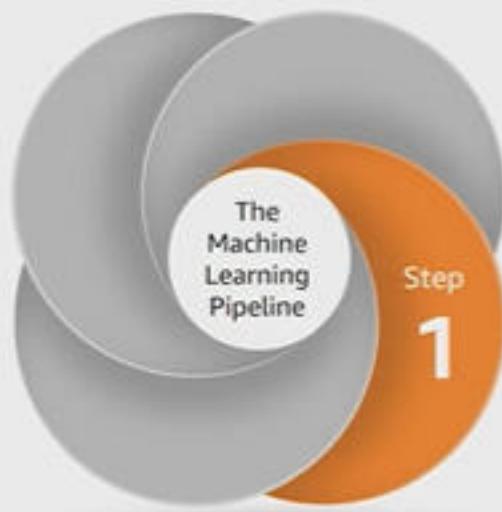
The Machine Learning Pipeline



The Machine Learning Pipeline

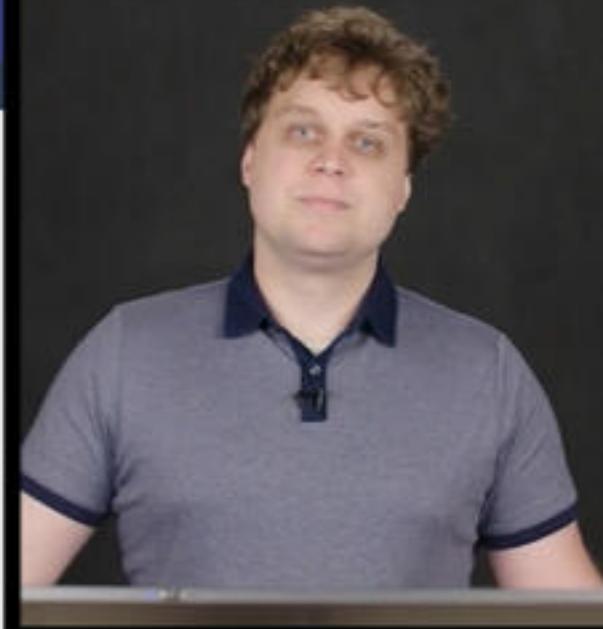


The Machine Learning Pipeline

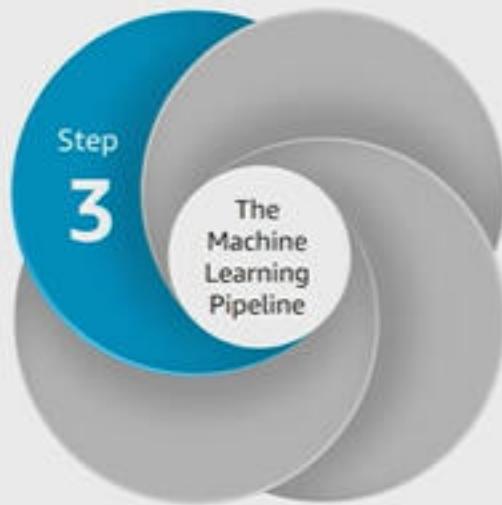


DATA PROCESSING

Collection
Formatting
Labelling



The Machine Learning Pipeline



MODELING



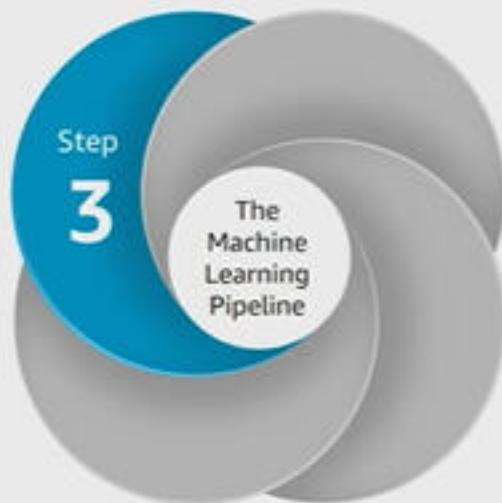
The Machine Learning Pipeline



MODELING



The Machine Learning Pipeline



MODELING



The Machine Learning Pipeline

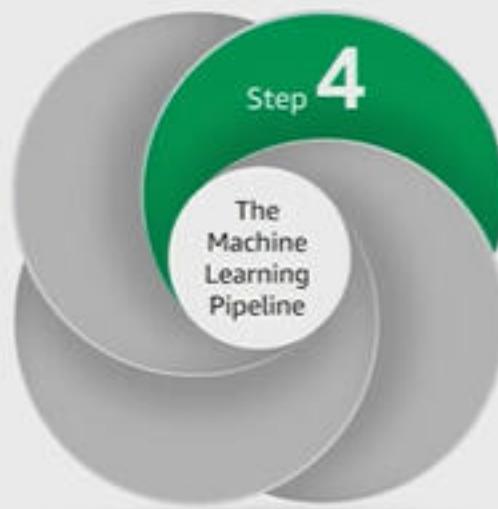


MODELING

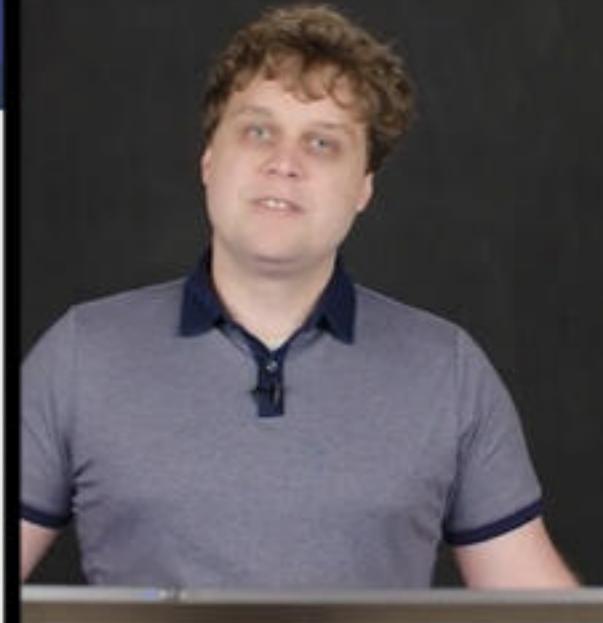
Geometry



The Machine Learning Pipeline



OPTIMIZATION



The Machine Learning Pipeline

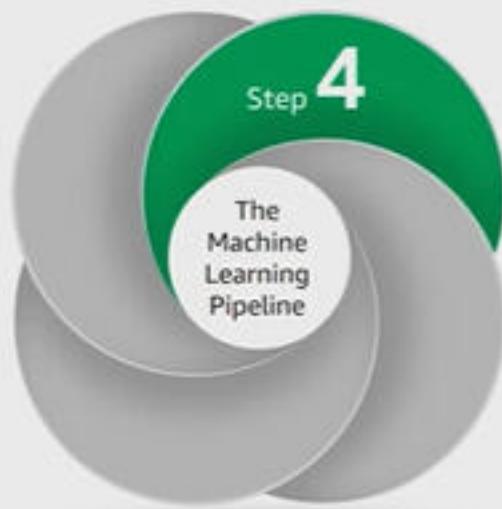


OPTIMIZATION

Training Phase



The Machine Learning Pipeline



OPTIMIZATION

Training Phase
Data Evaluation
Predictions



Mathematics in Context

What is ML-ready Data?



Data that is represented as either:

What is ML-ready Data?

Data that is represented as either:



Vectors

or

$$\begin{bmatrix} 1 & 5 \\ 7 & 9 \end{bmatrix}$$

Matrices

Transforming Data and Extracting Features



How is data transformed?

How are features extracted from the data?

You can use linear algebra to perform **multiplication** and **addition** on:

- Vectors
- Matrices

▲ Loss Functions

- Probability
- Norms
- Statistics

Loss Functions

- Probability
- Norms
- Statistics

Geometry

Loss Functions

- Probability
- Norms
- Statistics

Geometry

GOAL

Learn what
is driving
observed
events





Vector Calculus

- Functions
- Derivatives



Vector Calculus

- Functions
- Derivatives



Numerical Methods

Definitions

- Vectors
- Dimensions
- Matrices

Operations

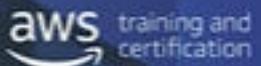
- Addition and the zero vector
- Scalar multiplication
- Transpose



-11:27



Row Vectors



$$\vec{w} = [w_1 \quad w_2 \quad \dots \quad w_n]$$



Row Vectors



$$\vec{w} = [w_1 \quad w_2 \quad \dots \quad w_n]$$

$$[1 \quad 0 \quad -1 \quad 2]$$



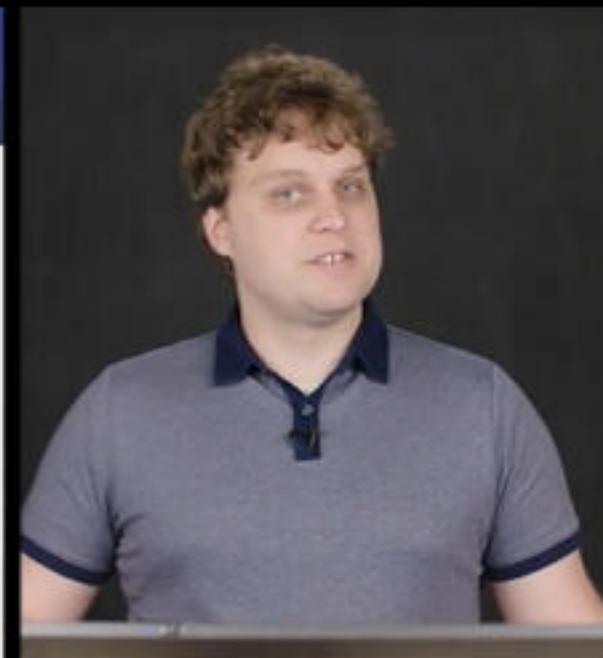
n-Dimensional



$$\vec{v} = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix}$$

A blue arrow points from the text "n-Dimensional" at the bottom left to the vector element v_n in the matrix.

n-Dimensional



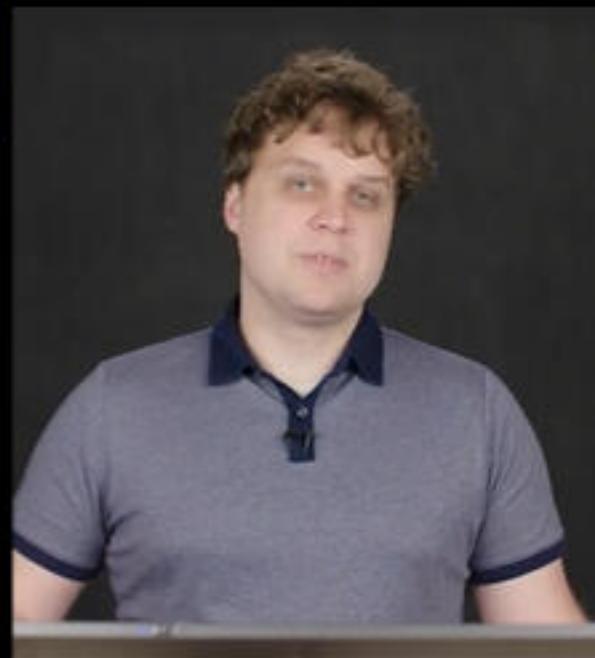
2-Dimensional



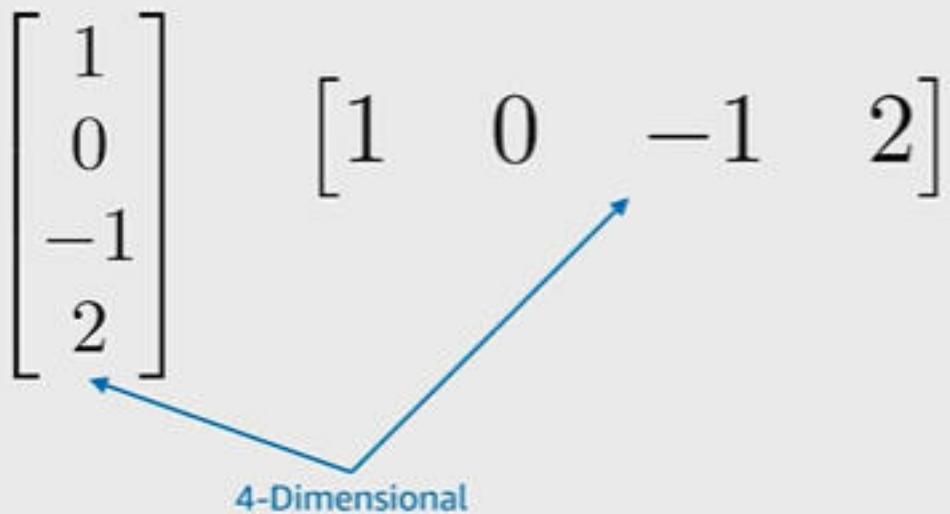
$$\begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 2 \end{bmatrix}$$

2-Dimensional



4-Dimensional



Matrices



$m \times n$ matrix

$$A = (a_{ij}) = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \dots & a_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & a_{m3} & \dots & a_{mn} \end{bmatrix}$$
$$\vec{v}_1 \quad \vec{v}_2 \quad \vec{v}_3 \quad \dots \quad \vec{v}_n$$

Matrices



$m \times n$ matrix

$$A = (a_{ij}) = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & a_{m3} & \cdots & a_{mn} \end{bmatrix}$$
$$\vec{v}_1 \quad \vec{v}_2 \quad \vec{v}_3 \quad \cdots \quad \vec{v}_n$$



Matrices

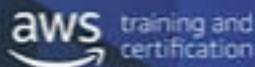
$m \times n$ matrix

$$A = (a_{ij}) = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ a_{m1} & a_{m2} & a_{m3} & \dots & a_{mn} \end{bmatrix}$$

$\vec{v}_1 \quad \vec{v}_2 \quad \vec{v}_3 \quad \dots \quad \vec{v}_n$

$1 \times n$
①
Row

Addition and The Zero Vector



$$\vec{v} = \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix}$$

$$\vec{w} = \begin{pmatrix} w_1 \\ \vdots \\ w_n \end{pmatrix}$$

$$\vec{v} + \vec{w} = \begin{pmatrix} v_1 + w_1 \\ v_2 + w_2 \\ \vdots \\ v_n + w_n \end{pmatrix}$$



Addition and The Zero Vector



$$\vec{v} = \begin{pmatrix} v_1 \\ \vdots \\ v_N \end{pmatrix}$$

$$\vec{w} = \begin{pmatrix} w_1 \\ \vdots \\ w_m \end{pmatrix}$$

$$\vec{v} + \vec{w} = \begin{pmatrix} v_1 + w_1 \\ v_2 + w_2 \\ \vdots \\ v_n + w_m \end{pmatrix}$$

$N=M$

GIVEN TWO VECTORS
WE CAN DEFINE
IF THEY ARE
LENGTH. IT IS
BY

$\vec{v} = \vec{w}$,
 $\vec{v} + \vec{w}$,
THE SAME
GIVEN

Scalar Multiplication



$a \in \mathbb{R} \leftarrow \text{REAL}$

$$\vec{v} = \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix}$$

$$a\vec{v} = \begin{pmatrix} av_1 \\ \vdots \\ av_n \end{pmatrix}$$

$$[a\vec{v}]_i = a\vec{v}_i$$

Scalar Multiplication



$a \in \mathbb{R} \leftarrow \text{REAL}$

$$\vec{v} = \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix}$$

$$a\vec{v} = \begin{pmatrix} av_1 \\ \vdots \\ av_n \end{pmatrix}$$

$$[a\vec{v}]_i = a\vec{v}_i$$

$$0\vec{v} = \begin{pmatrix} 0v_1 \\ 0v_2 \\ \vdots \\ 0v_n \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \vec{0}$$

Scalar Multiplication



$a \in \mathbb{R} \leftarrow \text{REAL}$

$$\vec{v} = \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix}$$

$$a\vec{v} = \begin{pmatrix} av_1 \\ \vdots \\ av_n \end{pmatrix}$$

$$(\vec{v} + \vec{w}) + x$$

$$[a\vec{v}]_i = a\vec{v}_i$$

$$0\vec{v} = \begin{pmatrix} 0v_1 \\ 0v_2 \\ \vdots \\ 0v_n \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix} = \vec{0}$$

$$1\vec{v} = \begin{pmatrix} 1v_1 \\ 1v_2 \\ \vdots \\ 1v_n \end{pmatrix} = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix} = \vec{v}$$



-3:00



Transpose

$$A = \begin{bmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \dots & a_{mn} \end{bmatrix}$$

$$A^T = \begin{bmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \dots & a_{mn} \end{bmatrix}$$

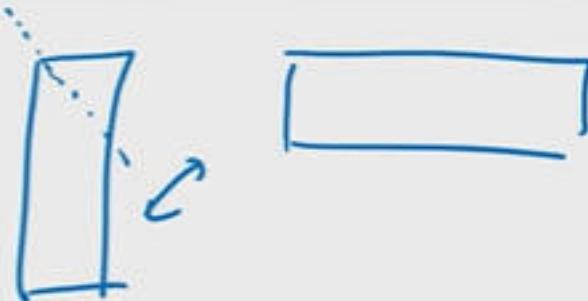
$$\vec{v} = \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix}$$

$$\begin{array}{c} A \\ A^T \end{array}$$

m × 1

n × n

MATRIX



Transpose

$$A = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix}$$

$$A^T = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix}$$

$$\vec{v} = \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix} \quad \vec{V} = (v_1, \dots, v_n)$$

$$(\alpha \vec{v})^T = \alpha \vec{v}^T$$

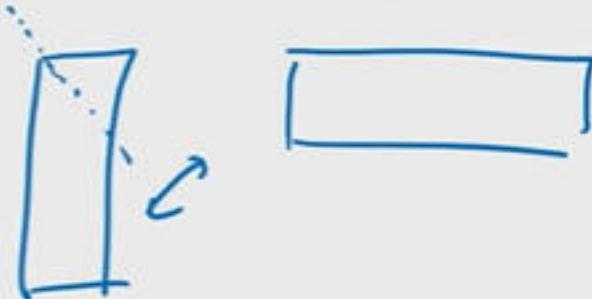
(

$$\begin{array}{c} A \\ A^T \end{array}$$

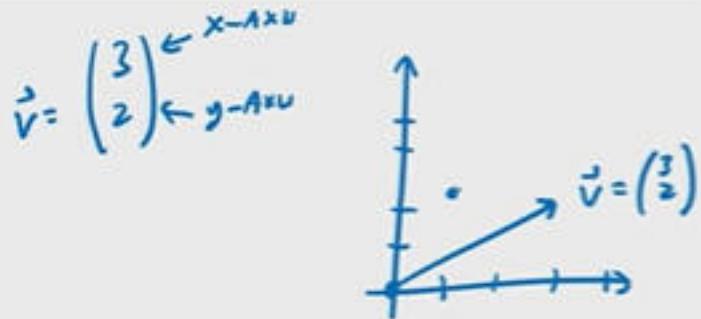
m × 1

n × n

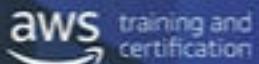
MATRIX



Vectors as Directions



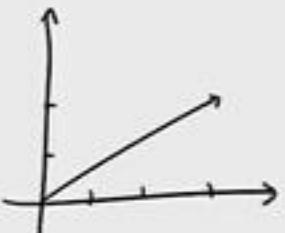
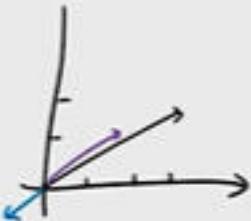
Addition as Displacement



$$\vec{v} = \begin{pmatrix} 3 \\ 2 \end{pmatrix}$$

$$\vec{w} = \begin{pmatrix} -1 \\ -1 \end{pmatrix}$$

$$\vec{v} + \vec{w} = \begin{pmatrix} 3-1 \\ 2-1 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$$



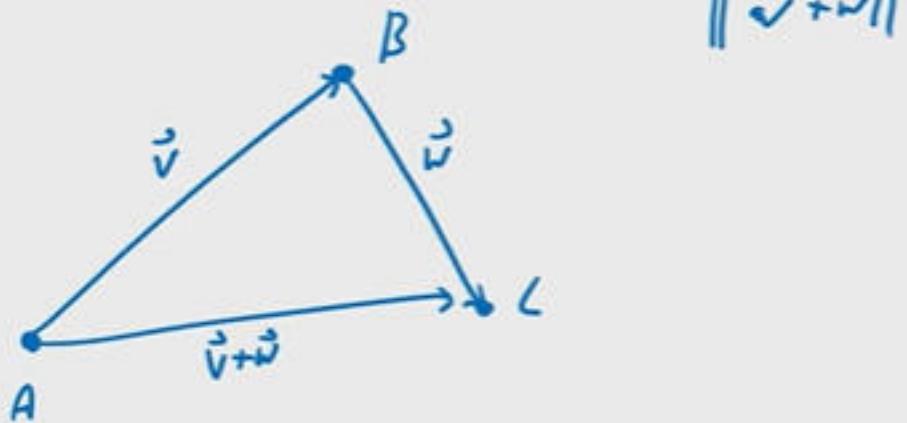
Topics

- Definition of Norms
- Norm Properties
- Euclidean Norm
- L_p -Norm
- L_1 -Norm
- L_∞ -Norm
- Geometry of Norms
- A Special Case: The L_0 -Norm



Norm Properties

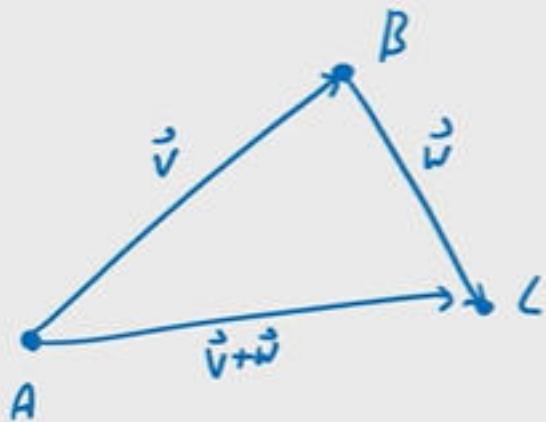
If I travel from A to B then B to C , that is at least as far as going from A to C . TRIANGLE INEQUALITY.



Norm Properties

If I travel from A to B then B to C , that is at least as far as going from A to C . *TRIANGLE NEARNESS.*

$$\|\vec{v} + \vec{w}\| \leq \|\vec{v}\| + \|\vec{w}\|$$



Euclidean Norm



$$\vec{v} = \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix}$$

$$\|\vec{v}\|_2 = \sqrt{v_1^2 + v_2^2 + v_3^2}$$

Euclidean Norm



$$\vec{v} = \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix}$$

$$\vec{v} = \begin{pmatrix} 4 \\ 3 \end{pmatrix}$$



$$\begin{aligned}\|\vec{v}\|_2 &= \sqrt{v_1^2 + v_2^2 + \dots + v_n^2} \\ &= \sqrt{v_1^2 + v_2^2 + v_3^2} \\ &= \sqrt{\sum_{i=1}^n v_i^2}\end{aligned}$$



L_p -Norm

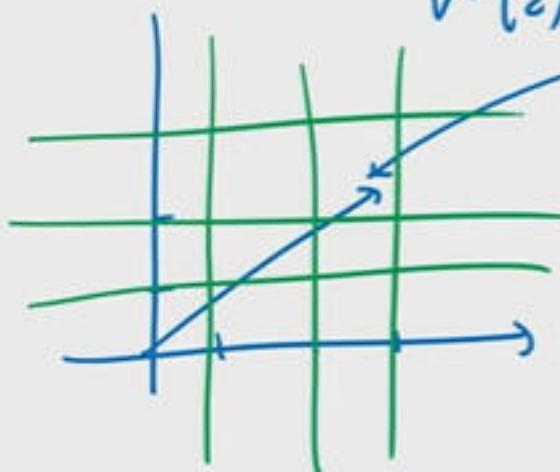
$$\|\vec{v}\|_p = \left(\sum_{i=1}^n \right)^{\frac{1}{p}}$$

L_1 -Norm

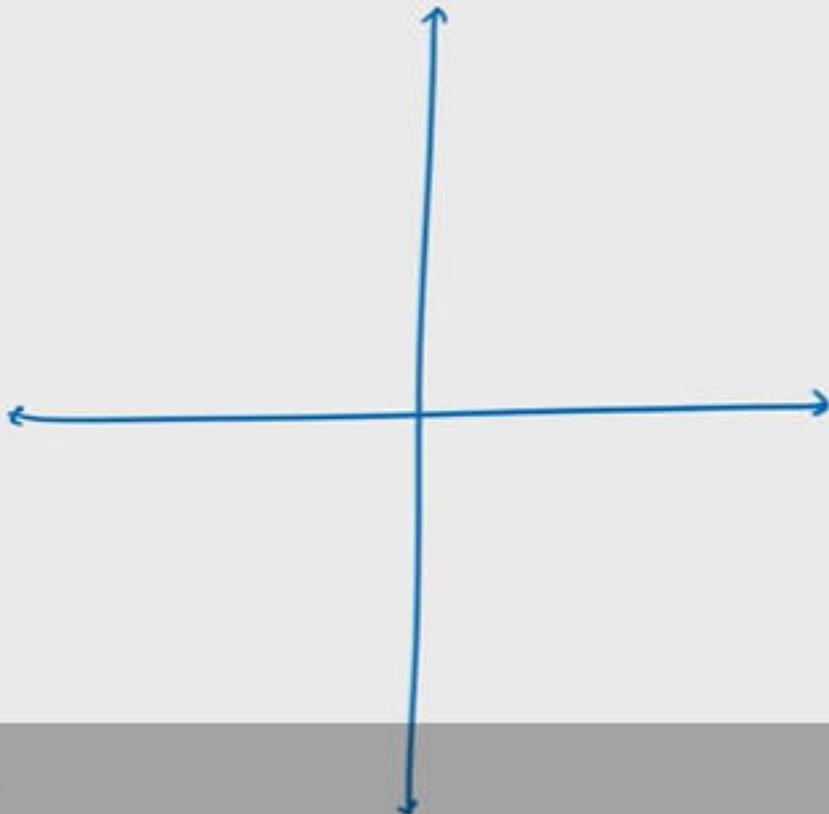
$$\| \vec{v} \|_1 = \left(\sum_{i=1}^n |v_i|^1 \right)^{\frac{1}{1}} = \sum_{i=1}^n |v_i| \rightarrow \begin{matrix} \text{"TAXICAB METRIC"} \\ \text{"MANHATTAN NORM"} \end{matrix}$$

$$\vec{v} = \begin{pmatrix} 3 \\ 2 \end{pmatrix}$$

$$\| \vec{v} \|_1 = \sqrt{3^1 + 2^1} = \sqrt{9+2} = \sqrt{13}$$



Geometry of Norms

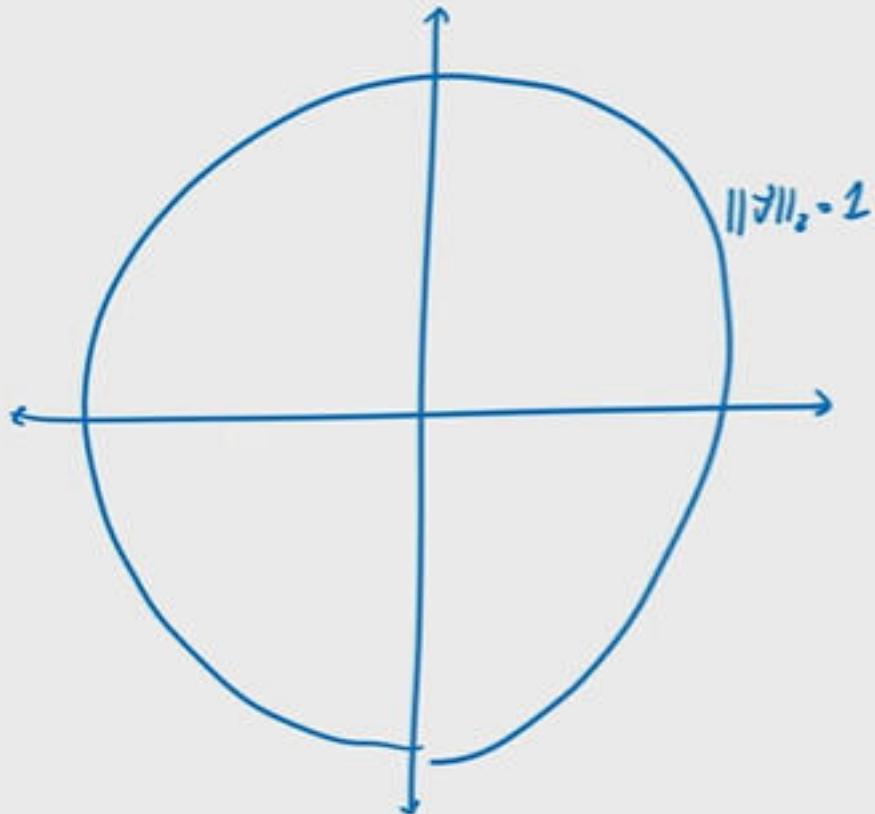


FOR A VARIETY OF NORMS
WHAT IS THE SET OF
VECTORS WITH $\|v\|=1$

- EUCLIDEAN



Geometry of Norms

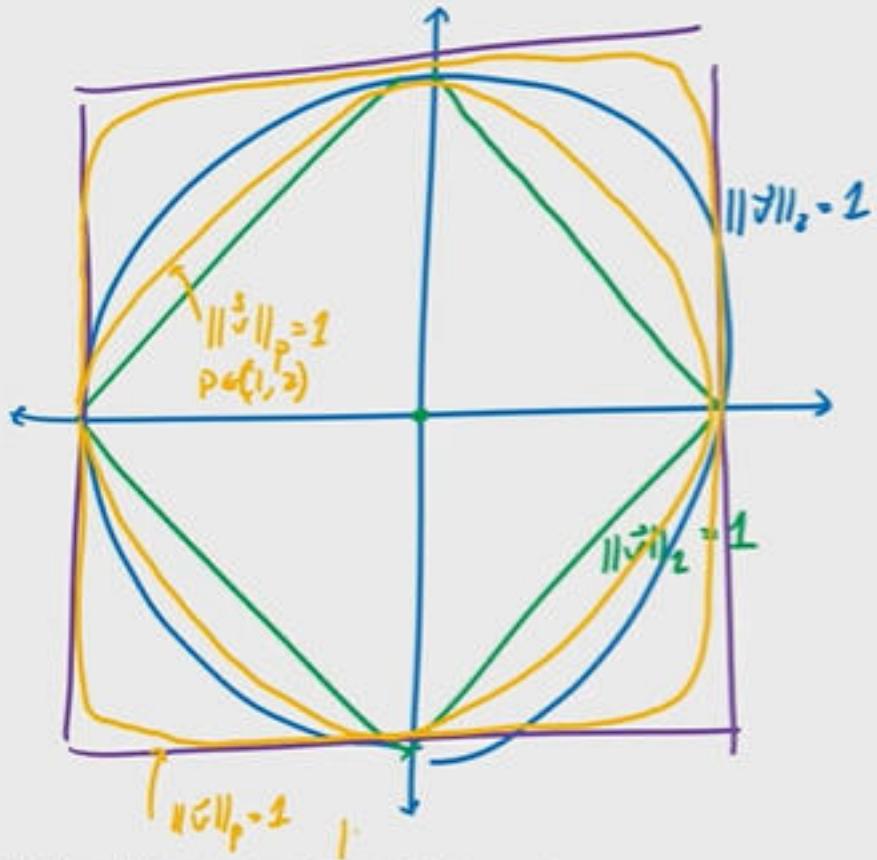


FOR A VARIETY OF NORMS
WHAT IS THE SET OF
VECTORS WITH $\|\vec{v}\| = 1$

- EUCLIDEAN $\|\vec{v}\|_2 = 1$

$$\sqrt{v_1^2 + v_2^2} = 1$$
$$\|\vec{v}\| = |v_1| + |v_2|$$

Geometry of Norms



FOR A VARIETY OF NORMS
WHAT IS THE SET OF
VECTORS WITH $\|v\|=1$

- EUCLIDEAN $\|\vec{v}\|_2 = 1$

$$\sqrt{v_1^2 + v_2^2} = 1$$

- L_1 $\|v\|_1 = |v_1| + |v_2| = 1$
 $v_1, v_2 \geq 0$
 $v_1 + v_2 = 1$
 $\Rightarrow v_2 = 1 - v_1$

- L_∞ $\|v\|_\infty = \max\{|v_1|, |v_2|\} = 1$
EITHER
 - $|v_1| = 1$ ($v_2 \leq 1$)
 - $|v_2| = 1$ ($v_1 \leq 1$)

L_0 -Norm



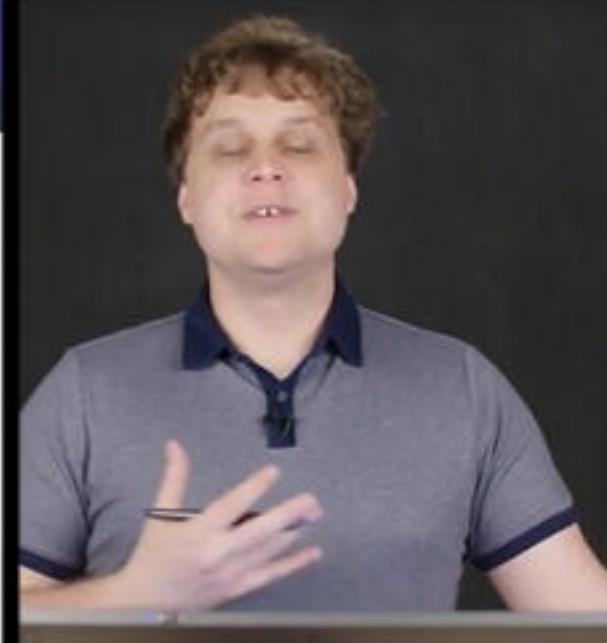
Despite the name, this is **not** a norm.

Defn: $\|\vec{v}\|_0 = \# \text{ OF } \text{NON-ZERO ELEMENTS} \text{ OF THE VECTOR } \vec{v}.$

$$\vec{v} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ -2 \\ -5 \\ 0 \\ 10 \end{pmatrix} \quad \|\vec{v}\|_0 = 4$$

$$\lim_{p \rightarrow \infty} \|\vec{v}\|_p^p = \|\vec{v}\|_0$$

$$\text{Also} \quad \|\alpha\vec{v}\|_0 = \|\vec{v}\|_0$$



machines using Python programming.

Topics



LATEX

Describing Math to People



Python

Describing Math to Machines



LATEX Pronunciation



LATEX is often pronounced *LAH-tekh*



This:

```
$\frac{d}{dx}f\left( x \right) = \mathop {\lim }\limits_{h \rightarrow 0} \frac{{f\left( {x + h} \right) - f\left( x \right)}}{h}$
```

Becomes This:

$$f'(x) = \lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h}$$

Input:

```
$y = \sqrt{x^3+2x^2-4x+7}$
```

Output:

$$y = \sqrt{x^3 + 2x^2 - 4x + 7}$$

Larger Equations in L^AT_EX

Input:

```
$$
\vec{v} =
\begin{bmatrix}
0 & 1 & 2
\end{bmatrix}
\\
M =
\begin{bmatrix}
0 & 2 & 4 \\
1 & 3 & 5 \\
4 & 5 & 6
\end{bmatrix}
$$
```

Output:

$$\vec{v} = \begin{bmatrix} 0 & 1 & 2 \end{bmatrix}$$

$$M = \begin{bmatrix} 0 & 2 & 4 \\ 1 & 3 & 5 \\ 4 & 5 & 6 \end{bmatrix}$$

Python: Describing Math to Machines

Python is the
lingua franca
of machine learning



Python Example



In [1]:

```
1 import numpy as np
2
3 # This defines a (row) vector
4 v = [1, 2, 3]
5
6 # This defines a matrix
7 A = [[1, 2, 3], [-1, 0, 1], [1, 1, 1]]
```



Python Example



In [1]:

```
1 import numpy as np
2
3 # This defines a (row) vector
4 v = [1, 2, 3]
5
6 # This defines a matrix
7 A = [[1, 2, 3], [-1, 0, 1], [1, 1, 1]]
```

np.array .

Python Example



In [2]: # L_p-Norms of vectors are done by the following:

```
1 print(np.linalg.norm(v,ord=1))  
2 print(np.linalg.norm(v,ord=2))  
3 print(np.linalg.norm(v,ord=np.inf))
```

```
6.0  
3.7416573867739413  
3.0
```



Python Example



In [3]:

```
1 # Python, in general, takes a different
2 # convention for matrix norms. Most will not
3 # do what you think they will from our
4 # notation. However, you can type the
5 # following for L_2 norms:
6
7 print(np.linalg.norm(A))
```

4.358898943540674



-0:23



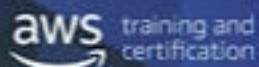
Motivation



- Motivating Example
- Definition



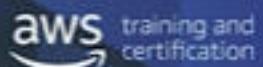
Motivation



- Motivating Example
- Definition



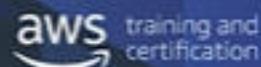
Motivation



- Motivating Example
- Definition



The Geometry of Dot Products



- Angles
- Key Consequences
 - Orthogonality: $\vec{v} \cdot \vec{w} = 0$
 - $\vec{v} \cdot \vec{w} > 0$
 - $\vec{v} \cdot \vec{w} < 0$
- Hyperplane Definition
 - Hyperplane
 - Decision plane
- Hyperplane Example



Product Properties



- Matrix Products
 - Distributivity
 - Associativity
 - Not commutativity



-1:16



Product Properties



- Matrix Products
 - Distributivity
 - Associativity
 - Not commutativity
- The Identity Matrix
- Properties of the Hadamard Product
 - Distributivity
 - Associativity
 - Commutativity



Determinate Computation



- The Two-By-Two
- Larger Matrices



© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.



-0:38



Invertibility



- When can you invert?
- How to compute the inverse



Motivating Example

Suppose you want to compute the **total number of families** that can live in Seattle given the **number of buildings** that hold **k families** for $k = 1, 2, \dots, 1000$.

$$\vec{v} = \begin{pmatrix} v_1 \\ \vdots \\ v_{1000} \end{pmatrix}$$

Motivating Example

Suppose you want to compute the total number of families that can live in Seattle given the number of buildings that hold k families for $k = 1, 2, \dots, 1000$.

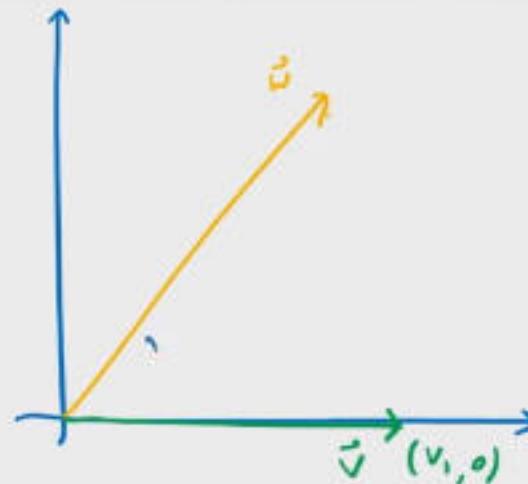
$$\vec{v} = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_{1000} \end{pmatrix} \leftarrow \# \text{ } 1 \text{ } \text{FAMILIES} \text{ } \text{BUILDINGS}$$
$$\# \text{FAMILIES}_j = 1 \cdot v_1 + 2 \cdot v_2 + \dots + 1000 \cdot v_{1000}$$
$$= \sum_{i=1}^{1000} i \cdot v_i$$
$$\leftarrow \# \text{ } 1000 \text{ } \text{FAMILIES} \text{ } \text{BUILDINGS}$$

Dot Products

Angles

$$\vec{v} = \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}$$

$$\vec{w} = \begin{pmatrix} w_1 \\ w_2 \end{pmatrix}$$



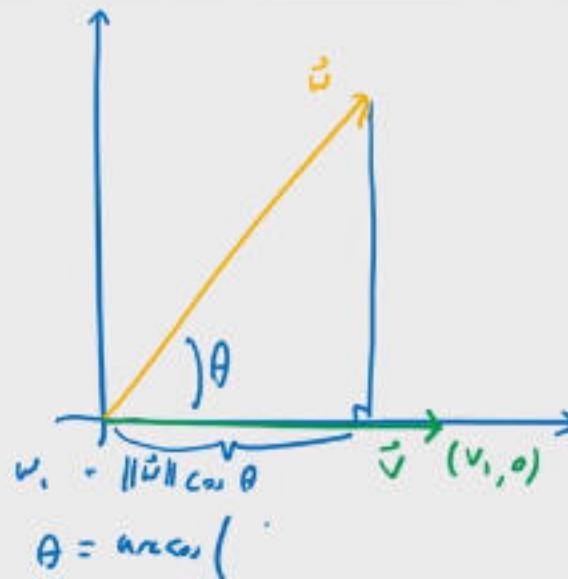
16:49



Angles

$$\vec{v} = \begin{pmatrix} v_1 \\ 0 \end{pmatrix}$$

$$\vec{w} = \begin{pmatrix} w_1 \\ w_2 \end{pmatrix}$$



$$v_1 = \|\vec{v}\| \cos \theta$$

$$\theta = \arccos \left(\frac{v_1}{\|\vec{v}\|} \right)$$

Angles

aws training and certification

$$\vec{v} = \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} \quad \vec{w} = \begin{pmatrix} w_1 \\ w_2 \end{pmatrix}$$

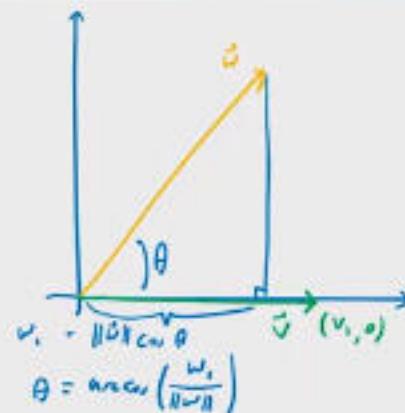
$$\vec{v} \cdot \vec{w} = v_1 \cdot w_1 + 0 \cdot w_2 = v_1 \cdot w_1$$

$$= v_1 \|\vec{w}\| \cos \theta$$

$$= \|\vec{v}\| \cdot \|\vec{w}\| \cos \theta$$

$$\vec{v} \cdot \vec{w} = \|\vec{v}\| \cdot \|\vec{w}\| \cos \theta$$

$$\theta = \arccos \left(\frac{\vec{v} \cdot \vec{w}}{\|\vec{v}\| \|\vec{w}\|} \right)$$



Orthogonality: $\vec{v} \cdot \vec{w} = 0$

 \vec{v}

WHAT

ARE

THE

VECTORS

 \vec{w}

ST.

$$\vec{v} \cdot \vec{w} = 0$$

0

 Assume $\|\vec{v}\|, \|\vec{w}\| \neq 0$

$$\|\vec{v}\| \cdot \|\vec{w}\| \cos \theta$$



Hyperplane

- IT IS THE THINK ORTHOGONAL TO A GIVEN VECTOR
- IN 2D:
- IN 3D:
- IN HIGHER: SIMILAR



-5:41



Hyperplane

- IT IS THE THINK ORTHOGONAL TO A GIVEN VECTOR
- IN 2D:
- OR A MEASURE TO A DIFFERENCE
- IN 3D:
- IN HIGHER: SIMILAR



-5:41



Decision Plane



$$\vec{w} = (v_1, \dots, v_n)$$

$$\vec{v} = \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix}$$

||

$$\vec{w} \cdot \vec{v} > c$$

$$c < \vec{w} \cdot \vec{v} = \|\vec{w}\| \cdot \|\vec{v}\| \cdot \cos \theta$$

\vec{w} fixed



Decision Plane

$$\vec{w} = (v_1 \dots v_n)$$

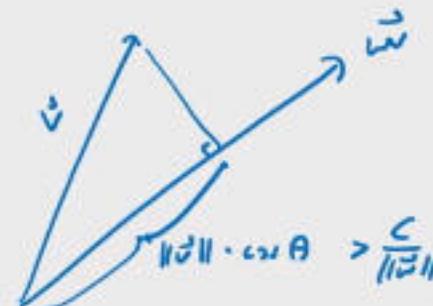
$$\vec{v} = \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix}$$

$$\vec{w} \cdot \vec{v} > c$$

$$c < \vec{w} \cdot \vec{v} = \|\vec{w}\| \cdot \|\vec{v}\| \cdot \cos \theta$$

\vec{w} FIXED

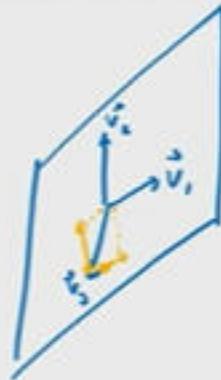
$$\|\vec{v}\| \cos \theta > \frac{c}{\|\vec{w}\|}$$



Definition

$\vec{v}_1, \vec{v}_2, \vec{v}_3$

3-Dm space



$$\vec{v}_3 = -\vec{v}_1 - \vec{v}_2$$

↑ ↑ ↑
two dimensions

$$\vec{v}_3 + \vec{v}_1 + \vec{v}_2 = \vec{0}$$

$$a_1 \vec{v}_1 + a_2 \vec{v}_2 + a_3 \vec{v}_3 = \vec{0}$$



-6:33



Example

$$\vec{v}_1 = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, \vec{v}_2 = \begin{pmatrix} 0 \\ -1 \\ -1 \end{pmatrix}, \vec{v}_3 = \begin{pmatrix} 1 \\ -2 \\ -1 \end{pmatrix}$$



Example

$$\vec{v}_1 = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, \vec{v}_2 = \begin{pmatrix} 0 \\ -1 \\ -1 \end{pmatrix}, \vec{v}_3 = \begin{pmatrix} 1 \\ -2 \\ -1 \end{pmatrix}$$

$a_1 \left(\begin{array}{c} \cdot \\ \cdot \\ \cdot \end{array} \right)$

Example

$$\vec{v}_1 = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, \vec{v}_2 = \begin{pmatrix} 0 \\ -1 \\ -1 \end{pmatrix}, \vec{v}_3 = \begin{pmatrix} 1 \\ -2 \\ -1 \end{pmatrix}$$

$$a_1 \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} + a_2 \begin{pmatrix} 0 \\ -1 \\ -1 \end{pmatrix} + a_3 \begin{pmatrix} 1 \\ -2 \\ -1 \end{pmatrix} = \vec{0}$$

$$\left\{ \begin{array}{l} a_1 + a_3 = 0 \\ -a_2 - 2a_3 = 0 \\ a_1 - a_2 - a_3 = 0 \end{array} \right. \quad \left. \begin{array}{l} a_3 = -a_1 \\ 2a_3 = -a_2 \\ a_1 = a_2 + a_3 \end{array} \right.$$

Example

$$\vec{v}_1 = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, \vec{v}_2 = \begin{pmatrix} 0 \\ -1 \\ -1 \end{pmatrix}, \vec{v}_3 = \begin{pmatrix} 1 \\ -2 \\ -1 \end{pmatrix}$$

$$a_1 \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} + a_2 \begin{pmatrix} 0 \\ -1 \\ -1 \end{pmatrix} + a_3 \begin{pmatrix} 1 \\ -2 \\ -1 \end{pmatrix} = \vec{0}$$

$$\begin{cases} a_1 + a_3 = 0 \\ -a_2 - 2a_3 = 0 \\ a_1 - a_2 - a_3 = 0 \end{cases} \quad \leftarrow \quad \begin{array}{l} a_3 = -a_1 \\ 2a_3 = -a_2 \end{array} \quad \Rightarrow \quad \begin{array}{l} a_3 = 1 \\ a_2 = -2 \\ a_1 = -1 \end{array}$$

$$-1 \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} - 2 \begin{pmatrix} 0 \\ -1 \\ -1 \end{pmatrix} + \begin{pmatrix} 1 \\ -2 \\ -1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} = \vec{0}$$



Example

$$\vec{v}_1 = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, \vec{v}_2 = \begin{pmatrix} 0 \\ -1 \\ -1 \end{pmatrix}, \vec{v}_3 = \begin{pmatrix} 1 \\ -2 \\ -1 \end{pmatrix}$$

$$a_1 \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} + a_2 \begin{pmatrix} 0 \\ -1 \\ -1 \end{pmatrix} + a_3 \begin{pmatrix} 1 \\ -2 \\ -1 \end{pmatrix} = \vec{0}$$

$$\left\{ \begin{array}{l} a_1 + a_3 = 0 \\ -a_2 - 2a_3 = 0 \\ a_1 - a_2 - a_3 = 0 \end{array} \right. \quad \left. \begin{array}{l} a_3 = -a_1 \\ 2a_3 = -a_2 \\ a_1 = a_2 \end{array} \right. \Rightarrow \begin{array}{l} a_3 = 1 \\ a_2 = -2 \\ a_1 = -1 \end{array}$$

$$-1 \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} - 2 \begin{pmatrix} 0 \\ -1 \\ -1 \end{pmatrix} + \begin{pmatrix} 1 \\ -2 \\ -1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} = \vec{0}$$



Example

$$\vec{v}_1 = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, \vec{v}_2 = \begin{pmatrix} 0 \\ -1 \\ -1 \end{pmatrix}, \vec{v}_3 = \begin{pmatrix} 1 \\ -2 \\ -1 \end{pmatrix}$$

$$a_1 \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} + a_2 \begin{pmatrix} 0 \\ -1 \\ -1 \end{pmatrix} + a_3 \begin{pmatrix} 1 \\ -2 \\ -1 \end{pmatrix} = \vec{0}$$

$$\left\{ \begin{array}{l} a_1 + a_3 = 0 \\ -a_2 - 2a_3 = 0 \\ a_1 - a_2 - a_3 = 0 \end{array} \right. \quad \left. \begin{array}{l} \xleftarrow{\hspace{1cm}} a_3 = -a_1 \\ \xleftarrow{\hspace{1cm}} 2a_3 = -a_2 \\ \xleftarrow{\hspace{1cm}} a_1 = a_2 \end{array} \right. \Rightarrow \begin{array}{l} a_3 = 1 \\ a_2 = -2 \\ a_1 = -1 \end{array}$$

$$-1 \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} - 2 \begin{pmatrix} 0 \\ -1 \\ -1 \end{pmatrix} + \begin{pmatrix} 1 \\ -2 \\ -1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} = \vec{0}$$

Linear Dependence

Definition

$\vec{v}_1, \vec{v}_2, \vec{v}_3$

3-Dm space



$$\vec{v}_3 = -\vec{v}_1 - \vec{v}_2$$

↑ ↑
two dimensions

$$\vec{v}_3 + \vec{v}_1 + \vec{v}_2 = 0$$

Definition

$\vec{v}_1, \vec{v}_2, \vec{v}_3$

3-Dm space



$$\vec{v}_3 = -\vec{v}_1 - \vec{v}_2$$

↑ ↑
two dimensions

$$\vec{v}_3 + \vec{v}_1 + \vec{v}_2 = 0$$

$$a_1 \vec{v}_1 + a_2 \vec{v}_2 + a_3 \vec{v}_3$$



-6:45



Example

$$\vec{v}_1 = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, \vec{v}_2 = \begin{pmatrix} 0 \\ -1 \\ -1 \end{pmatrix}, \vec{v}_3 = \begin{pmatrix} 1 \\ -2 \\ -1 \end{pmatrix}$$

$$a_1 \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} + a_2 \begin{pmatrix} 0 \\ -1 \\ -1 \end{pmatrix} + a_3 \begin{pmatrix} 1 \\ -2 \\ -1 \end{pmatrix} = \vec{0}$$



4:12



Example

$$\vec{v}_1 = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, \vec{v}_2 = \begin{pmatrix} 0 \\ -1 \\ -1 \end{pmatrix}, \vec{v}_3 = \begin{pmatrix} 1 \\ -2 \\ -1 \end{pmatrix}$$

$$a_1 \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} + a_2 \begin{pmatrix} 0 \\ -1 \\ -1 \end{pmatrix} + a_3 \begin{pmatrix} 1 \\ -2 \\ -1 \end{pmatrix} = \vec{0}$$

$$\begin{cases} a_1 + a_3 = 0 \\ -a_2 - 2a_3 = 0 \\ a_1 - a_2 - a_3 = 0 \end{cases}$$



Example

$$\vec{v}_1 = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, \vec{v}_2 = \begin{pmatrix} 0 \\ -1 \\ -1 \end{pmatrix}, \vec{v}_3 = \begin{pmatrix} 1 \\ -2 \\ -1 \end{pmatrix}$$

$$a_1 \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} + a_2 \begin{pmatrix} 0 \\ -1 \\ -1 \end{pmatrix} + a_3 \begin{pmatrix} 1 \\ -2 \\ -1 \end{pmatrix} = \vec{0}$$

$$\left\{ \begin{array}{l} a_1 + a_3 = 0 \\ -a_2 - 2a_3 = 0 \\ a_1 - a_2 - a_3 = 0 \end{array} \right. \quad \xleftarrow{a_3 = -a_1}$$



3:11



Example

$$\vec{v}_1 = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, \vec{v}_2 = \begin{pmatrix} 0 \\ -1 \\ -1 \end{pmatrix}, \vec{v}_3 = \begin{pmatrix} 1 \\ -2 \\ -1 \end{pmatrix}$$

$$a_1 \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} + a_2 \begin{pmatrix} 0 \\ -1 \\ -1 \end{pmatrix} + a_3 \begin{pmatrix} 1 \\ -2 \\ -1 \end{pmatrix} = \vec{0}$$

$$\left\{ \begin{array}{l} a_1 + a_3 = 0 \\ -a_2 - 2a_3 = 0 \\ a_1 - a_2 - a_3 = 0 \end{array} \right. \xleftarrow{\quad} \left. \begin{array}{l} a_3 = -a_1 \\ 2a_3 = -a_2 \\ a_1 - a_2 - a_3 = 0 \end{array} \right. \Rightarrow \begin{array}{l} a_3 = 1 \\ a_2 = -2 \\ a_1 = -1 \end{array}$$

$$-1 \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} - 2 \begin{pmatrix} 0 \\ -1 \\ -1 \end{pmatrix} + \begin{pmatrix} 1 \\ -2 \\ -1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$



-1:53



Example

$$\vec{v}_1 = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, \vec{v}_2 = \begin{pmatrix} 0 \\ -1 \\ -1 \end{pmatrix}, \vec{v}_3 = \begin{pmatrix} 1 \\ -2 \\ -1 \end{pmatrix}$$

$$a_1 \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} + a_2 \begin{pmatrix} 0 \\ -1 \\ -1 \end{pmatrix} + a_3 \begin{pmatrix} 1 \\ -2 \\ -1 \end{pmatrix} = \vec{0}$$

$$\begin{cases} a_1 + a_3 = 0 \\ -a_2 - 2a_3 = 0 \\ a_1 - a_2 - a_3 = 0 \end{cases} \xleftarrow{\quad 2a_3 = -a_2 \quad} \begin{array}{l} a_3 = -a_1 \\ a_2 = -2a_3 \\ a_1 - a_2 - a_3 = 0 \end{array} \Rightarrow \begin{array}{l} a_3 = 1 \\ a_2 = -2 \\ a_1 = -1 \end{array}$$

$$-1 \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} - 2 \begin{pmatrix} 0 \\ -1 \\ -1 \end{pmatrix} + \begin{pmatrix} 1 \\ -2 \\ -1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} = \vec{0}$$

$A \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \text{FIRST COLUMN OF } A$

$A \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix} = \text{SECOND COLUMN}$



The Two-by-Two

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \xrightarrow{A}$$

$\det(A)$ = scaling factor


$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \xrightarrow{A}$$



4:15



The Two-by-Two

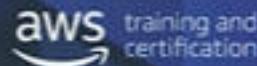
$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \xrightarrow{A}$$

$\det(A)$ = scaling factor

$$A \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \xrightarrow{A}$$

The Two-by-Two



$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \xrightarrow{A} \begin{matrix} A(1) \\ A(2) \end{matrix}$

$\det(A) = \text{scalars ratio}$



$$\det(A) = ad - bc$$



Larger Matrices

$$\det(A) = \det \begin{pmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mm} \end{pmatrix} = \underline{a_{11}} \cdot \det \underbrace{\begin{pmatrix} a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mm} \end{pmatrix}}_{\substack{(m-1) \times (m-1) \\ \text{matrix}}} - \underline{a_{12}} \cdot \det \begin{pmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1m} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2m} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & a_{m3} & \cdots & a_{mm} \end{pmatrix}$$

(square matrix: # rows = # columns)

$$+ \underline{a_{13}} \cdot \det \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} & \cdots & a_{1m} \\ a_{21} & a_{22} & a_{23} & a_{24} & \cdots & a_{2m} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & a_{m3} & a_{m4} & \cdots & a_{mm} \end{pmatrix} + \cdots + (-1)^{(m+1)} \underline{a_{1m}} \det \begin{pmatrix} a_{11} & \cdots & a_{1(m-1)} & a_{1m} \\ a_{21} & \cdots & a_{2(m-1)} & a_{2m} \\ \vdots & \ddots & \vdots & \vdots \\ a_{m1} & \cdots & a_{m(m-1)} & a_{mm} \end{pmatrix}$$

Larger Matrices

$$\det(A) = \det \begin{pmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mm} \end{pmatrix} = \underline{a_{11}} \cdot \det \underbrace{\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \ddots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mm} \end{pmatrix}}_{\substack{(m-1) \times (m-1) \\ \text{matrix}}} - \underline{a_{12}} \cdot \det \begin{pmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1m} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2m} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ a_{m1} & a_{m2} & a_{m3} & \cdots & a_{mm} \end{pmatrix}$$

(square matrix: # rows = # columns)

$$+ a_{13} \cdot \det \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} & \cdots & a_{1m} \\ a_{21} & a_{22} & a_{23} & a_{24} & \cdots & a_{2m} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & a_{m3} & a_{m4} & \cdots & a_{mm} \end{pmatrix} + \cdots + (-1)^{(m+1)} a_{1m} \det \begin{pmatrix} a_{11} & \cdots & a_{1(m-1)} & a_{1m} \\ a_{21} & \cdots & a_{2(m-1)} & a_{2m} \\ \vdots & \ddots & \vdots & \vdots \\ a_{m1} & \cdots & a_{m(m-1)} & a_{mm} \end{pmatrix}$$

Larger Matrices

$$\det(A) = \det \begin{pmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mm} \end{pmatrix} = \underline{a_{11}} \cdot \det \underbrace{\begin{pmatrix} a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \ddots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mm} \end{pmatrix}}_{(m-1) \times (m-1) \text{ matrix}} - \underline{a_{12}} \cdot \det \begin{pmatrix} a_{11} & a_{13} & \cdots & a_{1m} \\ a_{21} & a_{23} & \cdots & a_{2m} \\ \vdots & \ddots & \ddots & \vdots \\ a_{m1} & a_{m3} & \cdots & a_{mm} \end{pmatrix}$$

(square matrix: # rows = # columns)

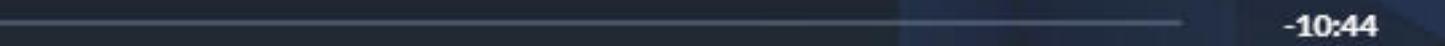
$$+ \underline{a_{13}} \cdot \det \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} & \cdots & a_{1m} \\ a_{21} & a_{22} & a_{23} & a_{24} & \cdots & a_{2m} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & a_{m3} & a_{m4} & \cdots & a_{mm} \end{pmatrix} + \cdots + (-1)^{(m+1)} \underline{a_{1m}} \det \begin{pmatrix} a_{11} & \cdots & a_{1(m-1)} & a_{1m} \\ a_{21} & \cdots & a_{2(m-1)} & a_{2m} \\ \vdots & \ddots & \vdots & \vdots \\ a_{m1} & \cdots & a_{m(m-1)} & a_{mm} \end{pmatrix}$$

m × m determinant → m (m-1) × (m-1) determinants

$n!$ TIME ALGORITHM

→ ACTUAL RESULT

Geometry of Matrix Operations



Topics



- Intuition from Two Dimensions
- The Determinant



Intuition from Two Dimensions

Suppose A is a 2×2 matrix (mapping \mathbb{R}^2 to itself). Any such matrix can be expressed uniquely as a **stretching**, followed by a **skewing**, followed by a **rotation**.

$$\textcircled{1} \quad \vec{v} = \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} v_1 \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ v_2 \end{pmatrix} = v_1 \begin{pmatrix} 1 \\ 0 \end{pmatrix} + v_2 \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

*↑
ANY VECTOR
[
SCALAR MULTIPLES OF TWO SPECIFIC VECTORS
/ SUM /*

(2)

Intuition from Two Dimensions

Suppose A is a 2×2 matrix (mapping \mathbb{R}^2 to itself). Any such matrix can be expressed uniquely as a **stretching**, followed by a **skewing**, followed by a **rotation**.

$$\textcircled{1} \quad \vec{v} = \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} v_1 \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ v_2 \end{pmatrix} = v_1 \begin{pmatrix} 1 \\ 0 \end{pmatrix} + v_2 \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

\uparrow ANY VECTOR \uparrow SCALAR MULTIPLES OF TWO SPECIFIC VECTORS
 / SUM /

$$\textcircled{2} \quad A\vec{v} = A \left(v_1 \begin{pmatrix} 1 \\ 0 \end{pmatrix} + v_2 \begin{pmatrix} 0 \\ 1 \end{pmatrix} \right) = v_1 \left(A \begin{pmatrix} 1 \\ 0 \end{pmatrix} \right) + v_2 \left(A \begin{pmatrix} 0 \\ 1 \end{pmatrix} \right)$$

\uparrow A APPL.

Intuition from Two Dimensions

Suppose A is a 2×2 matrix (mapping \mathbb{R}^2 to itself). Any such matrix can be expressed uniquely as a **stretching**, followed by a **skewing**, followed by a **rotation**.

$$\textcircled{1} \quad \vec{v} = \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} v_1 \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ v_2 \end{pmatrix} = v_1 \begin{pmatrix} 1 \\ 0 \end{pmatrix} + v_2 \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

↑
ANY VECTOR

[↑
SCALAR MULTIPLES OF TWO SPECIFIC VECTORS
] /

$$\textcircled{2} \quad A\vec{v} = A \left(v_1 \begin{pmatrix} 1 \\ 0 \end{pmatrix} + v_2 \begin{pmatrix} 0 \\ 1 \end{pmatrix} \right) = v_1 (A \begin{pmatrix} 1 \\ 0 \end{pmatrix}) + v_2 (A \begin{pmatrix} 0 \\ 1 \end{pmatrix})$$

↑
P APPLED

Intuition from Two Dimensions

Suppose A is a 2×2 matrix (mapping \mathbb{R}^2 to itself). Any such matrix can be expressed uniquely as a **stretching**, followed by a **skewing**, followed by a **rotation**.

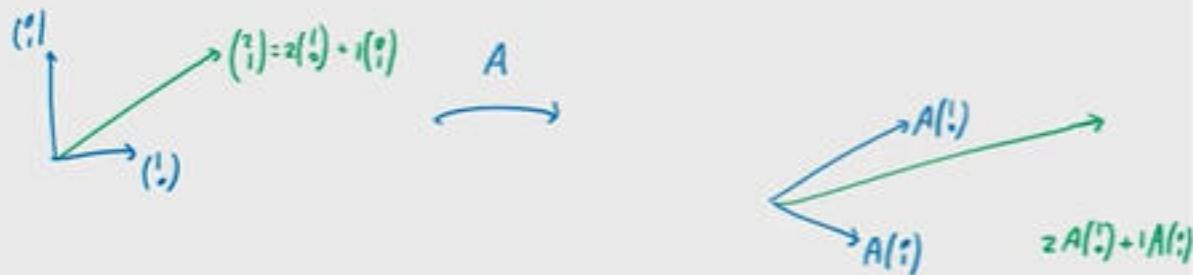
$$\textcircled{1} \quad \vec{v} = \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} v_1 \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ v_2 \end{pmatrix} = v_1 \begin{pmatrix} 1 \\ 0 \end{pmatrix} + v_2 \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

↑
ANY VECTOR
↑
SCALAR MULTIPLES OF TWO SPECIFIC VECTORS
[]
sum
/ \

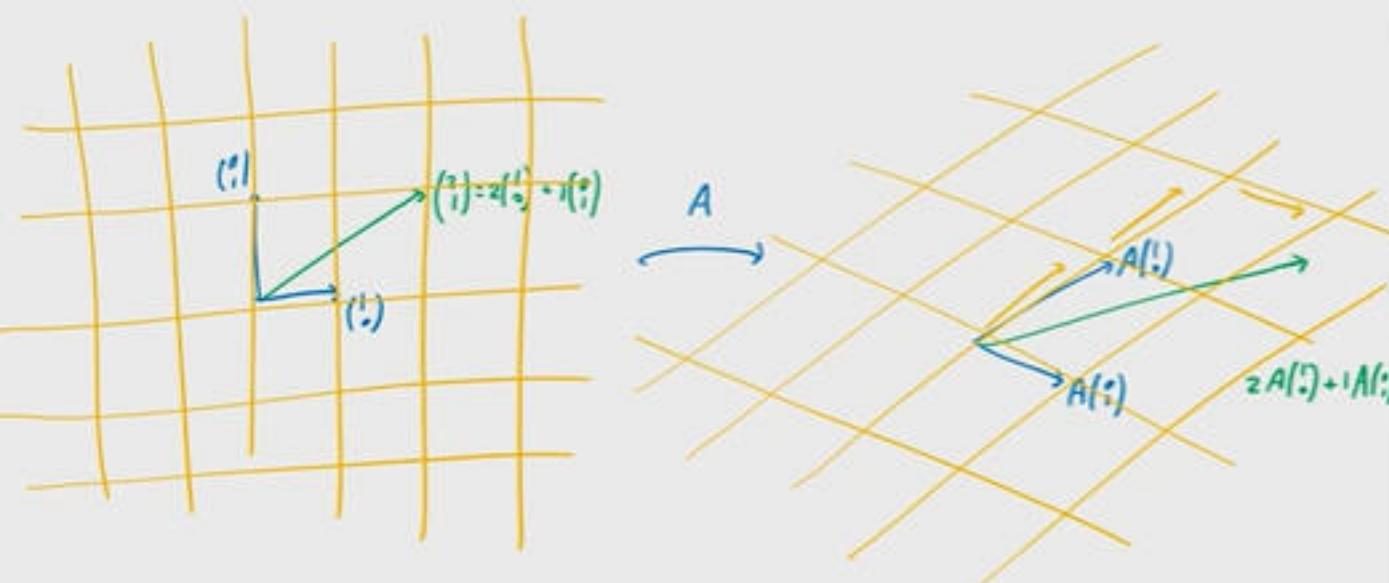
$$\textcircled{2} \quad A\vec{v} = A \left(v_1 \begin{pmatrix} 1 \\ 0 \end{pmatrix} + v_2 \begin{pmatrix} 0 \\ 1 \end{pmatrix} \right) = v_1 (A \begin{pmatrix} 1 \\ 0 \end{pmatrix}) + v_2 (A \begin{pmatrix} 0 \\ 1 \end{pmatrix})$$

↑
P APPLIED TO ANY VECTOR
↑
A APPLIED TO TWO SPECIFIC VECTORS

Intuition from Two Dimensions



Intuition from Two Dimensions

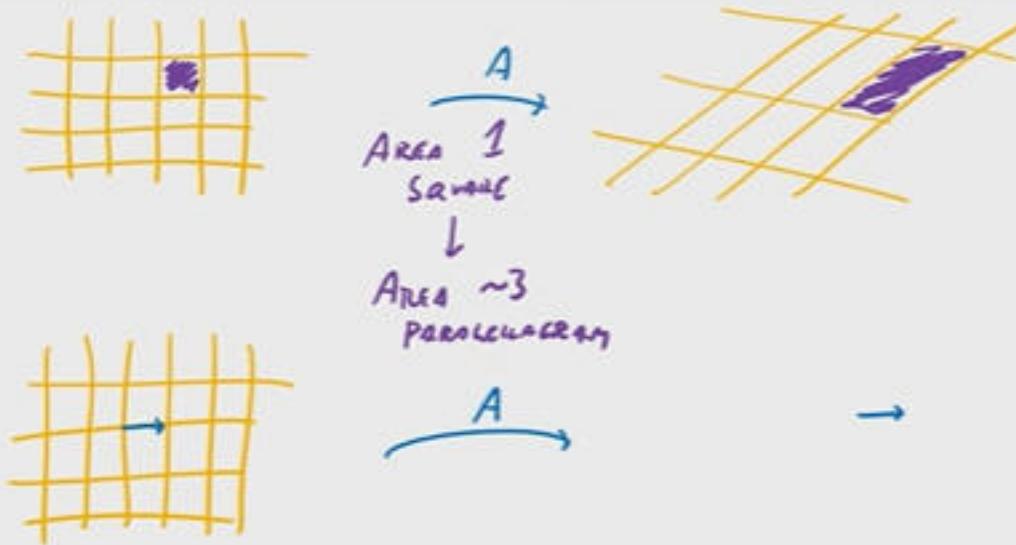


© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

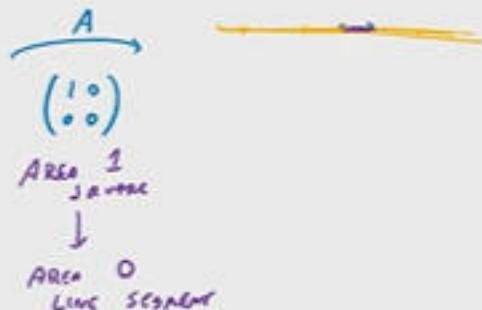
-4:32



The Determinant



The Determinant

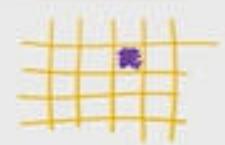


$\text{det}(A)$ is THE
FACTOR THE AREA
IS MULTIPLIED BY.

$\text{det}(A)$ is NEGATIVE
IF IT FLIPS THE
PLANE OVER



The Determinant



$\overset{A}{\curvearrowright}$
Area 1
square
 \downarrow
Area ~3
parallelogram



$\overset{A}{\curvearrowright}$
 $(1 \ 0)$
 $(0 \ 0)$
Area 1
square
 \downarrow
Area 0
line segment



$\text{det}(A)$ is THE
FACTOR THE AREA
IS MULTIPLIED BY.

$\text{det}(A)$ is NEGATIVE
IF IT FLIPS THE
PLANE OVER



Topics



- Matrix Products
 - Distributativity
 - Associativity
 - Not commutativity



© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.



-10:16



multiplication is commutative if - and only when - both of the matrices are diagonal and of equal dimensions.

Not Commutativity



$$A = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \quad B = \begin{pmatrix} 1 & 1 \end{pmatrix}$$

$$AB = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \begin{pmatrix} 1 & 1 \end{pmatrix} = \begin{pmatrix} 1 \cdot 1 & 1 \cdot 1 \\ 1 \cdot 1 & 1 \cdot 1 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$$

$$BA = \begin{pmatrix} 1 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} =$$

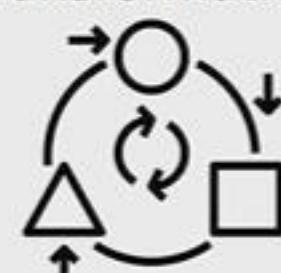
multiplication is commutative if - and only when - both of the matrices are diagonal and of equal dimensions.

Matrix Multiplication



The best way to think about matrix multiplication is as a

transformation



of your data



multiplication is commutative if - and only when - both of the matrices are diagonal and of equal dimensions.

Put on your Socks, then your Shoes



It is very different to:

put on your socks,



then your shoes

put on your shoes,



then your socks

than it is to



multiplication is commutative if - and only when - both of the matrices are diagonal and of equal dimensions.

The Identity Matrix

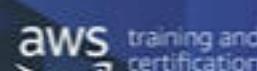


$$I A = A$$

$$\rightarrow \begin{pmatrix} & \end{pmatrix} \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

multiplication is commutative if - and only when - both of the matrices are diagonal and of equal dimensions.

The Identity Matrix



$$IA = A$$
$$\Rightarrow \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$



multiplication is commutative if - and only when - both of the matrices are diagonal and of equal dimensions.

The Identity Matrix

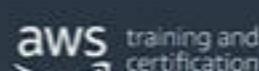


$$I A = A$$

$$\rightarrow \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

$$\begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & & & & \end{pmatrix} A$$

multiplication is commutative if - and only when - both of the matrices are diagonal and of equal dimensions.

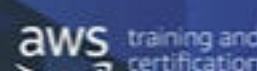


Properties of the Hadamard Product

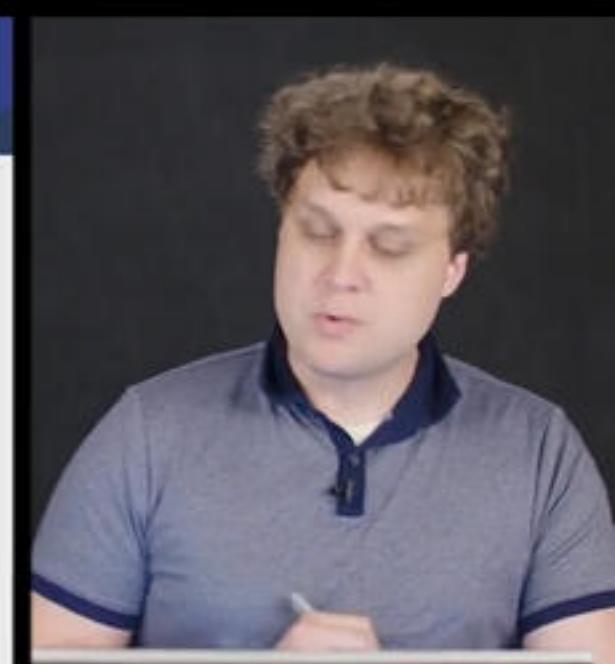


multiplication is commutative if - and only when - both of the matrices are diagonal and of equal dimensions.

Distributativity



$$A \cdot (B + C) = A \cdot B + A \cdot C$$

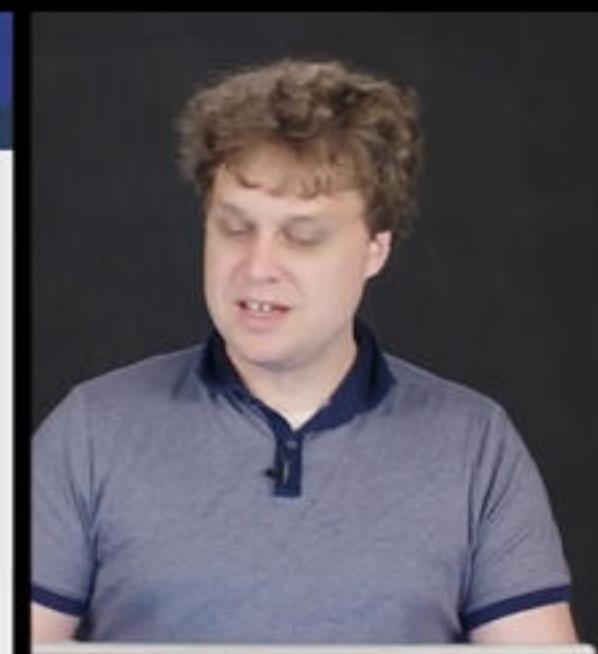


multiplication is commutative if - and only when - both of the matrices are diagonal and of equal dimensions.

Associativity

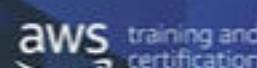


$$A \circ (B \circ C) = (A \circ B) \circ C$$



multiplication is commutative if - and only when - both of the matrices are diagonal and of equal dimensions.

Associativity



$$A \circ (B \circ C) = (A \circ B) \circ C$$



Probability Definitions



- Example
- Definitions
 - Outcome
 - Sample space
 - Event
 - Probability



Visualizing Probability



- General Picture



Entropy



- Intuition
- Three Examples
 - One coin
 - Two coins
 - A mixed case
- Examine the Trees



© 2018, Amazon Web Services, Inc., or its Affiliates. All rights reserved.



-1:58



Summary Statistics



- Definitions
 - Random variable
 - Expected value
 - Variance
 - Standard deviation
- Chebyshev's Inequality



The Gaussian



- Standard Gaussian (Normal Distribution) Density



© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.



-1:17



The Gaussian



- Standard Gaussian (Normal Distribution) Density
- General Gaussian Density
- Key Properties
 - Maximum entropy distribution
 - Central limit theorem



© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.



-1:01



The Gaussian



- Standard Gaussian (Normal Distribution) Density
- General Gaussian Density
- Key Properties
 - Maximum entropy distribution
 - Central limit theorem



Thank You

© 2016 Amazon Web Services, Inc. or its affiliates. All rights reserved. This work may not be reproduced or redistributed, in whole or in part, without prior written permission from Amazon Web Services, Inc. Commercial copying, lending, or selling is prohibited. Corrections or feedback on the course, please email us at aws-course-feedback@amazon.com. For all other questions, contact us at <https://aws.amazon.com/contact-us/training/>. All trademarks are the property of their owners.



-0:01



Topics



- Example
- Definitions
 - Outcome
 - Sample space
 - Event
 - Probability



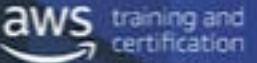
© 2016, Amazon Web Services, Inc. or its Affiliates. All rights reserved.



-6:08



Coin Flip Example



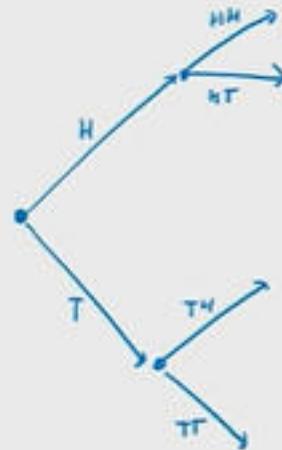
Suppose I flip three coins. What is the probability that I get **exactly** two heads?



Coin Flip Example



Suppose I flip three coins. What is the probability that I get **exactly** two heads?



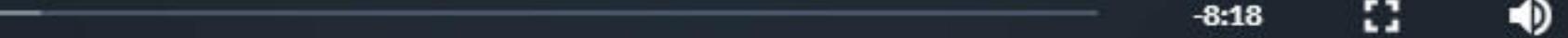
© 2016, Amazon Web Services, Inc. or its Affiliates. All rights reserved.



-5:08



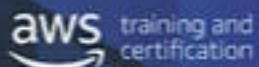
Axioms



8:18



Axiom #1



The fraction of the times an event occurs is between 0 and 1.



Axiom #2

Something always happens.

Sample space $\Omega = \{\text{set of } n\}$



6:26



In this video, you'll learn the three basic rules that every probability function satisfies, and how you use them to prove probability theorems.



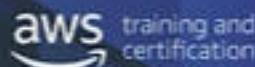
Axiom #3

If two events can't happen at the same time, then the fraction of the time that one of them occurs is the sum of the fraction of the time either one occurs separately.

E_1 E_2
" "

AT LEAST

Axiom #3 – Infinite Events



Also works for any number of events (including countably infinite)

$$\{E_i\}$$

$$E_i \cap E_j = \emptyset \quad \text{for } i \neq j$$

(all pairs are disjoint)

$$P\{\bigcup_i E_i\} = \sum_i P\{E_i\}$$



General Picture



$\frac{P_{KOB}}{\mathcal{G}}$



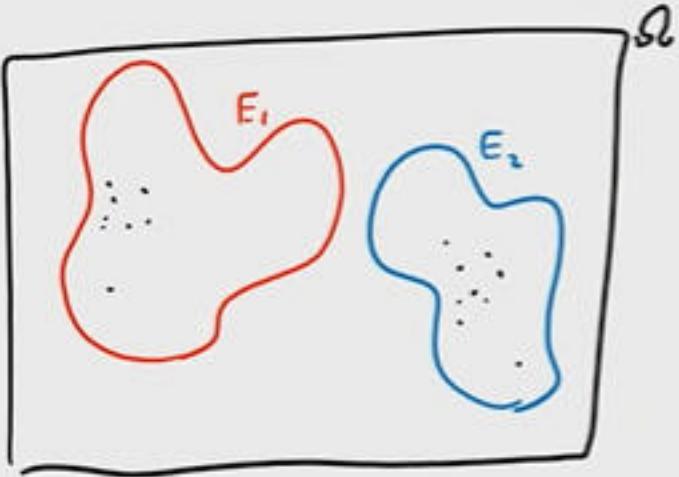
$\frac{\text{Geometri}}{\text{Recurs}}$



7:49

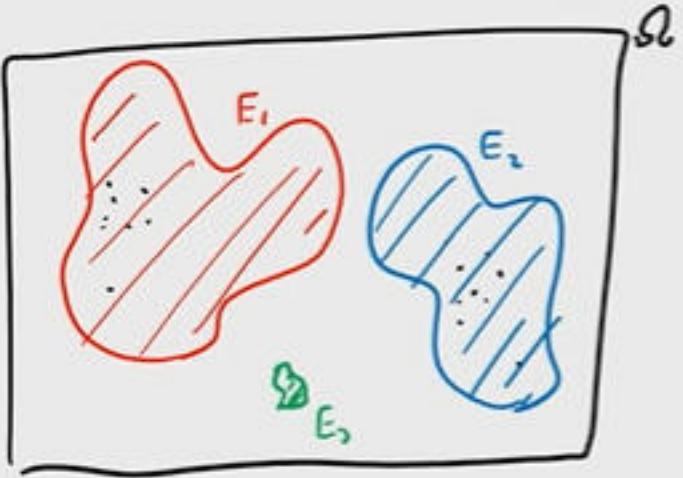


General Picture



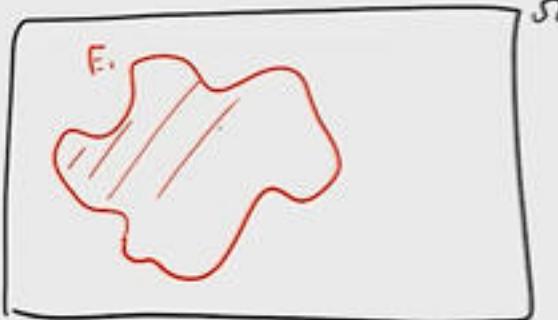
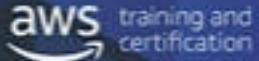
| <u>Prob</u> | <u>Geometri</u> |
|--------------|---------------------|
| Ω | Regions |
| outcomes | points |
| events | subregions |
| joint events | disjoint subregions |
| P | . |

General Picture



| $\frac{P_{KOB}}{\Omega}$ | \leftrightarrow | <u>Geometri</u> |
|--------------------------|-------------------|---------------------|
| OUTCOME | \leftrightarrow | REGIONS |
| EVENTS | \leftrightarrow | POINTS |
| DISJOINT EVENTS | \leftrightarrow | SUBREGIONS |
| P | \rightarrow | DISJOINT SUBREGIONS |
| $P\{\cap\} = 1$ | \leftrightarrow | AREA |
| | | \cap |

Inclusion-Exclusion



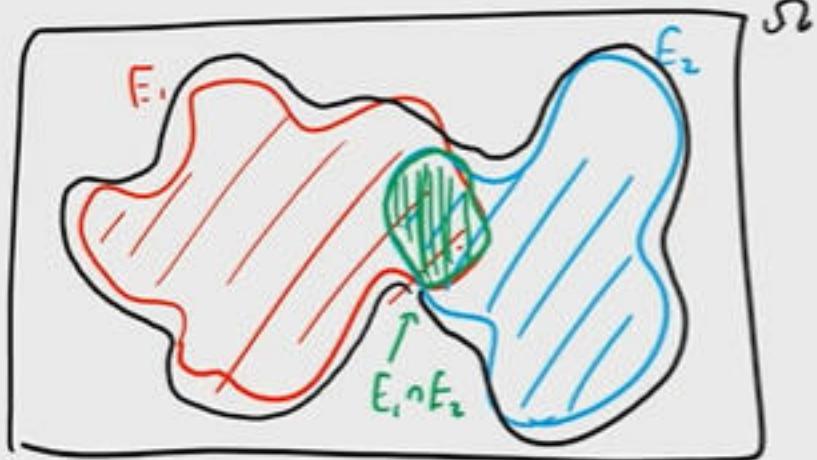
$$P\{E_1 \cup E_2\} = P\{E_1\} + P\{E_2\}$$

when $E_1 \cap E_2 = \emptyset$

$$P\{E_1 \cup E_2\} = ?$$

when $E_1 \cap E_2 \neq \emptyset$?





$$P\{E_1 \cup E_2\} = P\{E_1\} + P\{E_2\}$$

when $E_1 \cap E_2 = \emptyset$

$$P\{E_1 \cup E_2\} = ?$$

when $\text{not } E_1 \cap E_2 = \emptyset$?

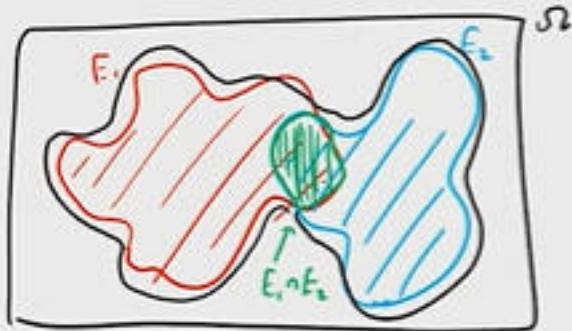
$$P\{E_1 \cup E_2\} = P\{E_1\} + P\{E_2\}$$



-2:25



Inclusion-Exclusion



$$P\{E_1 \cup E_2\} = P\{E_1\} + P\{E_2\}$$

$$P\{E_1 \cup E_2\} = P\{E_1\} + P\{E_2\}$$

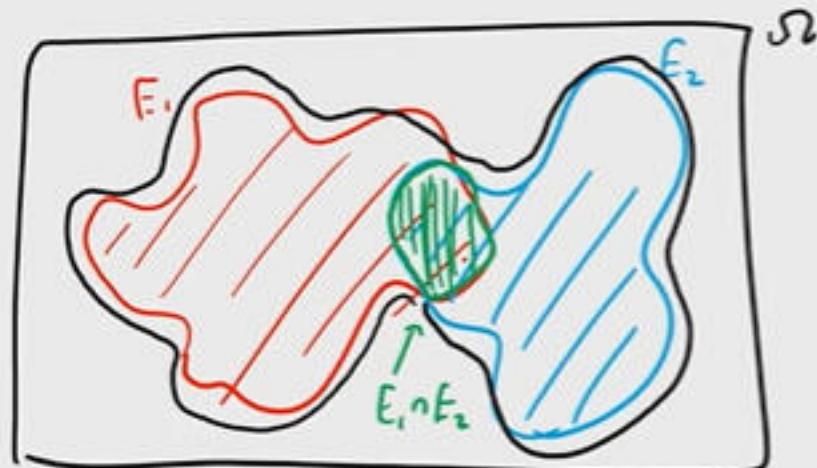
when $E_1 \cap E_2 = \emptyset$

$$P\{E_1 \cup E_2\} = ?$$

when $E_1 \cap E_2 \neq \emptyset$?



Inclusion-Exclusion



$$P\{E_1 \cup E_2\} = P\{E_1\} + P\{E_2\}$$

when $E_1 \cap E_2 = \emptyset$

$$P\{E_1 \cup E_2\} = ?$$

when not?

$$P\{E_1 \cup E_2\} = P\{E_1\} + P\{E_2\} - P\{E_1 \cap E_2\}$$

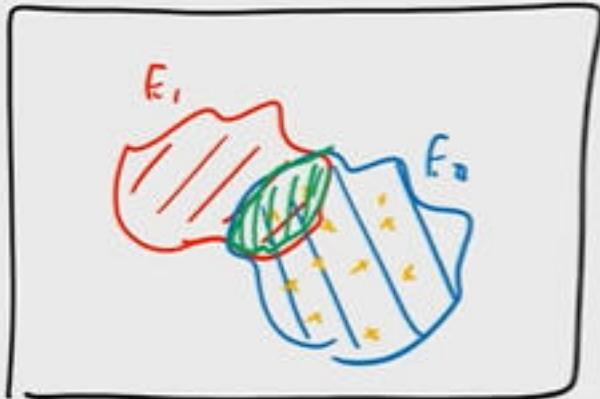
$$P\{E_1 \cup E_2 \cup E_3\} = P\{E_1\} + P\{E_2\} + P\{E_3\} - P\{E_1 \cap E_2\} - P\{E_1 \cap E_3\}$$



-0:49



Motivation: Partial Information



S

If I know E_2 occurred
⇒ we know our outcome comes
from E_2 .

GIVEN E_2 OCCURRED, WHAT IS
THE PROBABILITY E_1 OCCURRED?

$P\{ \text{seeing } E_1 \mid \text{Given saw } E_2 \}$

= FRACTION OF THE AREA OF E_2 WHICH
IS OCCUPIED BY E_1

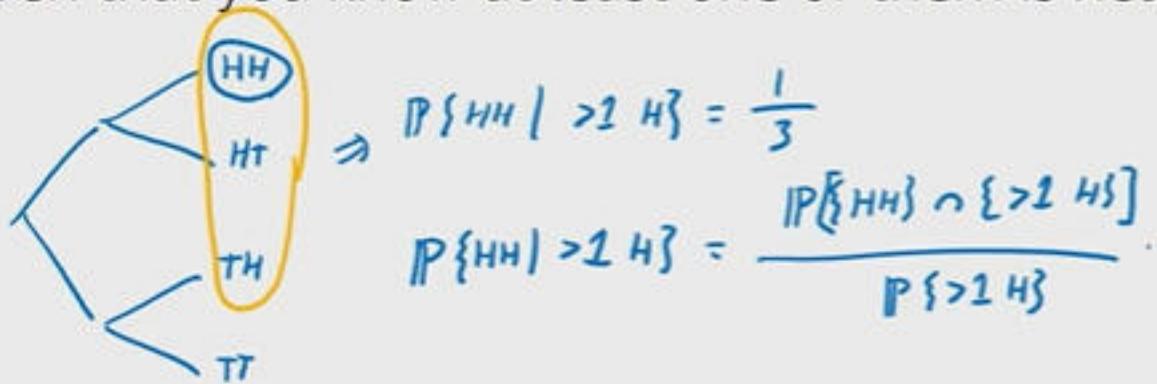
Area = AREA (O)

6:37



Example

Suppose you flip two coins. What is the probability that both are heads, given that you know at least one of them is heads?



Example

Suppose you flip two coins. What is the probability that both are heads, given that you know at least one of them is heads?

$$P\{HH \mid >1 H\} = \frac{1}{3}$$
$$P\{HH \mid >1 H\} = \frac{P\{\text{HH}\} \cap \{>1 H\}}{P\{>1 H\}} = \frac{P\{\text{HH}\}}{P\{>1 H\}}$$
$$= \frac{\frac{1}{4}}{\frac{3}{4}} = \frac{1}{3}$$



Bayes' Rule



Bayes' Rule can be leveraged to understand **competing hypotheses**.

$$H_1 \quad H_2$$

Suppose you have two hypotheses, H_1 and H_2 , and you have some prior belief about the odds that each is true.

If you observe some data, how do you understand the **new odds** of the two hypotheses under the new data?



Bayes' Rule



Bayes' Rule can be leveraged to understand **competing hypotheses**.

H_1 H_2

Suppose you have two hypotheses, H_1 and H_2 , and you have some prior belief about the odds that each is true.

$$\frac{P(H_1)}{P(H_2)}$$

If you observe some data, how do you understand the **new odds** of the two hypotheses under the new data?

Bayes' Rule

Bayes' Rule can be leveraged to understand **competing hypotheses**.

H_1 H_2

Suppose you have two hypotheses, H_1 and H_2 , and you have some prior belief about the odds that each is true.

$$\frac{P(H_1)}{P(H_2)} = \frac{2/3}{1/3} = \frac{2}{1}$$

If you observe some data, how do you understand the **new odds** of the two hypotheses under the new data?



Bayes' Rule

$$\frac{P(H_1 | D)}{P(H_2 | D)} = \frac{\frac{P(H_1 \wedge D)}{P(D)}}{\frac{P(H_2 \wedge D)}{P(D)}}$$

Bayes' Rule

$$\frac{P(H_1 | D)}{P(H_2 | D)} = \frac{\frac{P(H_1 \cap D)}{P(D)}}{\frac{P(H_2 \cap D)}{P(D)}} = \frac{P(H_1 \cap D)}{P(H_2 \cap D)}$$

$$= \frac{P(H_1 \cap D)}{P(H_2 \cap D)}.$$

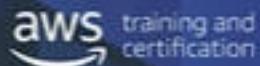
Bayes' Rule

$$\begin{aligned}\frac{P(H_1 | D)}{P(H_2 | D)} &= \frac{\frac{P(H_1 \cap D)}{P(D)}}{\frac{P(H_2 \cap D)}{P(D)}} = \frac{P(H_1 \cap D)}{P(H_2 \cap D)} \\&= \frac{P(H_1 \cap D)}{P(H_2 \cap D)} \cdot \frac{P(H_1)}{P(H_2)} \cdot \frac{P(H_2)}{P(H_1)} = \left[\frac{\frac{P(H_1 \cap D)}{P(H_1)}}{\frac{P(H_2 \cap D)}{P(H_2)}} \right] \cdot \frac{P(H_1)}{P(H_2)} \\&= \frac{P(D | H_1)}{P(D | H_2)} \cdot \frac{P(H_1)}{P(H_2)}\end{aligned}$$



$$\frac{P(H_1 \mid D)}{P(H_2 \mid D)} = \frac{P(D \mid H_1)}{P(D \mid H_2)} \cdot \frac{P(H_1)}{P(H_2)}$$

Bayes' Rule



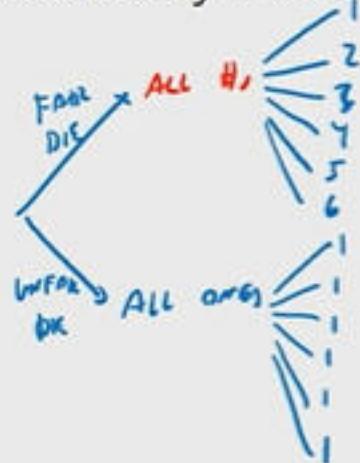
$$\frac{6}{1} = \frac{3}{1} + \frac{2}{1}$$

$$\frac{P(H_1 | D)}{P(H_2 | D)} = \underbrace{\frac{P(D | H_1)}{P(D | H_2)}}_{\text{RATIO OF PROBABILITY OF OBSERVING DATA.}} \cdot \underbrace{\frac{P(H_1)}{P(H_2)}}_{\text{PRIOR ODDS}}$$



Example

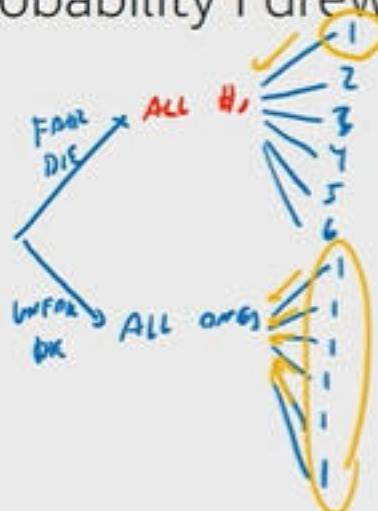
Suppose I have a bag with 2 dice. The red one is fair, the blue one is only ones. I draw a random die, roll it, and get a one. What is the probability I drew the fair die?



$$\begin{aligned} H_1 &= \text{BLUE DIE} & H_2 &= \text{RED DIE} \\ D &= \text{ROLLED} & A &= 1 \\ \frac{P(H_1|D)}{P(H_2|D)} &= \frac{P(D|H_1)}{P(D|H_2)} \cdot \frac{P(H_1)}{P(H_2)} \\ &= \frac{\frac{1}{6}}{\frac{1}{2}} \cdot \frac{\frac{1}{2}}{\frac{1}{2}} \\ &= \frac{6}{1} \cdot \frac{1}{1} \end{aligned}$$

Example

Suppose I have a bag with 2 dice. The red one is fair, the blue one is only ones. I draw a random die, roll it, and get a one. What is the probability I drew the fair die?

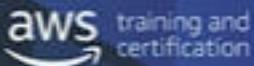


$$\begin{aligned} H_1 &= \text{BLUE DIE} & H_2 &= \text{RED DIE} \\ D &= \text{ROLLED} & A &= 1 \\ P(H_1|D) &= \frac{P(D|H_1)}{P(D|H_2)} \cdot \frac{P(H_1)}{P(H_2)} & P(H_1|D) &= \frac{6}{7} \\ &= \frac{1}{6} \cdot \frac{1}{\frac{1}{2}} & P(H_2|D) &= \frac{1}{7} \\ \frac{6}{1} &= \frac{6}{1} \cdot \frac{1}{1} \end{aligned}$$

$E_1 \ \& \ E_2$ ← "No" REGULATOR"

$$\left. \begin{array}{l} P\{E_1 | E_2\} = P\{E_1\} \\ \frac{P\{E_1 \cap E_2\}}{P\{E_2\}} \end{array} \right\} =$$

Intuition



$E_1 \text{ & } E_2$ ← "no overlap"

$$\left. \begin{aligned} P\{E_1 | E_2\} &= P\{E_1\} \\ \frac{P\{E_1 \cap E_2\}}{P\{E_2\}} & \end{aligned} \right\} \Rightarrow P\{E_1 \cap E_2\} = P\{E_1\} \cdot P\{E_2\}$$



Definition

Two events are **independent** if one event doesn't influence the other.

$$E_1, E_2 \text{ are independent if } P\{E_1 \cap E_2\} = P\{E_1\} \cdot P\{E_2\}$$

$$\overline{E_1, E_2, \dots, E_k \text{ are independent if } P\{\bigcap \text{ ANY SET OF THE } E_i\} = \prod P\{\text{THE SET}\}}$$
$$P\{E_1 \cap E_2 \cap E_3\} = P\{E_1\} \cdot P\{E_2\} \cdot P\{E_3\}$$

Definition

Two events are **independent** if one event doesn't influence the other.

$$E_1, E_2 \text{ ARE INDEPENDENT IF } P\{E_1 \cap E_2\} = P\{E_1\} \cdot P\{E_2\}$$

$$\begin{aligned} E_1, E_2, \dots, E_k \text{ ARE INDEPENDENT IF } & P\{\bigcap \text{ ANY SET OF THE } E_i\} \\ & = \prod P\{\text{THE SET}\} \\ P\{E_1 \cap E_2 \cap E_3\} & = P\{E_1\} \cdot P\{E_2\} \cdot P\{E_3\} \end{aligned}$$

Definition

Two events are **independent** if one event doesn't influence the other.

$$E_1, E_2 \text{ are independent if } P\{E_1 \cap E_2\} = P\{E_1\} \cdot P\{E_2\}$$

$$\overline{E_1, E_2, \dots, E_k \text{ are independent if } P\{\bigcap \text{ ANY SET OF THE } E_i\} = \prod P\{\text{THE SET}\}}$$
$$P\{E_1 \cap E_2 \cap E_3\} = P\{E_1\} \cdot P\{E_2\} \cdot P\{E_3\}$$

$$P\{E_1 \cap E_2 \cap \dots \cap E_k\} = P\{E_1\} \cdot P\{E_2\} \cdot \dots \cdot P\{E_k\}$$

Definition



Two events are **independent** if one event doesn't influence the other.

$$E_1, E_2 \text{ are independent if } P\{E_1 \cap E_2\} = P\{E_1\} \cdot P\{E_2\}$$

E_1, E_2, \dots, E_k are independent if

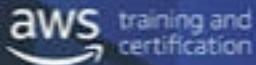
$$\begin{aligned} &P\{\bigcap \text{ ANY SET OF THE } E_i\} \\ &= \prod P\{\text{THE SET}\} \end{aligned}$$

$$P\{E_1 \cap E_2 \cap \dots \cap E_k\} = P\{E_1\} \cdot P\{E_2\} \cdot \dots \cdot P\{E_k\}$$

$$P\{E_1 \cap E_2 \cap \dots \cap E_k\} = P\{E_1\} \cdot P\{E_2\} \cdot \dots \cdot P\{E_k\}$$



Definition



Two events are **independent** if one event doesn't influence the other.

$$E_1, E_2 \text{ are independent if } P\{E_1 \cap E_2\} = P\{E_1\} \cdot P\{E_2\}$$

E_1, E_2, \dots, E_k are independent if

$$\begin{aligned} &P\{\bigcap \text{ ANY SET OF THE } E_i\} \\ &= \prod P\{\text{THE SET}\} \end{aligned}$$

$$e.g. P\{E_1 \cap E_2 \cap E_3\} = P\{E_1\} \cdot P\{E_2\} \cdot P\{E_3\}$$

$$P\{E_1 \cap E_2 \cap \dots \cap E_k\} = P\{E_1\} \cdot P\{E_2\} \cdot \dots \cdot P\{E_k\}$$



Topics



- Definitions
 - Discrete Random variable
 - Expected value
 - Variance
 - Standard deviation
- Chebyshev's Inequality



Topics



- Definitions
 - Discrete Random variable
 - Expected value
 - Variance
 - Standard deviation
- Chebyshev's Inequality



Discrete Random Variable



DFA

DID I GET 3 HEADS? → How many heads?

$\Omega \ni \omega$
outcome HHHHTTHH



17:40



Discrete Random Variable



DFA

DID I GET 3 HEADS? → How many heads?

$\Omega \ni \omega$
outcome HHHHTTHH



17:40



Discrete Random Variable



DFA

DID I GET 3 HEADS? → How many heads?

$\Omega \ni \omega$
outcome $HHHHTTHH$ ↘ real #

$H: \omega \in \Omega \rightarrow \mathbb{R}$

$$H(HHHHTTHH) = 4$$

Discrete R.V.
A function
A outcome
 X takes
at ω gives x
THAT A number back



-16:19



Expected Value

$$\begin{aligned} X &\leftarrow \text{R.V.} \\ &= \text{MEAN} \\ \mu_x = \mathbb{E}[X] &= \sum_x x \cdot P\{X=x\} \end{aligned}$$



-14:16



Expected Value

$$X \leftarrow R.V.$$
$$\mu_x = E[X] = \sum_x x \cdot P\{X=x\} = \sum_{\omega} X(\omega) \cdot P\{\omega\}$$

e.g. FLIPPING 4 coins AND getting the number of heads H is the R.V. which $H \in \{0, 1, 2, 3, 4\}$

$$E[H] = \sum_x x \cdot P\{H=x\} = 1 \cdot P\{H=1\} + 2 \cdot P\{H=2\} + 3 \cdot P\{H=3\} + 4 \cdot P\{H=4\}$$
$$= 1 \cdot \frac{1}{16}$$



-12:21



Expected Value



$$X \leftarrow \text{R.V.} \\ = \text{MEAN}$$
$$\mu = \mathbb{E}[X] = \sum_x x \cdot P\{X=x\} = \sum_{\omega} X(\omega) \cdot P\{\omega\}$$

e.g. Flipping four coins and let H be the R.V. which gives the number of heads. $H \in \{0, 1, 2, 3, 4\}$

$$\mathbb{E}[H] = \sum_x x \cdot P\{H=x\} = 1 \cdot P\{H=1\} + 2 \cdot P\{H=2\} + 3 \cdot P\{H=3\} + 4 \cdot P\{H=4\}$$
$$= 1 \cdot \frac{4}{16} + 2 \cdot \frac{6}{16} + 3 \cdot \frac{4}{16} + 4 \cdot \frac{1}{16} = 2$$



Variance



$$\begin{aligned}\sigma_x^2 &= \text{Var}(X) = E[(X - \mu_x)^2] \\ &= E[X^2 - 2X\mu_x + \mu_x^2] \\ &= E[X^2] - 2\mu_x E[X] + \mu_x^2 \\ &= E[X^2] - \mu_x^2\end{aligned}$$

FLAV

$X \leftarrow \text{income}$
 $X - \mu_x \leftarrow \text{inverty}$
 $(X - \mu_x)^2 \leftarrow \text{square income}$
 $\Rightarrow \text{Var}(X) \leftarrow \text{square inverty}$



Standard Deviation

$$\sigma_x = sd(\bar{X}) = \sqrt{Var(\bar{X})}$$

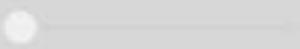


422



Standard Deviation

$$\sigma_x = sd(\bar{X}) = \sqrt{Var(\bar{X})}$$



422



Standard Deviation

$$\sigma_x = \text{sd}(\bar{X}) = \sqrt{\text{Var}(\bar{X})}$$

σ_x is in the same units as X
and thus can be interpreted as a measure



-3:40



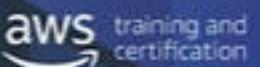
Chebyshev's Inequality



-2:31



Chebyshev's Inequality



For *any* random variable X (no assumptions) at least 99% of the time

$$X \in [\mu_X - 10\sigma_X, \mu_X + 10\sigma_X]$$



Chebyshev's Inequality

For *any* random variable X (no assumptions) at least 99% of the time

$$X \in [\mu_X - 10\sigma_X, \mu_X + 10\sigma_X]$$



Topics

- Intuition
- Three Examples
 - One coin
 - Two coins
 - A mixed case
- Examine the Trees
- Definition
- The Only Choice was the Units



Intuition



We want to be able to quantify the amount of randomness in a distribution: A six-sided die should be more random than a coin. Entropy will turn out to be the unique answer to this question.

Declare one coin flip = 1 bit of randomness



Three Examples

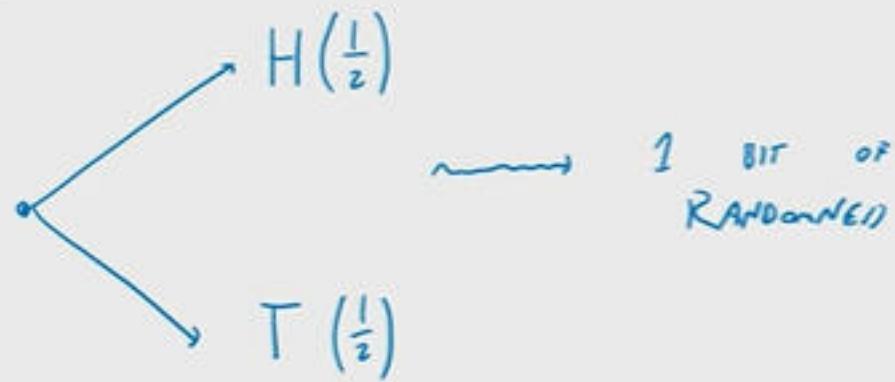


One coin

Three Examples



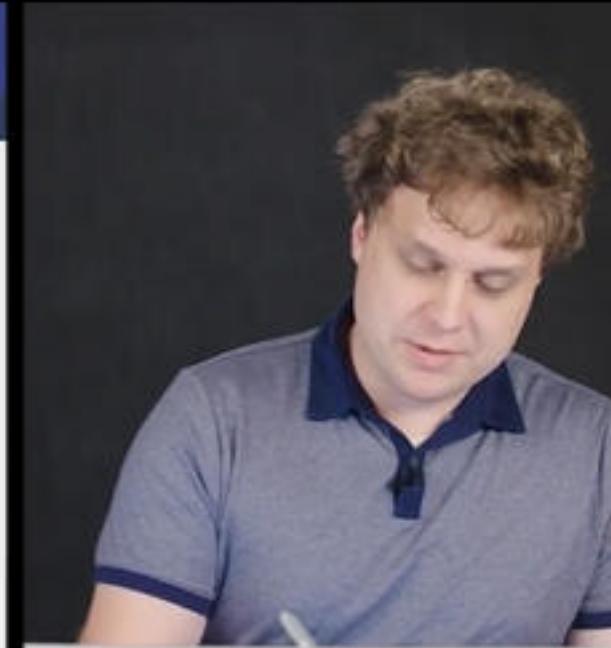
One coin



Three Examples

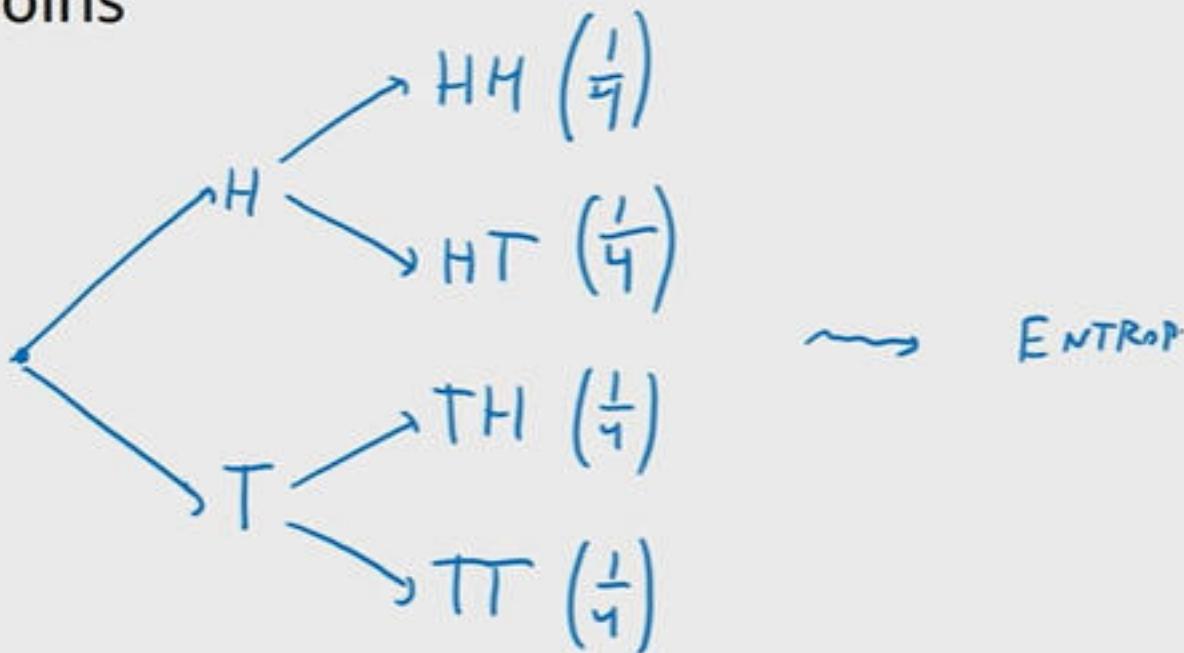


Two coins

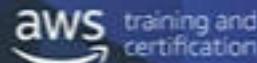


Three Examples

Two coins



Definition



The entropy of a probability distribution is:

$$\underline{H} = - \sum p_i \log_2(p_i)$$



Definition

The entropy of a probability distribution is:

Color |

$$\underline{H} = - \sum p_i \log_2(p_i)$$

3 outcomes
 $\left(\begin{array}{c} \frac{1}{2} \\ \frac{1}{2} \\ \frac{1}{2} \\ \frac{1}{2} \end{array} \right)$

$$H = - \sum^3 \frac{1}{2} \log_2 \left(\frac{1}{2} \right)$$

3 terms
= $- \log_2 \left(\frac{1}{2} \right) = 1$ bit (3)

3:28



The Only Choice was the Units



1 Unit of Randomness \iff 1 Roll

The Only Choice was the Units

1 unit of Randomness \iff 1 Role or a 10 sided die.



$$\frac{1}{10^k} \text{ Row}$$

$$\# \text{Rows} = -\log_{10}(p)$$

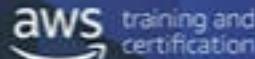
$$H = - \sum_i p_i \log_{10}(p_i)$$



-1:17



The Only Choice was the Units



1 unit of randomness \iff 1 bit or a 10 sized dic.

$$\left\{ \text{to learn} \right. \quad \frac{1}{10^k} \text{ raw}$$

$$H = - \sum_i p_i \log_{10}(p_i)$$

$$= \frac{1}{\log_2(10)} H' \quad \text{"bit"}$$

$$\# \text{Recs} = -\log_{10}(p)$$

$$\log_{10}(x) = \frac{\log_2(x)}{\log_2(10)}$$



Definitions

aws training and certification

Probability density function



© 2016, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

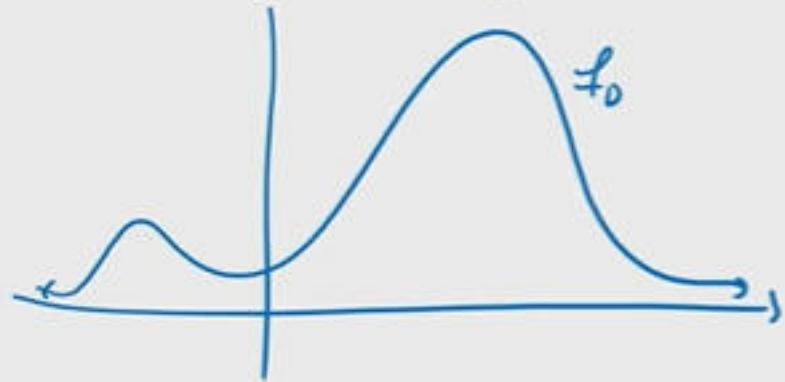


-3:10



Probability density function

f_D

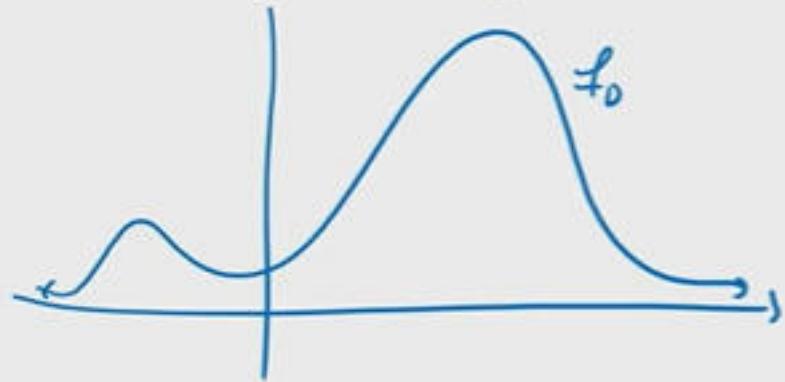


234



Probability density function

f_D



234



Definitions

Probability density function



$$f_D$$
$$\mathbb{P}\{D \in [a, b]\} = \text{AREA}(\textcircled{O})$$

$$\mathbb{P}\{D \in [x, x + 0.00001]\} = f_D(x) \cdot 0.00001$$



Topics



- Standard Gaussian (Normal Distribution) Density
- General Gaussian Density
- Key Properties
 - Maximum entropy distribution
 - Central limit theorem



© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.



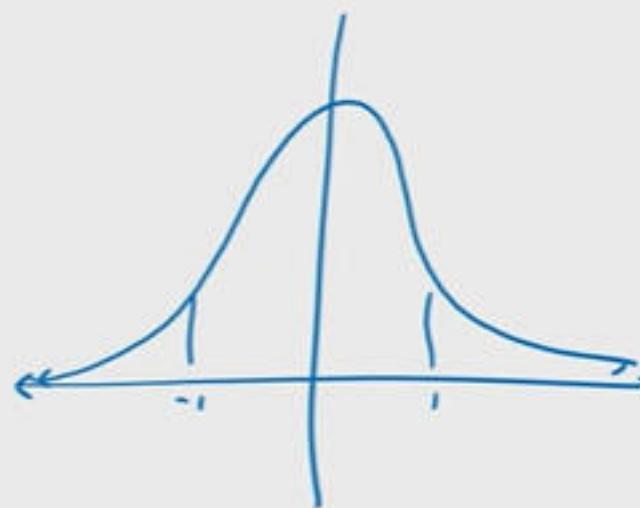
7:11



Standard Gaussian (Normal Distribution) Density



$$f_x(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

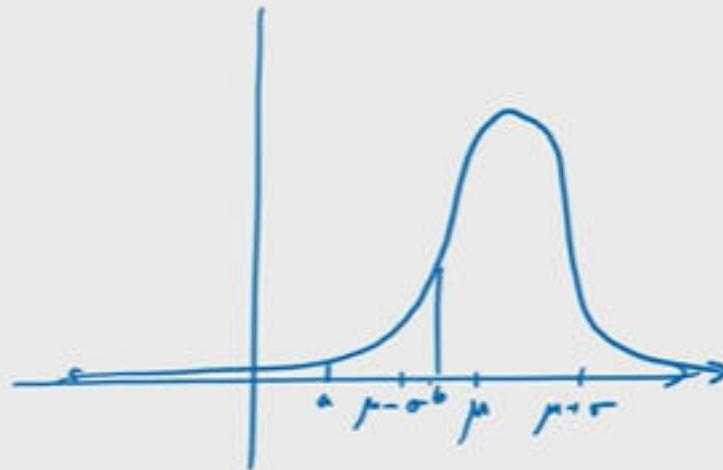


6:27



General Gaussian Density

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

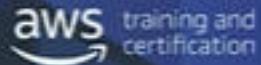


General Gaussian Density

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

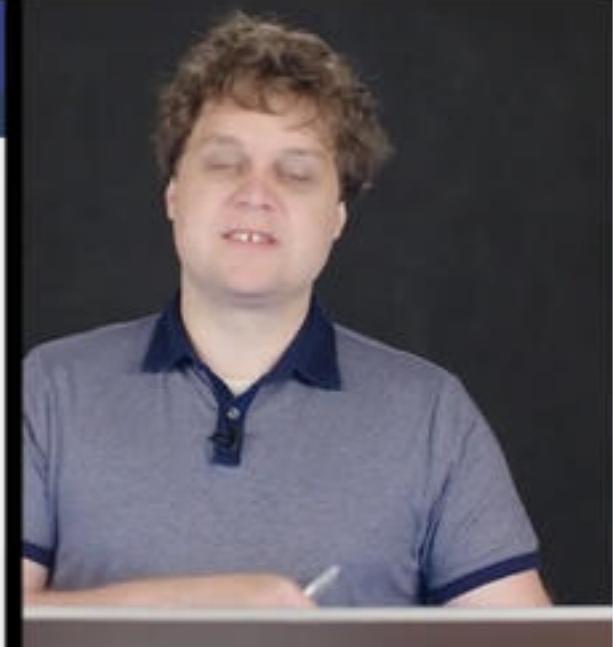


Key Properties



Maximum Entropy Distribution

A random acc continuous RV WITH $E[X] = \mu$
 $V_{\text{AR}}[X] = 1$ $H(X)$ is maximized UNIQUELY
For $X \sim N(\mu, \sigma^2)$

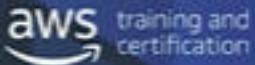


Maximum Entropy Distribution

Amongst all continuous RV with $E[X] = \mu$
 $Var[X] = 1$ $H(X)$ is maximized uniquely
For $X \sim N(\mu, 1)$

Gaussian is the most random RV
variable with fixed mean (μ)

Key Properties

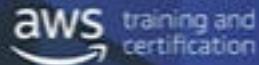


Central Limit Theorem

X_i if a sequence of independent R.V.s



Key Properties



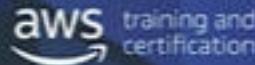
Central Limit Theorem

X_i if a sequence of independent R.V.s.
 $E[X_i] = \mu$ $\text{Var}[X_i] = \sigma^2$

$$X_1 + X_2 + X_3 + \dots + X_n \sim N(\mu, n\sigma^2)$$



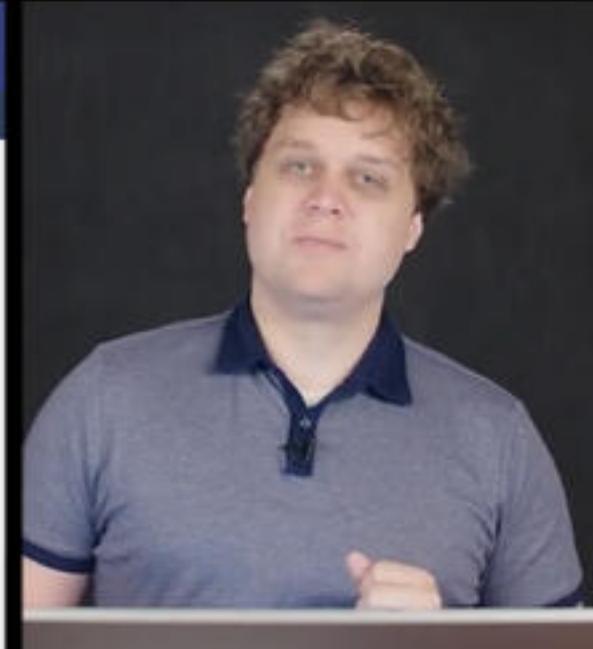
Key Properties



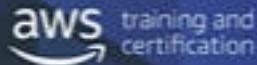
Central Limit Theorem

X_i if a sequence of independent R.V.s.
 $E[X_i] = \mu$ $\text{Var}[X_i] = \sigma^2$

$$X_1 + X_2 + X_3 + \dots + X_n \sim N(\mu, n\sigma^2)$$



Maximum Likelihood Estimation



When building a model, it is often done by describing a process to produce our data. This will almost always incorporate random variables to account for un-modeled variables. You will then be able to evaluate the parameters of the models by how likely they were to produce the given data.



Maximum Likelihood Estimation



When building a model, it is often done by describing a process to produce our data. This will almost always incorporate random variables to account for un-modeled variables. You will then be able to evaluate the parameters of the models by how likely they were to produce the given data.



Maximum Likelihood Estimation

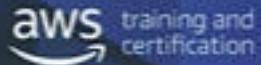


When building a model, it is often done by describing a process to produce our data. This will almost always incorporate random variables to account for un-modeled variables. You will then be able to evaluate the parameters of the models by how likely they were to produce the given data.

$$\{\tilde{x}_i\}_{i=1}^n, \quad \{y_i\}_{i=1}^n$$

$$y \sim \omega \cdot \tilde{x} + N(0, 1)$$

Maximum Likelihood Estimation



Given a probability model with some vector of parameters $\vec{\theta}$, and observed data D , the best fitting model is the one that maximizes

$$P_{\vec{\theta}}(D).$$

The parameters that maximize this probability are called the *maximum likelihood estimator*.



Maximum Likelihood Estimation



Given a probability model with some vector of parameters $\vec{\theta}$, and observed data D , the best fitting model is the one that maximizes

$$P_{\vec{\theta}}(D).$$

FIND $\hat{\theta}$ S

The parameters that maximize this probability are called the *maximum likelihood estimator*.

Learning Objectives



- See how:

probability theory + optimization = learning



Derivatives



- Slope and Optimization
- Definition
- Common Derivatives
- Derivatives Rules
- Back to the Example
- Interpreting Derivatives



© 2016, Amazon Web Services, Inc. or its Affiliates. All rights reserved.



-2:11



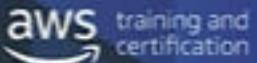
Numerical Methods



- A Quick Note on Negative Log-Likelihood
- Intuition: Gradient Descent



Gradient Descent



- Definition
- Example
- This is Actively Used
- Issue



Newton's Method



- Issue
- Idea
- Pictorially
- Computing the Line

Newton's Method

- Issue
- Idea
- Pictorially
- Computing the Line
- Update Step for Zero Finding
- Update Step for Minimization
- Relationship to Gradient Descent

Summary

- Summary





Thank You

© 2018 Amazon Web Services, Inc. or its affiliates. All rights reserved. This work may not be reproduced or redistributed, in whole or in part, without prior written permission from Amazon Web Services, Inc. Commercial copying, lending, or selling is prohibited. Corrections or feedback on the course, please email us at gcp-course-feedback@amazon.com. For all other questions, contact us at <http://aws.amazon.com/contact-us-training/>. All trademarks are the property of their owners.



Motivating Example

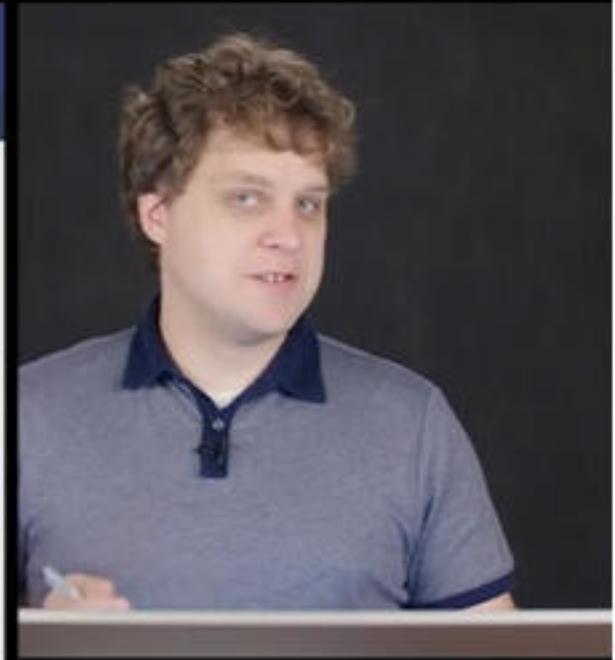


As an example, here is a simple toy learning task:

A coin comes up heads with probability p and tails with probability $1 - p$, but you do not know p . You observe a sequence of flips to be:

HHHTTHTTHHTHTT

How can you use this data to learn p ?



Intuition



You can probably guess the answer, but let's generalize this to make it applicable to more complex questions. (Think about trying to model what cat pictures look like, and looking at 1,000 photos of cats) Let's dive into why the answer will be:

$$P = \frac{\# \text{ Heads}}{\# \text{ Flips}} = \frac{6}{13}$$



6 HEADS

7 TAILS

$$P_p(D) = p^6 (1-p)^7$$

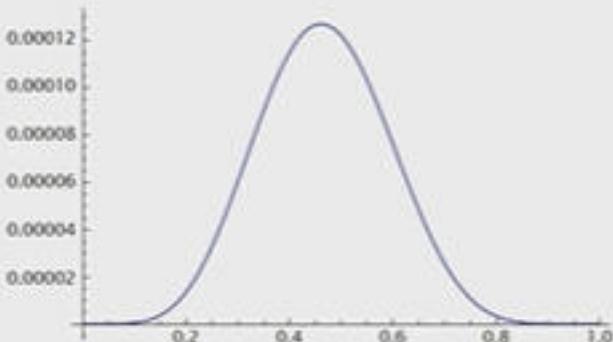


-3:54



Example

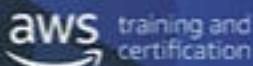
$$P_p(H H H T T H T T H T H T T) = P^6(1)$$



-3:02



Topics



- Definition
- Common Derivatives
- Derivatives Rules
- Back to the Example
- Interpreting Derivatives
- The Example Completed



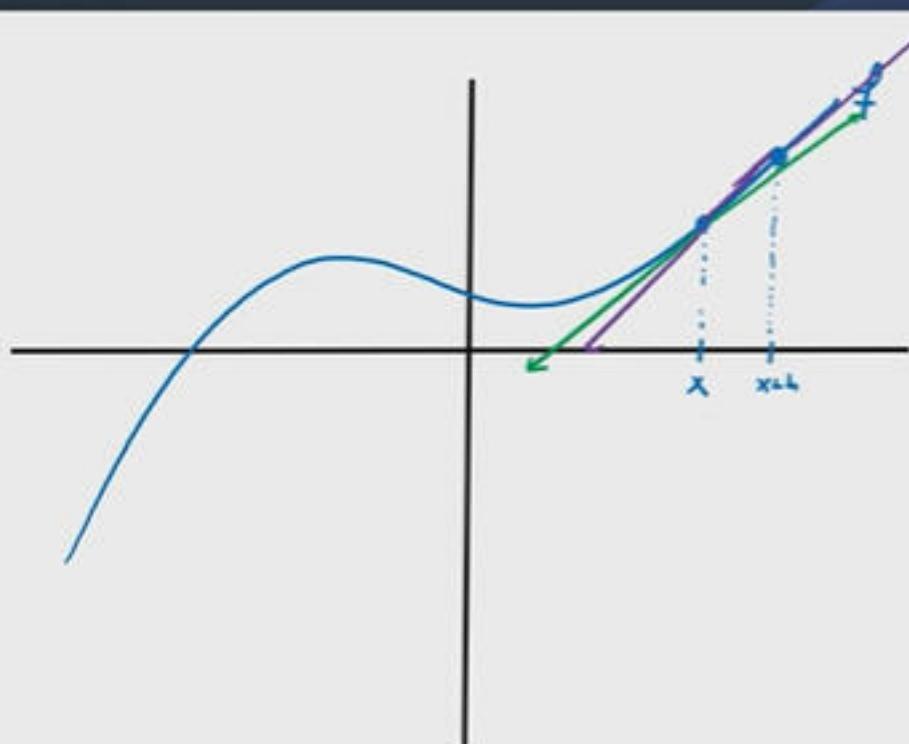
© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.



-24:02

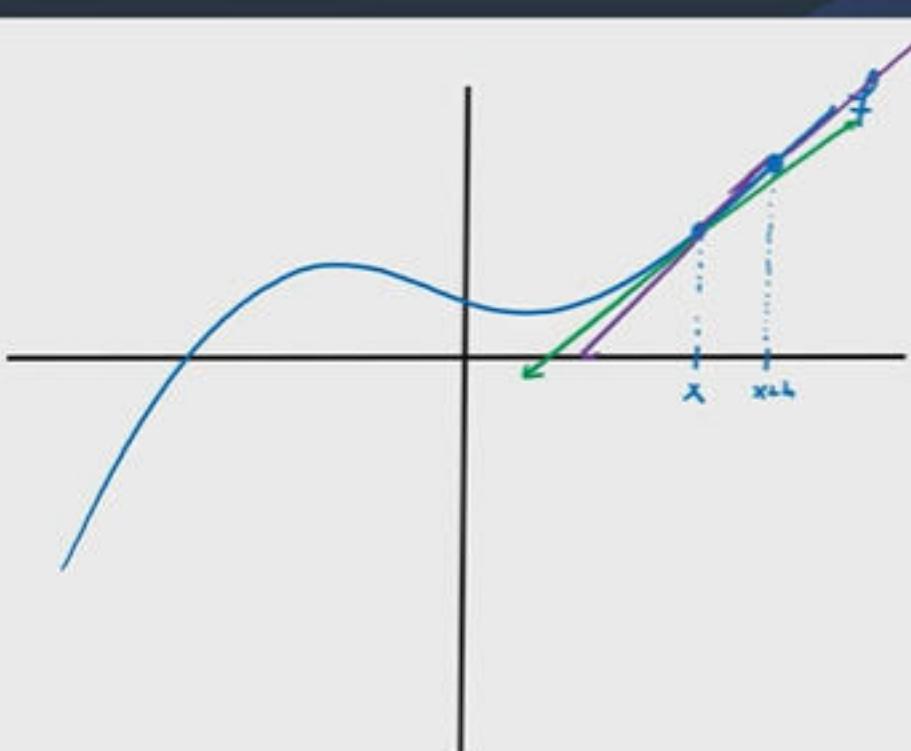


Definition



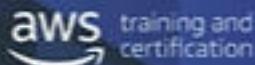
$$\begin{aligned} \text{Approximate} &= \frac{f(x+h) - f(x)}{(x+h) - x} \\ &= \frac{f(x+h) - f(x)}{h} \end{aligned}$$

Definition



$$\begin{aligned} \text{Approximation} &= \frac{f(x+h) - f(x)}{(x+h) - x} \\ &= \frac{f(x+h) - f(x)}{h} \\ \text{GET BETTER APPROX.} & \quad h \rightarrow 0 \\ f'(x) &= \lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h} \end{aligned}$$

Common Derivatives



Most ML tasks thankfully need comparatively few derivatives.

| | |
|-------------------|--|
| f | $f' = \frac{df}{dx} = \frac{d}{dx} f = \partial_x f = f_x$ |
| $(n \neq 0) x^n$ | nx^{n-1} |
| e^x | e^x |
| $(x > 0) \log(x)$ | $\frac{1}{x}$ |



Note: There are many notations you may encounter for derivatives.
The list above is not exhaustive.

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.



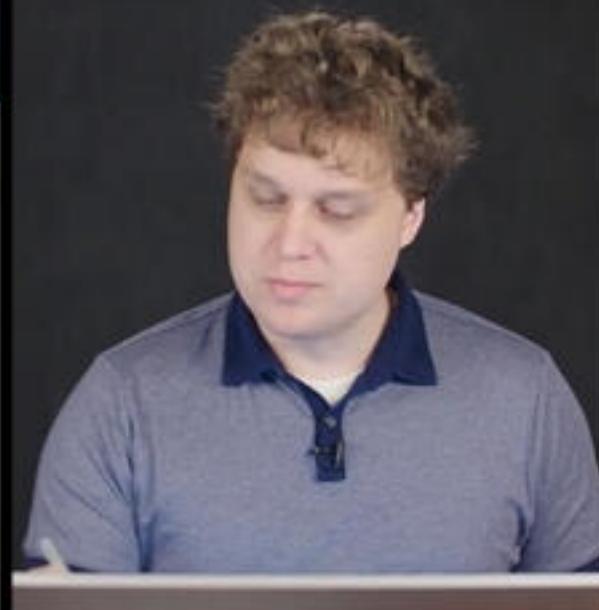
-19:06



Derivative Rules



$$[f(x) + g(x)]' = f'(x) + g'(x)$$



Derivative Rules



$$[f(x) + g(x)]' = f'(x) + g'(x) \quad \leftarrow \text{sum}$$

$$\underline{f(x+\epsilon) + g(x+\epsilon)} = \underline{f(x) + g(x)}$$

Derivative Rules



$$[f(x) + g(x)]' = f'(x) + g'(x) \quad \leftarrow \text{sum}$$

$$f(x+\epsilon) + g(x+\epsilon) = f(x) + g(x) + f'(x)\epsilon + g'(x)\epsilon$$

Derivative Rules



$$[f(x) + g(x)]' = f'(x) + g'(x) \quad \leftarrow \text{sum}$$

$$f(x+\epsilon) + g(x+\epsilon) = f(x) + g(x) + f'(x)\epsilon + g'(x)\epsilon$$

Derivative Rules

$$[f(x) + g(x)]' = f'(x) + g'(x) \quad \leftarrow \text{sum}$$

$$\begin{aligned}f(x+\epsilon) + g(x+\epsilon) &= f(x) + g(x) + f'(x)\epsilon + g'(x)\epsilon \\&= (f(x) + g(x)) + \underbrace{(f'(x) + g'(x))\epsilon}_{1} \\&= [f(x) + g(x)]'\end{aligned}$$

$$\frac{d}{dx}$$

Derivative Rules



$$[f(x) + g(x)]' = f'(x) + g'(x)$$

← sum

$$\begin{aligned}f(x+\epsilon) + g(x+\epsilon) &= f(x) + g(x) + f'(x)\epsilon + g'(x)\epsilon \\&= (f(x) + g(x)) + \underbrace{(f'(x) + g'(x))\epsilon}_{1} \\&= [f(x) + g(x)]'\end{aligned}$$

$$\frac{d}{dx}[x^2 + e^x] = 2x + e^x$$



Derivative Rules



$$[f(x) \cdot g(x)]' = f(x)g'(x) + f'(x)g(x)$$

← power rule



Derivative Rules



$$[f(x) \cdot g(x)]' = f(x)g'(x) + f'(x)g(x) \quad \leftarrow \text{product rule}$$

$$\begin{aligned}f(x+\epsilon)g(x+\epsilon) &= [f(x) + f'(\omega)\epsilon] \cdot [g(x) + g'(\omega)\epsilon] \\&= f(x)g(x) + [f(x)g'(\omega) + g(x)f'(\omega)]\epsilon + f'(\omega)g'(\omega)\epsilon^2\end{aligned}$$



Derivative Rules

$$[f(x) \cdot g(x)]' = \underline{f(x)g'(x) + f'(x)g(x)} \quad \leftarrow \text{product rule}$$

$$\begin{aligned} f(x+\epsilon) \cdot g(x+\epsilon) &= [f(x) + f'(x)\epsilon] \cdot [g(x) + g'(x)\epsilon] \\ &= \underbrace{f(x)g(x)}_{f(x) \cdot g(x)} + \underbrace{[f'(x)g(x) + g'(x)f(x)]\epsilon}_{[f(x) \cdot g(x)]'} + \cancel{f'(x)g'(x)\epsilon^2} \end{aligned}$$

$x^2 e$

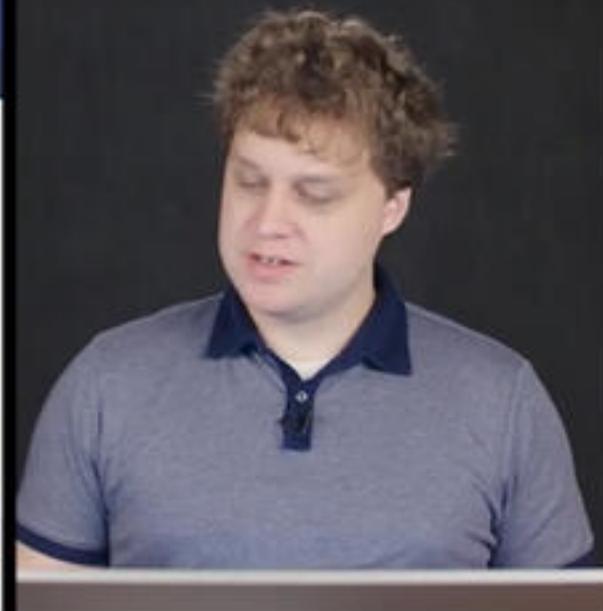
Derivative Rules



$$[f(x) \cdot g(x)]' = \underline{f(x)g'(x)} + f'(x)g(x) \quad \leftarrow \text{product rule}$$

$$\begin{aligned} f(x+\varepsilon)g(x+\varepsilon) &= [f(x) + f'(x)\varepsilon] \cdot [g(x) + g'(x)\varepsilon] \\ &= \underbrace{f(x)g(x)}_{f(x)g(x)} + \underbrace{[f'(x)g(x) + g(x)f'(x)]\varepsilon}_{[f(x)g(x)]'} + \underbrace{f'(x)g'(x)\varepsilon^2}_0 \end{aligned}$$

$$\frac{d}{dx}[x^2 e^x] = 2x e^x + x^2 e^x$$



Derivative Rules



$$[f(g(x))]' = f'(g(x)) g'(x) \quad \leftarrow \text{CHAIN RULE}$$



© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.



-12:48



Derivative Rules



$$[f(g(x))]' = f'(g(x)) g'(x) \quad \leftarrow \text{CHAIN RULE}$$

$$f(g(x+\epsilon)) = f(g(x) + g'(x)\epsilon)$$

Derivative Rules



$$[f(g(x))]' = f'(g(x)) g'(x) \quad \leftarrow \text{CHAIN RULE}$$

$$f(g(x+\epsilon)) = f(g(x) + g'(x)\epsilon)$$

Derivative Rules



$$[f(g(x))]' = f'(g(x)) g'(x) \quad \leftarrow \text{CHAIN RULE}$$

$$\begin{aligned} f(g(x+\epsilon)) &= f\left(g(x) + \underbrace{g'(x)\epsilon}_{\text{error term}}\right) \\ &= \underbrace{f(g(x))}_{\text{original function}} + f'(g(x))g'(x)\epsilon \end{aligned}$$



Derivative Rules



$$[f(g(x))]' = f'(g(x)) g'(x) \quad \leftarrow \text{CHAIN RULE}$$

$$\begin{aligned} f(g(x+\epsilon)) &= f\left(g(x) + \underbrace{g'(x)\epsilon}_{\text{error term}}\right) \\ &= \underbrace{f(g(x))}_{\text{original function}} + \underbrace{f'(g(x))g'(x)\epsilon}_{\text{derivative of the composition}} \\ &\quad [f(g(x))]' \end{aligned}$$

$$\frac{d}{dx} [e^{x^2}] = e^{x^2} 2x$$

Derivative Rules

$$[f(g(x))]' = f'(g(x)) g'(x) \quad \leftarrow \text{CHAIN RULE}$$

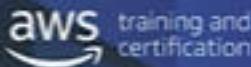
$f(g(x))$

$$\frac{df}{dx} = \frac{df}{dg} \cdot \frac{dg}{dx}$$

$$\begin{aligned} f(g(x+\epsilon)) &= f\left(g(x) + \underbrace{g'(x)\epsilon}_{\text{error term}}\right) \\ &= \underbrace{f(g(x))}_{\text{outer function}} + \underbrace{f'(g(x))g'(x)\epsilon}_{\text{inner function derivative}} \\ &\quad [f(g(x))]' \end{aligned}$$

$$\frac{d}{dx} [e^{x^2}] = e^{x^2} 2x$$

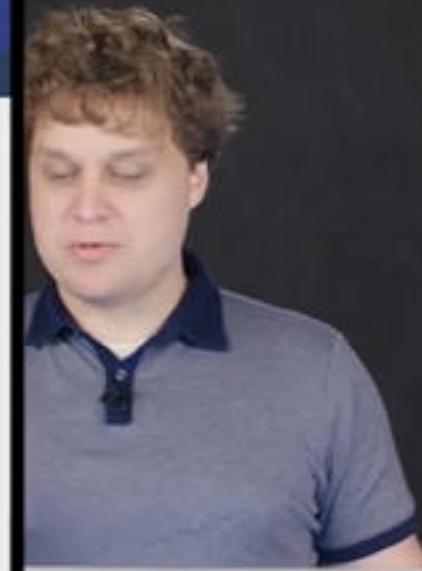
Derivative Rules



$$[f(g(x))]' = f'(g(x)) g'(x) \quad \leftarrow \text{CHAIN RULE}$$

$$\begin{aligned} f(g(x+\epsilon)) &= f(g(x) + \underbrace{g'(x)\epsilon)} \\ &= \underbrace{f(g(x))}_{f(g(x))} + \underbrace{f'(g(x))g'(x)\epsilon}_{[f(g(x))]'}. \end{aligned}$$

$$\boxed{\frac{df}{dx} = \frac{df}{dg} \cdot \frac{dg}{dx}}$$



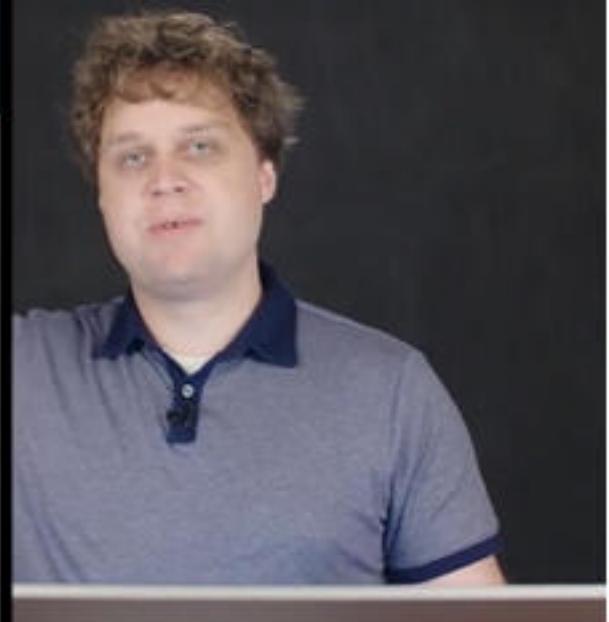
$$\frac{d}{dx} [e^{x^2}] = e^{x^2} 2x$$

Back to the Example



1) Cov PWIN 6 HSAR

$$f(p) = P_p(0) = p^6(1-p)^6$$
$$f'(p) = [p^6(1-p)^6]' = [p^6]'(1-p)^6 + p^6[(1-p)^6]' \quad (\text{product rule})$$
$$= 6p^5(1-p)^6 + p^6$$



Back to the Example

13 Core FLIPS

6 HEADS

$$f(p) = P_p(0) = p^6(1-p)^7$$

$$\begin{aligned} f'(p) &= [p^6(1-p)^7]' = [p^6]'(1-p)^7 + p^6[(1-p)^7]' \quad (\text{product rule}) \\ &= 6p^5(1-p)^7 + 7p^6(1-p)^6 \end{aligned}$$

Back to the Example

13 Core FLIPS 6 HEADS

$$f(p) = P_p(6) = p^6(1-p)^6$$
$$f'(p) = [p^6(1-p)^6]' = [p^6]'(1-p)^6 + p^6[(1-p)^6]' \quad (\text{PRODUCT RULE})$$
$$= 6p^5(1-p)^6 + 7p^6(1-p)^5 \quad (\text{POWER RULE, CHAIN RULE})$$
$$= 6p^5(1-p)^6 + 7p^6(1-p)^5$$

Interpreting Derivatives



$f' > 0 \longrightarrow$



$f' < 0 \longrightarrow$



$f' = 0 \longrightarrow$



© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.



-4:03



Our Example Completed

$$0 = f'(p) = p^5(1-p)^4 [6 - 13p]$$

Find p s.t. $f'(p) = 0$

\downarrow

$$p^5 = 0 \quad (1-p)^4 = 0 \quad 6 - 13p = 0$$

$\Leftrightarrow p = 0$
 $\Leftrightarrow f(0)$

Our Example Completed

$$0 = f'(p) = p^5(1-p)^4 [6 - 13p]$$

Find p s.t. $f'(p) = 0$

$$\begin{array}{lll} p^5 = 0 & (1-p)^4 = 0 & 6 - 13p = 0 \\ \Leftrightarrow p = 0 & \Leftrightarrow 1-p = 0 & \Leftrightarrow 6 = 13p \\ \Leftrightarrow f(p) = 0 & \Leftrightarrow p = 1 & \Leftrightarrow p = \frac{6}{13} \\ \text{min} & \text{min} & \Leftrightarrow f(p) > 0 \end{array}$$



-0:46



Our Example Completed



$$0 = f'(p) = p^5(1-p)^4 [6 - 13p]$$

$f' \leftarrow p \text{ is } \text{LT.} \quad f'(p) > 0$

$$\begin{array}{lll} p^5 = 0 & (1-p)^4 > 0 & 6 - 13p = 0 \\ \Leftrightarrow p = 0 & \Leftrightarrow 1-p > 0 & \Leftrightarrow 6 = 13p \\ \Leftrightarrow f(p) = 0 & \Leftrightarrow p = 1 & \Leftrightarrow p = \frac{6}{13} \\ \text{min} & \text{max} & \text{max} \end{array}$$



Our Example Completed



$$0 = f'(p) = p^5(1-p)^4 [6 - 13p]$$

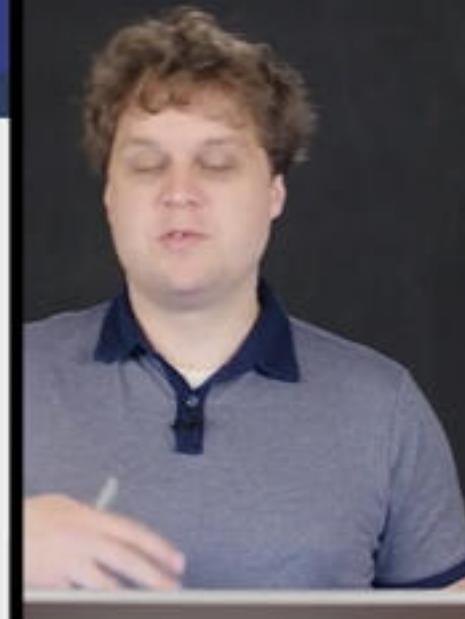
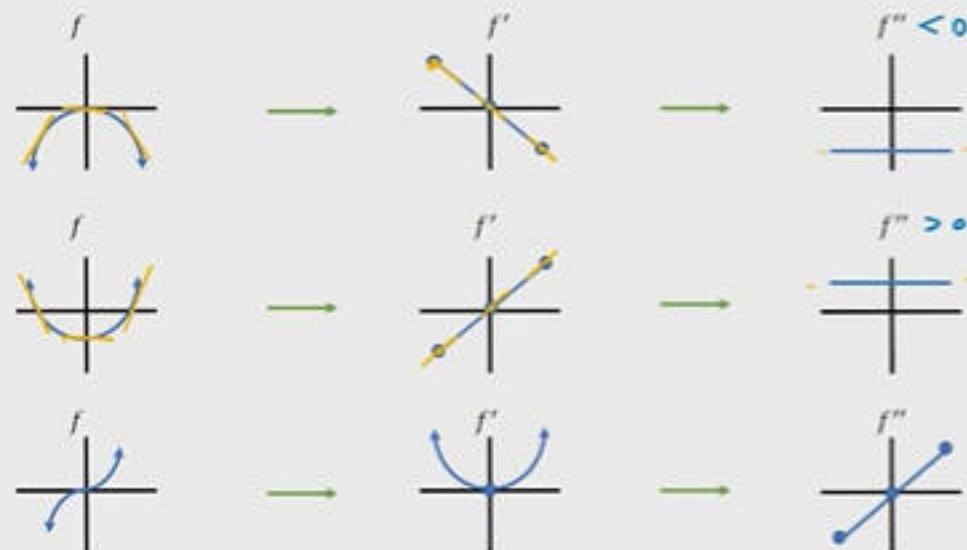
For $p \in [0, 1]$, $f'(p) = 0$

$$\begin{aligned} p^5 = 0 &\Leftrightarrow p = 0 \\ \Leftrightarrow f(p) = 0 &\Leftrightarrow f(p) = 0 \\ \text{min} & \end{aligned}$$
$$\begin{aligned} (1-p)^4 > 0 &\Leftrightarrow 1-p > 0 \\ \Leftrightarrow p < 1 &\Leftrightarrow f(p) > 0 \\ \text{min} & \end{aligned}$$
$$\begin{aligned} 6 - 13p = 0 &\Leftrightarrow 6 = 13p \\ \Leftrightarrow p = \frac{6}{13} &\Leftrightarrow f(p) > 0 \\ \text{max} & \end{aligned}$$

$$\boxed{\hat{p} = \frac{6}{13}}$$



Second Derivatives



How to Identify Extrema



f ← Max OR Min

$f' \neq 0$ → NEITHER A MAX OR MIN



-1:37



How to Identify Extrema



$f \leftarrow$ Max OR Min

$f' \neq 0 \rightarrow$ NEITHER A MAX OR MIN

$f' = 0 \rightarrow$ can HAVE A MAX OR MIN

$f'' > 0$

$f'' < 0$

f

Summary



$f' > 0$ f is locally increasing

$f' < 0$ f is locally decreasing

$f' = 0$ $\begin{cases} f'' > 0 & f \text{ has a local minimum} \\ f'' < 0 & f \text{ has a local maximum} \\ f'' = 0 & \text{Can't tell} \end{cases}$



Definition



To minimize $f(x)$

- START WITH A GUESS x_0

- ITERATE

$$x_{i+1} = x_i - \gamma f'(x_i)$$

- STOP AFTER



Definition

To minimize $f(x)$

- START WITH A GUESS x_0
- ITERATE

$$x_{i+1} = x_i - \gamma f'(x_i)$$

- STOP AFTER some condition is MET.
 - IF THE VALUE OF x Doesn't change by more than 0.001
 -
 -

Example



Let's minimize $f(x) = e^x - x$ with an initial guess of $x_0 = 1$ and learning rate of

$$\eta = \frac{1}{2}$$

$$x_0 = 1$$

$$x_1 \approx 0.14$$

$$x_2 \approx 0.065$$

$$x_3 \approx 0.032$$

$$x_4 \approx 0.016$$

Note: The error is cut roughly in half each time, which means that each step essentially gives one more correct digit.

Example

Let's minimize $f(x) = e^x - x$ with an initial guess of $x_0 = 1$ and learning rate of $\eta = \frac{1}{2}$.

$$f'(x) = e^x - 1$$

$$\begin{aligned}x_{i+1} &= x_i - \eta f'(x_i) \\&= x_i - \frac{1}{2}(e^{x_i} - 1)\end{aligned}$$

$$x_0 = 1$$

$$x_1 \approx 0.14$$

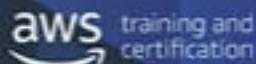
$$x_2 \approx 0.065$$

$$x_3 \approx 0.032$$

$$x_4 \approx 0.016$$

Note: The error is cut roughly in half each time, which means that each step essentially gives one more correct digit.

Example



Let's minimize $f(x) = e^x - x$ with an initial guess of $x_0 = 1$ and learning rate of

$$\eta = \frac{1}{2}$$

$$f'(x) = e^x - 1 \quad \longrightarrow \quad e^x - 1 = 0 \Rightarrow \frac{e^x - 1}{x - l_2(2)} = 0$$

$$x_{i+1} = x_i - \eta f'(x_i)$$

$$x_{i+1} = x_i - \frac{1}{2}(e^{x_i} - 1)$$

$$x_0 = 1 \quad \leftarrow$$

$$x_1 \approx 0.14 \quad \leftarrow$$

$$x_2 \approx 0.065$$

$$x_3 \approx 0.032$$

$$x_4 \approx 0.016$$

Note: The error is cut roughly in half each time, which means that each step essentially gives one more correct digit.

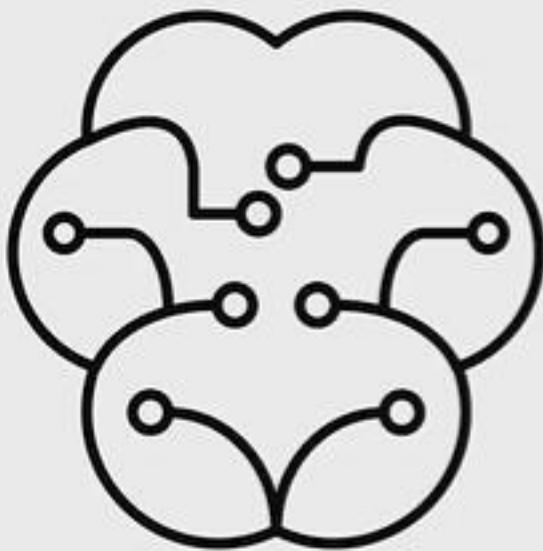
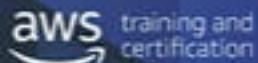
© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.



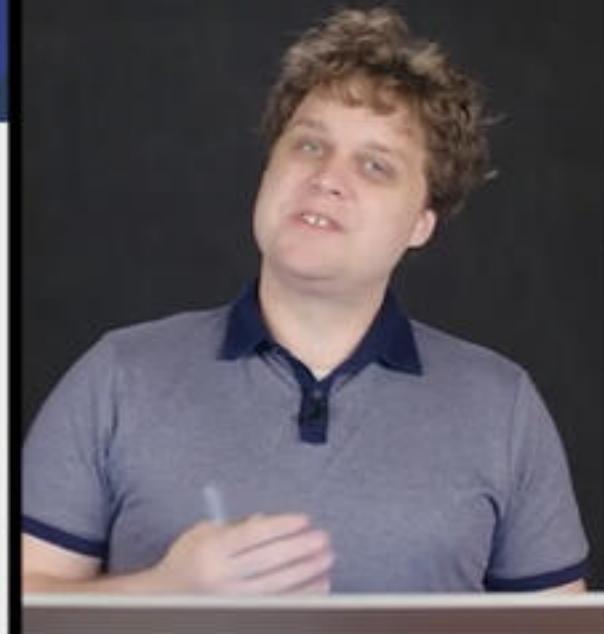
4:41



This Is Actively Used



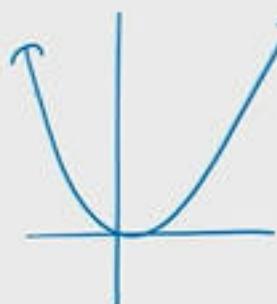
As simplistic as this is, **almost all** machine learning you have heard of use **some version** of this in the learning process.



Issue

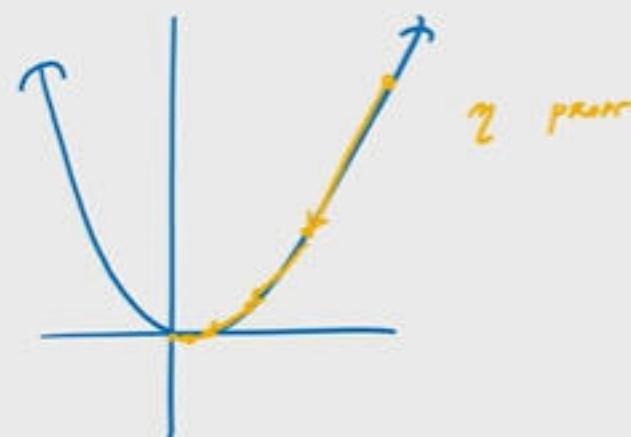


This has one major issue: **how to pick η** . An improperly chosen learning rate will cause the entire optimization procedure to either **fail** or operate **too slowly** to be of practical use. *A priori*, there is no way to know which rate is correct.



Issue

This has one major issue: **how to pick η .** An improperly chosen learning rate will cause the entire optimization procedure to either **fail** or operate **too slowly** to be of practical use. *A priori*, there is no way to know which rate is correct.

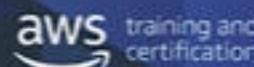


Issue

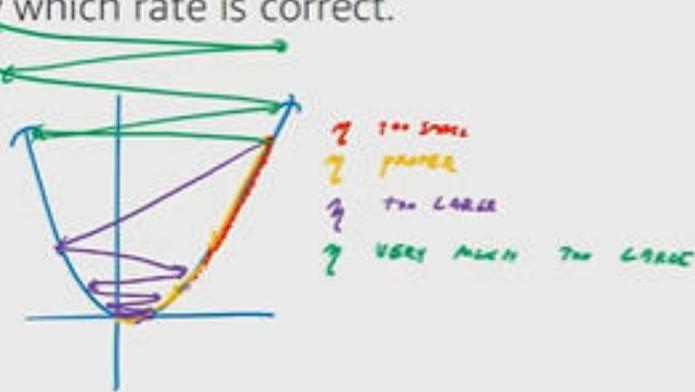
This has one major issue: **how to pick η** . An improperly chosen learning rate will cause the entire optimization procedure to either **fail** or operate **too slowly** to be of practical use. *A priori*, there is no way to know which rate is correct.



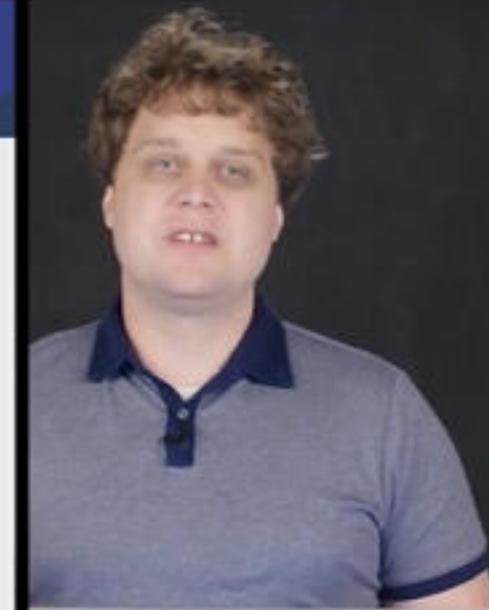
Issue



This has one major issue: **how to pick η** . An improperly chosen learning rate will cause the entire optimization procedure to either **fail** or operate **too slowly** to be of practical use. *A priori*, there is no way to know which rate is correct.



© 2018, Amazon Web Services, Inc. or its affiliates. All rights reserved.



-0:04





Thank You

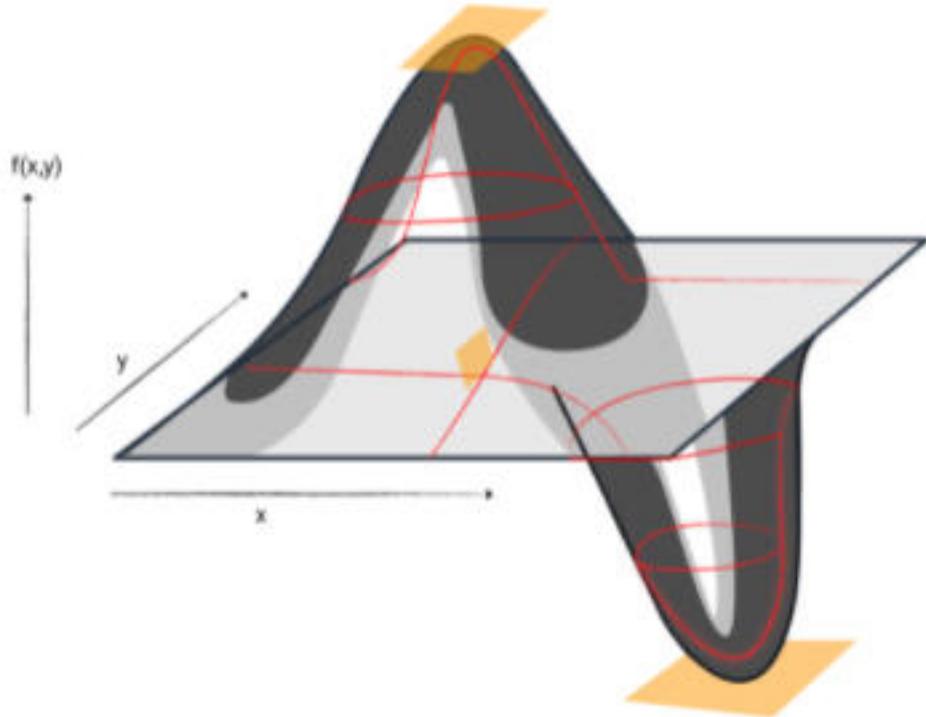
© 2018 Amazon Web Services, Inc. or its affiliates. All rights reserved. This work may not be reproduced or redistributed, in whole or in part, without prior written permission from Amazon Web Services, Inc. Commercial copying, lending, or selling is prohibited. Comments or feedback on the course, please email us at aws-course-feedback@amazon.com. For all other questions, contact us at aws.amazon.com/training/. All trademarks are the property of their owners.



-0:00

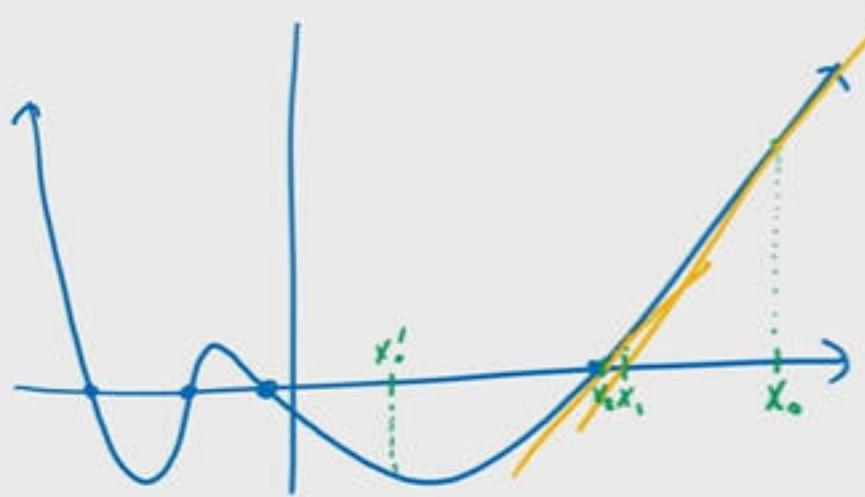


Click in the maximum, midpoint and minimum of the gradient descent to learn more about each calculation.



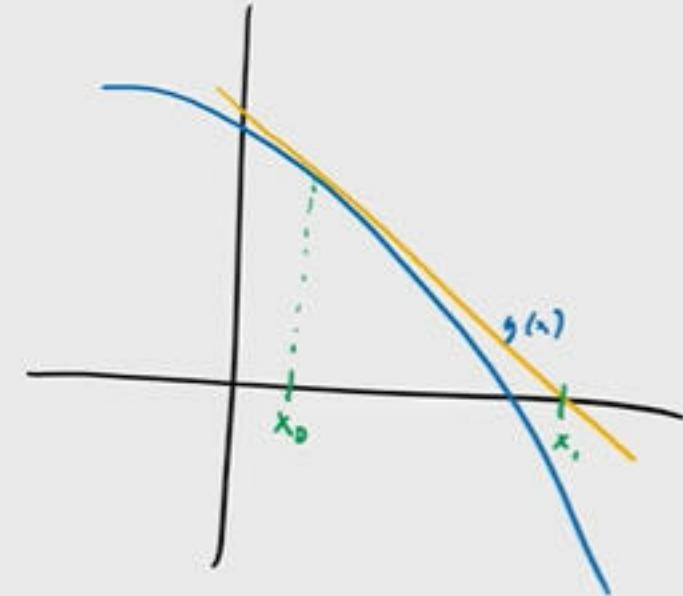
Pictorially

$g(x)$ x st $g(x) < 0$

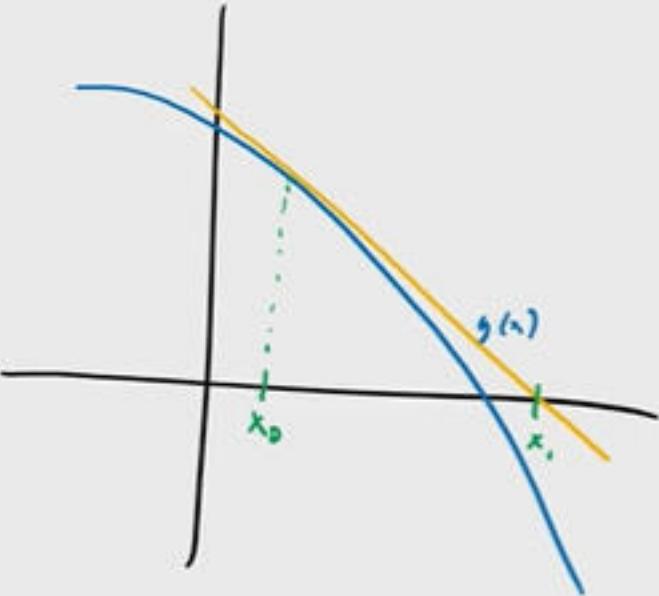


Computing the Line

line: on $(x_0, g(x_0))$
slope



Computing the Line



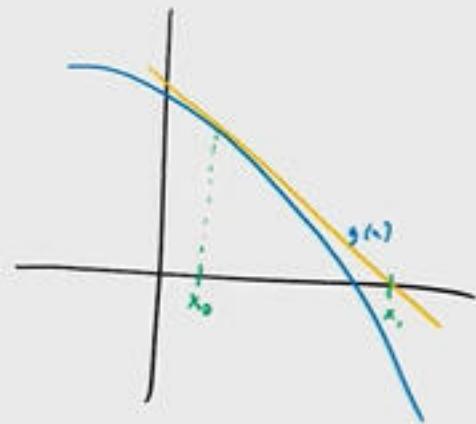
line: on $(x_0, g(x_0))$
slope $g'(x_0)$

$$y = g'(x_0)(x - x_0) + g(x_0)$$

$$-g(x_0) = g'(x_0)(x_0 - x_0)$$

(*)

Computing the Line



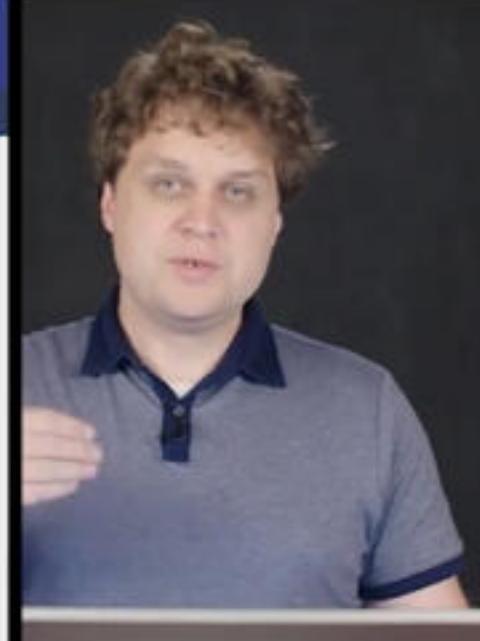
$$\text{line: } \text{on } (x_0, g(x_0)) \\ \text{slope } g'(x_0)$$

$$\text{or } y = g'(x)(x - x_0) + g(x_0)$$

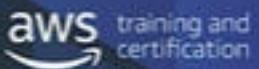
$$-g(x_0) = g'(x_0)(x_0 - x_0)$$

$$\frac{-g(x_0)}{g'(x_0)} = (x_0 - x_0)$$

$$\boxed{x_1 = x_0 - \frac{g(x_0)}{g'(x_0)}}$$



Update Step for Zero Finding



We want to find where $g(x) = 0$ and we start with some initial guess x_0 and then iterate

$$x_{i+1} = x_i - \frac{g(x_i)}{g'(x_i)}$$

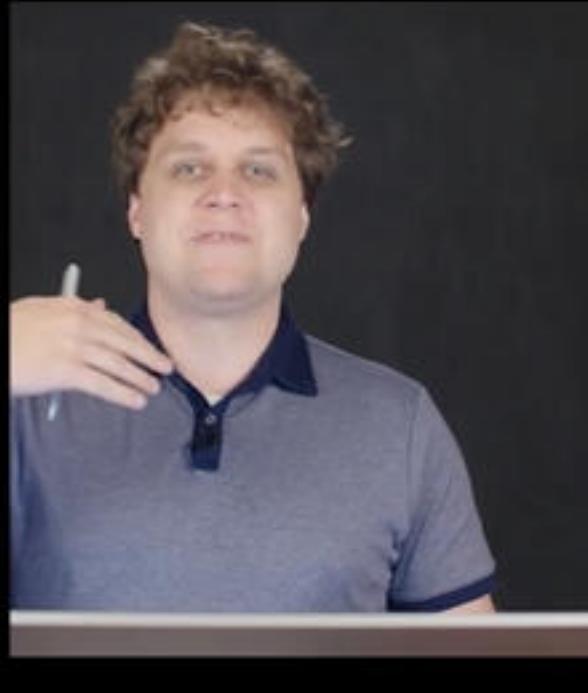


Update Step for Minimization



To minimize f , we want to find where $f'(x) = 0$ and thus we may start with some initial guess x_0 and then iterate Newton's Method on f' to get

$$x_{i+1} = x_i - \frac{f'(x_i)}{f''(x_i)}$$



Update Step for Minimization



To minimize f , we want to find where $f'(x) = 0$ and thus we may start with some initial guess x_0 and then iterate Newton's Method on f' to get

$$x_{i+1} = x_i - \frac{f'(x_i)}{f''(x_i)}$$

Learning Rate

$$g(x) = f'(x)$$



Relationship to Gradient Descent



$$x_{i+1} = x_i - \frac{f'(x_i)}{f''(x_i)}$$

Example

Let's again minimize $f(x) = e^x - x$ with an initial guess of $x_0 = 1$.

$$x_0 = 1$$

$$x_1 \approx 0.36$$

$$x_2 \approx 0.060$$

$$x_3 \approx 0.0018$$

$$x_4 \approx 0.0000016$$

Note: In this case, the number of correct digits approximately doubles.



Example

Let's again minimize $f(x) = e^x - x$ with an initial guess of $x_0 = 1$.

$$f'(x) = e^x - 1$$

$$f''(x) = e^x$$

$$x_{i+1} = x_i - \frac{e^{x_i} - 1}{e^{x_i}}$$

$$x_{i+1} = x_i - 1 + e^{-x_i}$$

$$x_0 = 1$$

$$x_1 \approx 0.36$$

$$x_2 \approx 0.060$$

$$x_3 \approx 0.0018$$

$$x_4 \approx 0.0000016$$

Note: In this case, the number of correct digits approximately doubles.



Learning Outcomes



In this track, you learned:

- Linear Algebra - Vectors and Linear Spaces
 - Vector representation, norms (L_1 , L_2 , L -infinity), inner products
 - Linear independence, orthogonality, hyperplanes, subspaces
- Linear Algebra - Matrix Theory
 - Basic matrix operations (addition, multiplication, Inverse)
 - Matrices as linear operators, rank
 - Span, linear dependence
 - Solving systems of linear equations
- Probability Fundamentals
 - Rules of probability (independence, conditional and marginal probability)
 - Bayes rule
 - Entropy of discrete probability spaces
 - Basic probability distributions (uniform, binomial, Bernoulli, Poisson, Gaussian)
 - Likelihoods, log-likelihoods, and loss functions
- Single Variable Calculus
 - Derivatives of functions of a single variable
 - Gradient descent, and Newton's method in a single variable
 - Maximum likelihood estimation of parameters
- Multi-Variable Calculus
 - Derivatives of functions of several variables
 - Derivatives of scalar functions of matrices with applications
 - Classification of stationary points
 - Gradient descent in many variables

The Gradient



- Definitions
 - Vector derivatives
 - The gradient
 - Matrix derivatives



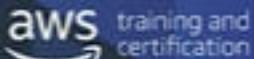
© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.



-2:53



Second Derivatives



- Note
- Definition
 - Hessian
- 2⁺D Intuition
- Example: Not Always So Simple
- An Extra Definition
 - Trace
- Classification of Critical Points in 2D
- Example



Newton's Method



- Extension of Newton's Method
- An Issue

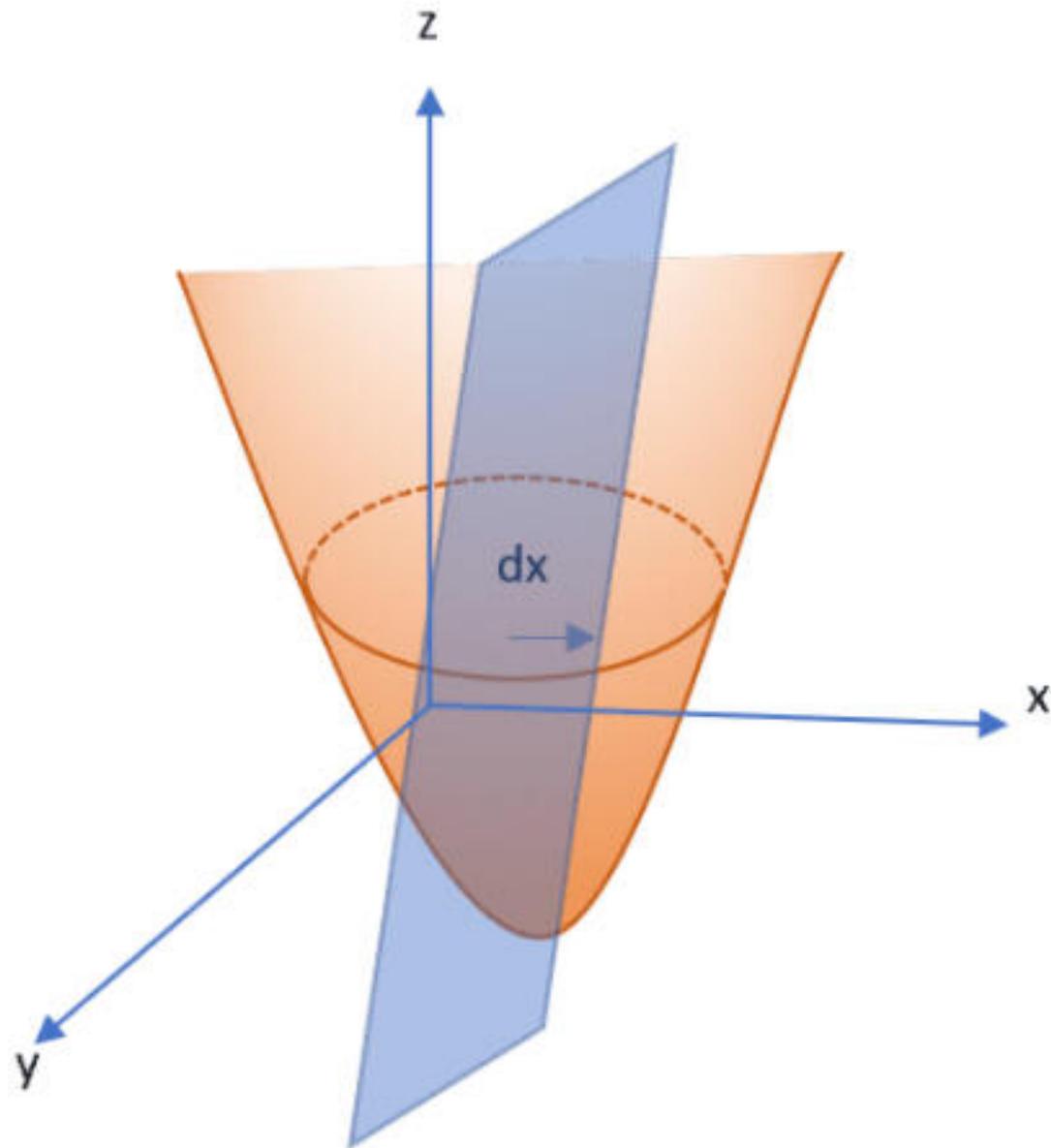


© 2018, Amazon Web Services, Inc., or its Affiliates. All rights reserved.



-1:08



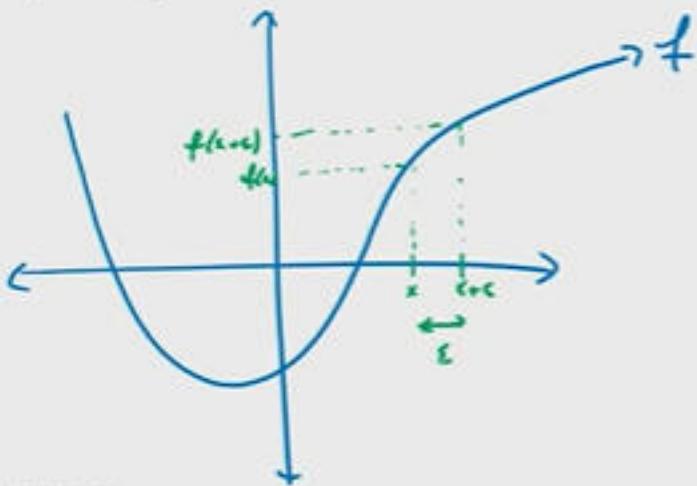


We want to see how a function responds to many variables, not just one. We need to keep track of how these variables change, and how they influence each other.

What would happen if you nudged x or y, just a little....

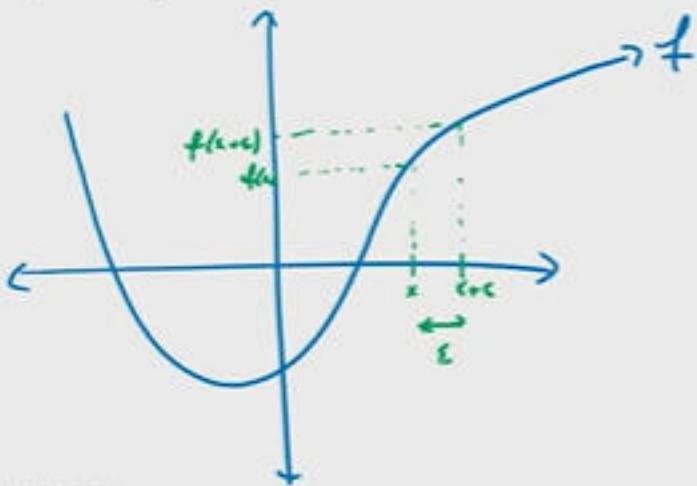
Recall

The derivative of a function encodes how that function changes when you change the inputs by a tiny bit.



Recall

The derivative of a function encodes how that function changes when you change the inputs by a tiny bit.



Issue



If you have a function:

$$f : \mathbb{R}^n \rightarrow \mathbb{R}$$

A handwritten green arrow points from the "n" in \mathbb{R}^n to the word "many" in the adjacent text.

Then it is a function of many variables.

You need to know how the function responds to changes in all of them.

The majority of this will be just bookkeeping, but will be terribly messy bookkeeping.

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.



-1:41

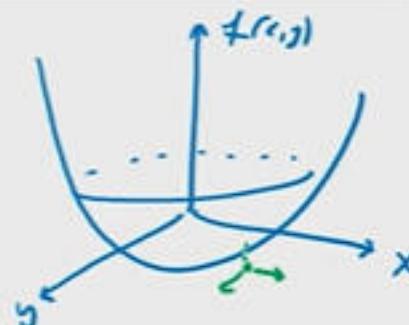


If you have a function:

$$f : \mathbb{R}^n \rightarrow \mathbb{R}$$

n-dim vector

Then it is a function of many variables.



You need to know how the function responds to changes in all of them.

The majority of this will be just bookkeeping, but will be terribly messy bookkeeping.



-1:02



$$f(x, y) = e^{x^2 + \sqrt{y}}$$

$$\frac{\partial f}{\partial x} = e^{x^2 + \sqrt{y}} \cdot (2x)$$

$$\frac{\partial f}{\partial y} = e^{x^2 + \sqrt{y}} \left(\frac{1}{2\sqrt{y}} \right)$$

As you watch the video, keep in mind that a partial derivative of a function (of two variables x and y) is a measure of the rate of change of the function... when one of the variables is subjected to a small change but the others are kept constant.

Example:

$$f(x, y) = e^{x^2 + \sqrt{y}}$$

$$\frac{\partial f}{\partial x} = e^{x^2 + \sqrt{y}} \cdot (2x)$$

$$\frac{\partial f}{\partial x}(1) =$$

Partial Derivative

$$f(x, y)$$

$\frac{\partial f}{\partial x} = \text{derivative of } f \text{ where we think of}$
 y as fixed



-2:47



Definition

Partial Derivative

$f(x, y)$

$\frac{\partial f}{\partial x} = \text{derivative of } f$ where we think of
y as fixed and x as changing



Example

$$f(x, y) = e^{x^2 + \sqrt{y}}$$



Example

$$f(x, y) = e^{x^2 + \sqrt{y}}$$

$$\frac{\partial f}{\partial x} = e^{x^2 + \sqrt{y}} \cdot (2x)$$

$$\frac{\partial f}{\partial y} =$$



Topics



- Definitions
 - Vector derivatives
 - The gradient
 - Matrix derivatives
- Example
- Visualizing the Gradient
- Key Properties
- Gradient Descent
- Pictorially: Level Sets
- Example

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.



-19:02



Definitions

Vector Derivative

$f(\vec{x})$

$\vec{x} = (x_1, x_2, \dots, x_n)$

$\frac{\partial f}{\partial \vec{x}} = \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \dots, \frac{\partial f}{\partial x_n} \right)$



Definitions



The Gradient

$f(\vec{x})$

$\vec{x} = (x_1, \dots, x_n)$

∇f

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.



-17:05



Definitions

Matrix Derivatives

$$f(A)$$

$$A = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix}$$

$$\frac{\partial f}{\partial A} = \begin{pmatrix} \frac{\partial f}{\partial a_{11}} & \cdots & \frac{\partial f}{\partial a_{1n}} \\ \vdots & & \vdots \\ \frac{\partial f}{\partial a_{n1}} & \cdots & \frac{\partial f}{\partial a_{nn}} \end{pmatrix}$$

Example

$$f(x, y) = e^{x^2 + \sqrt{y}}, \quad f_x = \frac{\partial f}{\partial x} = 2xe^{x^2 + \sqrt{y}}, \quad f_y = \frac{\partial f}{\partial y} = \frac{e^{x^2 + \sqrt{y}}}{2\sqrt{y}}$$

Example

$$f(x, y) = e^{x^2 + \sqrt{y}}, \quad f_x = \frac{\partial f}{\partial x} = 2xe^{x^2 + \sqrt{y}}, \quad f_y = \frac{\partial f}{\partial y} = \frac{e^{x^2 + \sqrt{y}}}{2\sqrt{y}}$$

 $f(v)$ $v = (x, y)$

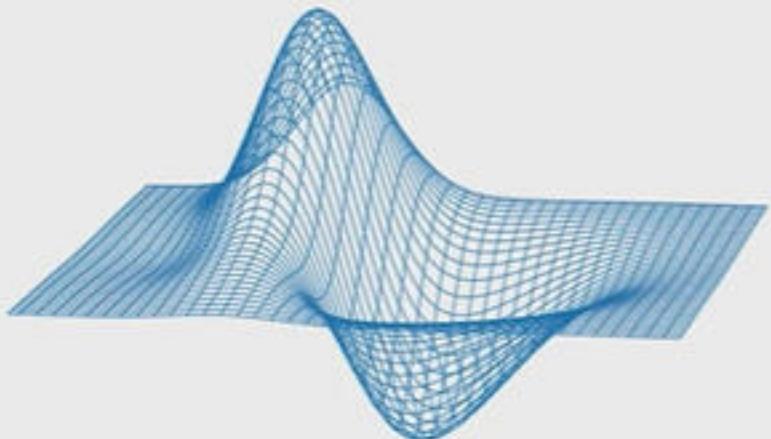
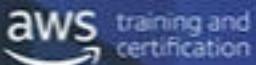
$$\frac{\partial f}{\partial v} = \left(2x e^{x^2 + \sqrt{y}}, \frac{e^{x^2 + \sqrt{y}}}{2\sqrt{y}} \right) = e^{x^2 + \sqrt{y}} (2,$$



13:23



Visualizing the Gradient



© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.



-12:11



Visualizing the Gradient

$$\frac{\partial f}{\partial \vec{x}} = \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2} \right)$$

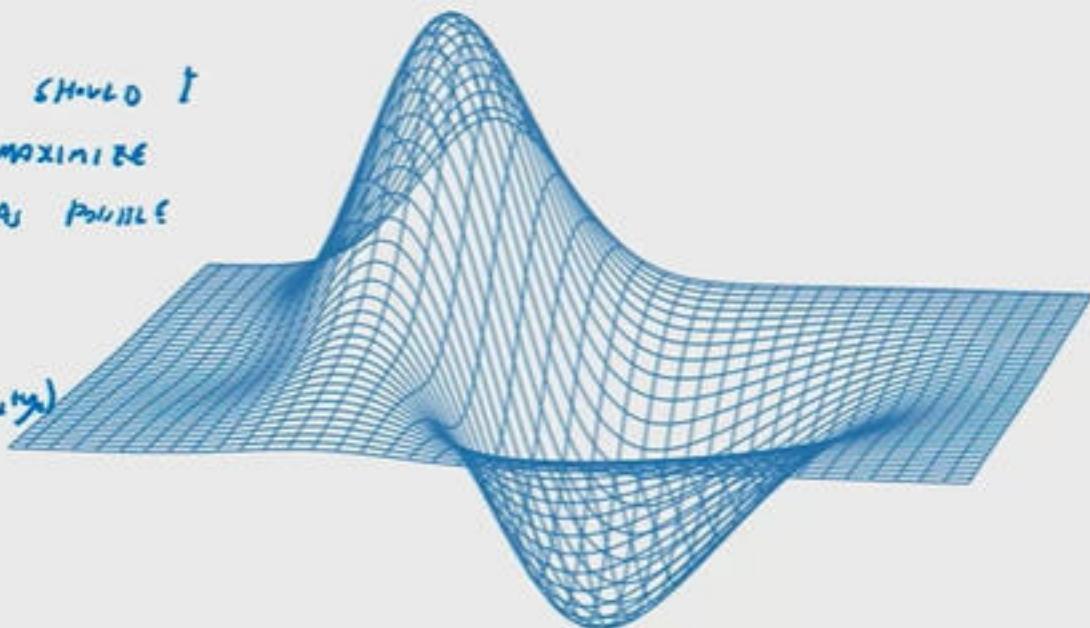
WHAT DIRECTION \vec{y} SHOULD I

HEAD IN TO MAXIMIZE

f AS QUICKLY AS POSSIBLE

$$\vec{y} = (y_1, y_2)$$

$$f(x_1, x_2) \rightarrow f(x_1 + y_1, x_2 + y_2)$$



Visualizing the Gradient

$$\frac{\partial \mathbf{f}}{\partial \vec{x}} = \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2} \right)$$

WHAT DIRECTION \vec{y} SHOULD I
HEAD IN TO MAXIMIZE
 f AS QUICKLY AS POSSIBLE

$$y = (y_1, y_2)$$

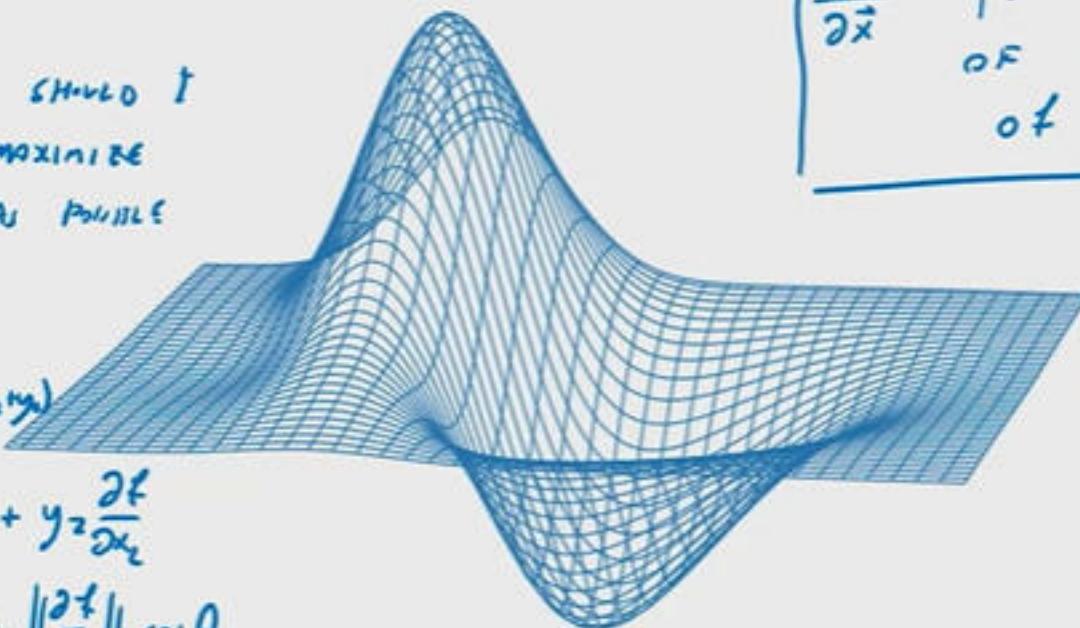
$$f(x_1, x_2) \rightarrow f(x_1 + y_1, x_2 + y_2)$$

$$\nabla f(x_1, x_2) + y_1 \frac{\partial f}{\partial x_1} + y_2 \frac{\partial f}{\partial x_2}$$

$$\vec{y} \cdot \frac{\partial \mathbf{f}}{\partial \vec{x}} = \|\vec{y}\| \cdot \left\| \frac{\partial \mathbf{f}}{\partial \vec{x}} \right\| \cdot \cos \theta$$

$\theta = 0$

$\frac{\partial \mathbf{f}}{\partial \vec{x}}$ POINTS IN THE DIRECTION
OF MAXIMUM INCREASE
of f



Visualizing the Gradient

$$\frac{\partial \mathbf{f}}{\partial \vec{x}} = \left(\frac{\partial \mathbf{f}}{\partial x_1}, \frac{\partial \mathbf{f}}{\partial x_2} \right)$$

WHAT DIRECTION \vec{y} SHOULD I
HEAD IN TO MAXIMIZE
 f AS QUICKLY AS POSSIBLE

$$y = (y_1, y_2)$$

$$f(x_1, x_2) \rightarrow f(x_1 + y_1, x_2 + y_2)$$

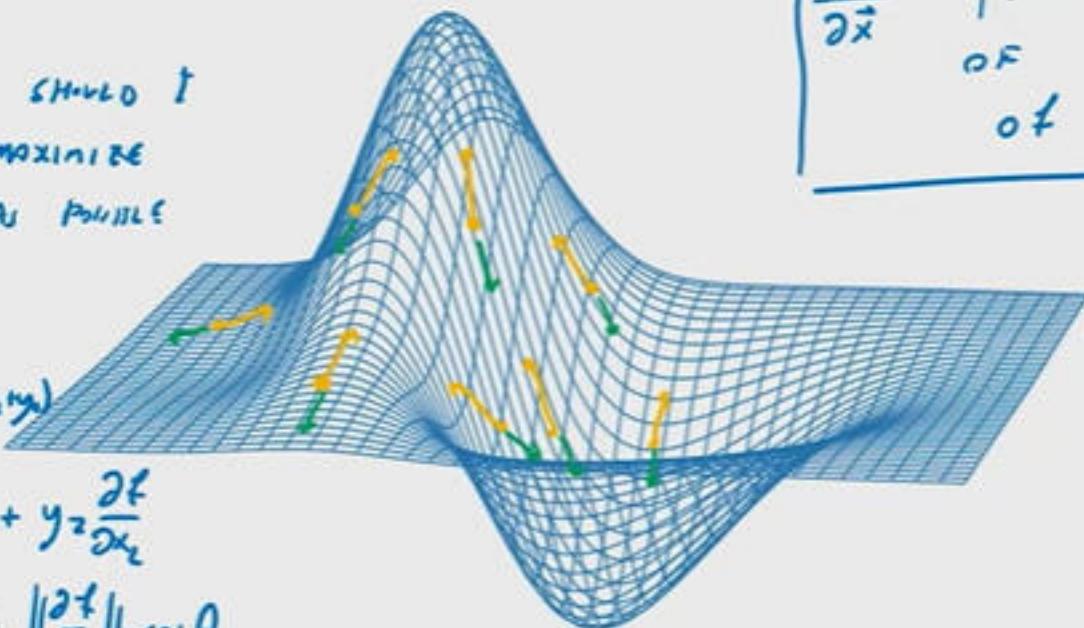
$$\nabla f(x_1, x_2) + y_1 \frac{\partial f}{\partial x_1} + y_2 \frac{\partial f}{\partial x_2}$$

$$\vec{y} \cdot \frac{\partial \mathbf{f}}{\partial \vec{x}} = \|\vec{y}\| \cdot \left\| \frac{\partial \mathbf{f}}{\partial \vec{x}} \right\| \cdot \cos \theta$$

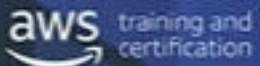
$\theta = 0$

$\frac{\partial \mathbf{f}}{\partial \vec{x}}$ POINTS IN THE DIRECTION
OF MAXIMUM INCREASE
of f

$$\frac{\partial \mathbf{f}}{\partial \vec{x}} = r$$



Key Properties



- The gradient points in the direction of **maximum increase**.
- $-\nabla f$ points in the direction of **maximum decrease**.
- **Maximums** and **minimums** have $\nabla f = \frac{\partial f}{\partial x} = 0$.



Gradient Descent

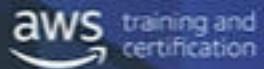


Given an initial guess \vec{x}_0 and a learning rate η , you can attempt to find a minimum for f by iterating

$$\vec{x}_{i+1} = \vec{x}_i - \eta \nabla f(\vec{x}_i).$$



Gradient Descent



Given an initial guess \vec{x}_0 and a learning rate η , you can attempt to find a minimum for f by iterating

$$\vec{x}_{i+1} = \vec{x}_i - \eta \nabla f(\vec{x}_i).$$



© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.



-5:34



Gradient Descent



Given an initial guess \dot{x}_0 and a learning rate η , you can attempt to find a minimum for f by iterating

$$\dot{x}_{i+1} = \dot{x}_i - \eta \nabla f(x_i).$$

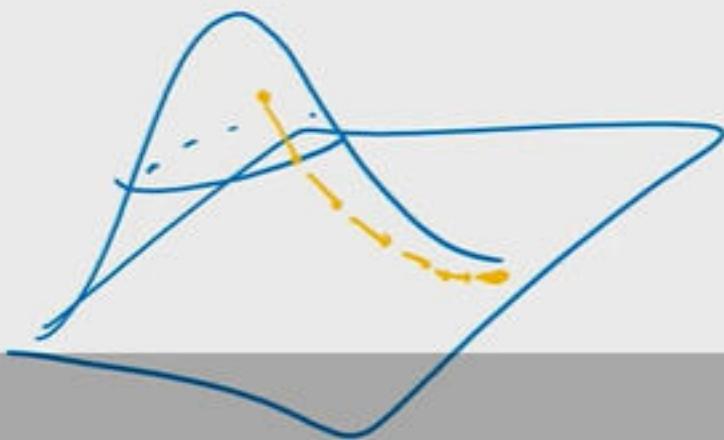


Gradient Descent

Given an initial guess \vec{x}_0 and a learning rate η , you can attempt to find a minimum for f by iterating

$$\vec{x}_{i+1} = \vec{x}_i - \eta \nabla f(\vec{x}_i).$$

$$\begin{pmatrix} x_{i+1,1} \\ x_{i+1,2} \\ \vdots \\ x_{i+1,n} \end{pmatrix}$$



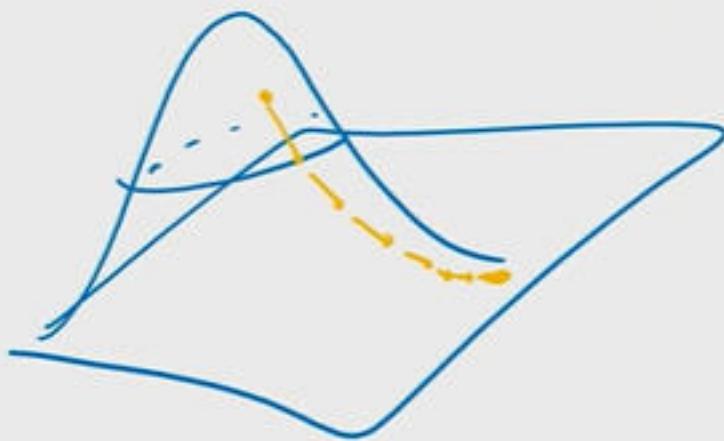
4:57



Gradient Descent

Given an initial guess \vec{x}_0 and a learning rate η , you can attempt to find a minimum for f by iterating

$$\vec{x}_{i+1} = \vec{x}_i - \eta \nabla f(\vec{x}_i).$$

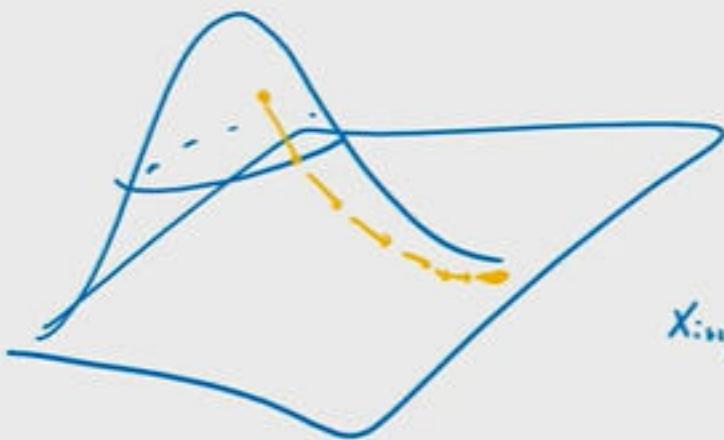


$$\begin{pmatrix} x_{i+1,1} \\ x_{i+1,2} \\ \vdots \\ x_{i+1,n} \end{pmatrix} = \begin{pmatrix} x_{i,1} \\ x_{i,2} \\ \vdots \\ x_{i,n} \end{pmatrix} - \eta \begin{pmatrix} \frac{\partial L}{\partial x_1} \\ \frac{\partial L}{\partial x_2} \\ \vdots \\ \frac{\partial L}{\partial x_n} \end{pmatrix}$$

Gradient Descent

Given an initial guess \vec{x}_0 and a learning rate η , you can attempt to find a minimum for f by iterating

$$\vec{x}_{i+1} = \vec{x}_i - \eta \nabla f(\vec{x}_i).$$



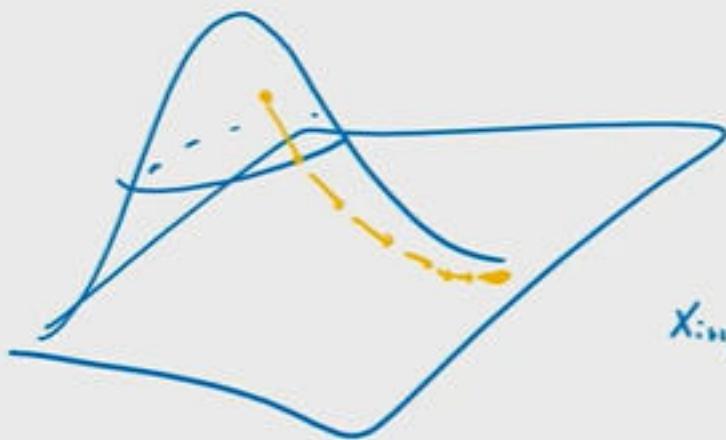
$$x_{i+1,j} = x_{i,j} - \eta \frac{\partial L}{\partial x_j}$$

$$\begin{pmatrix} x_{i+1,1} \\ x_{i+1,2} \\ \vdots \\ x_{i+1,n} \end{pmatrix} = \begin{pmatrix} x_{i,1} \\ x_{i,2} \\ \vdots \\ x_{i,n} \end{pmatrix} - \eta \begin{pmatrix} \frac{\partial L}{\partial x_1} \\ \frac{\partial L}{\partial x_2} \\ \vdots \\ \frac{\partial L}{\partial x_n} \end{pmatrix}$$

Gradient Descent

Given an initial guess \vec{x}_0 and a learning rate η , you can attempt to find a minimum for f by iterating

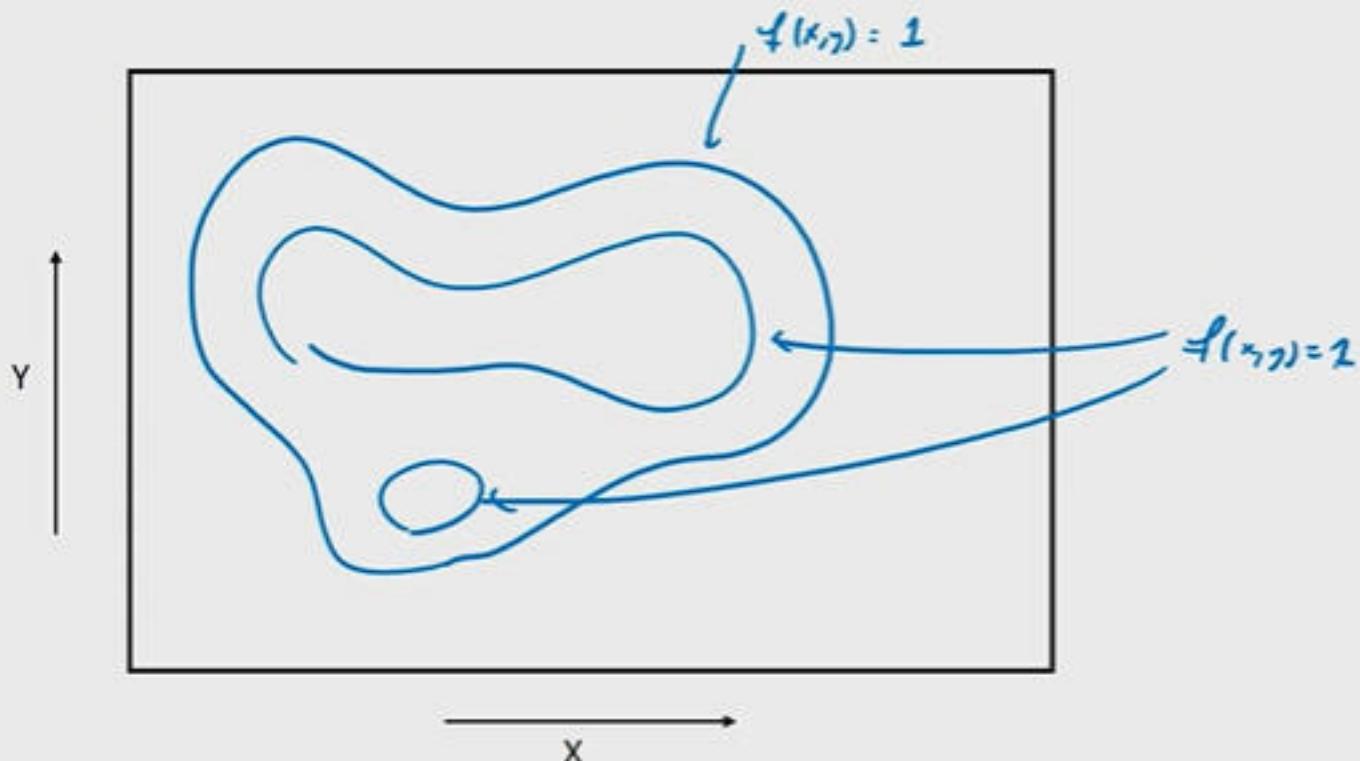
$$\vec{x}_{i+1} = \vec{x}_i - \eta \nabla f(\vec{x}_i).$$



$$x_{i+1,j} = x_{i,j} - \eta \frac{\partial L}{\partial x_j}$$

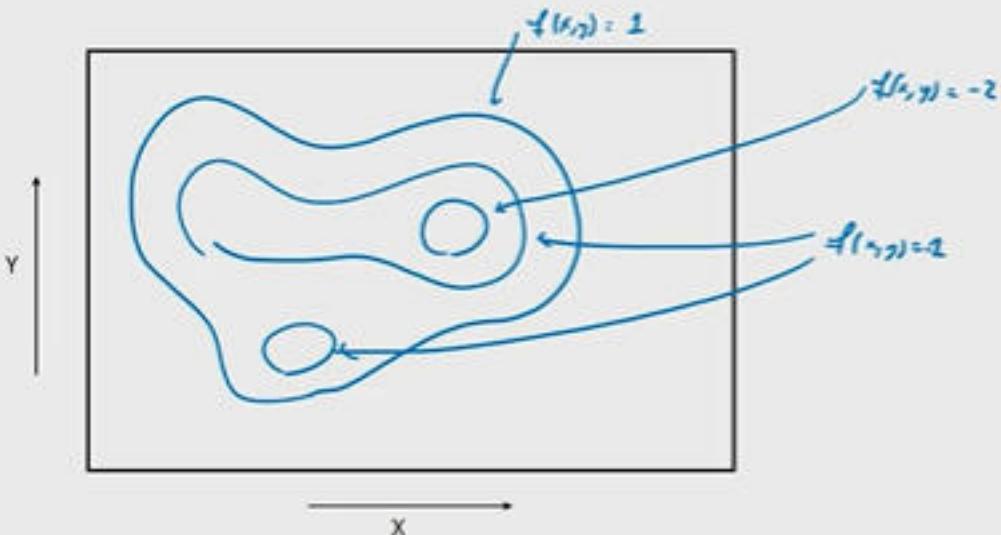
$$\begin{pmatrix} x_{i+1,1} \\ x_{i+1,2} \\ \vdots \\ x_{i+1,n} \end{pmatrix} = \begin{pmatrix} x_{i,1} \\ x_{i,2} \\ \vdots \\ x_{i,n} \end{pmatrix} - \eta \begin{pmatrix} \frac{\partial L}{\partial x_1} \\ \frac{\partial L}{\partial x_2} \\ \vdots \\ \frac{\partial L}{\partial x_n} \end{pmatrix}$$

Pictorially: Level Sets



Pictorially: Level Sets

aws training and certification



© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

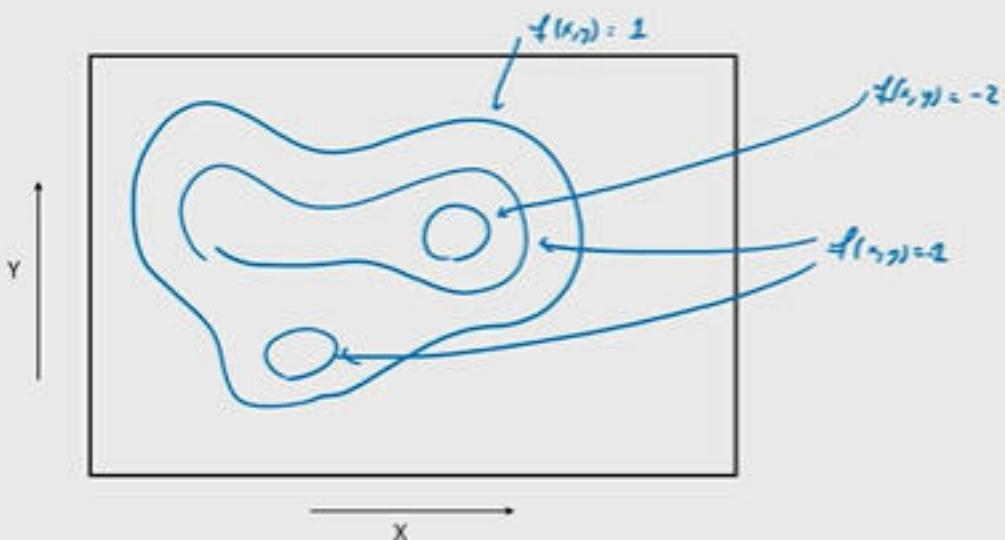


-2:57



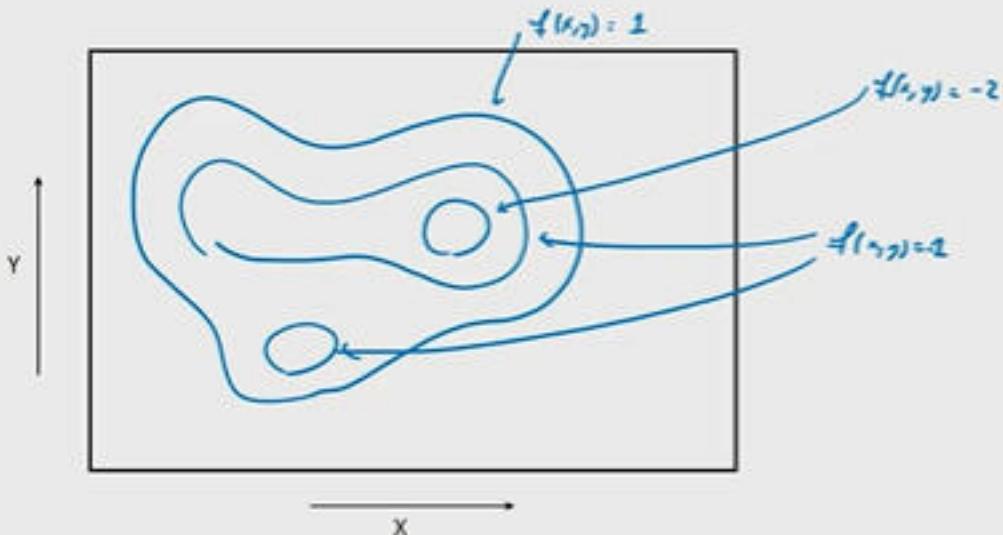
Pictorially: Level Sets

aws training and certification



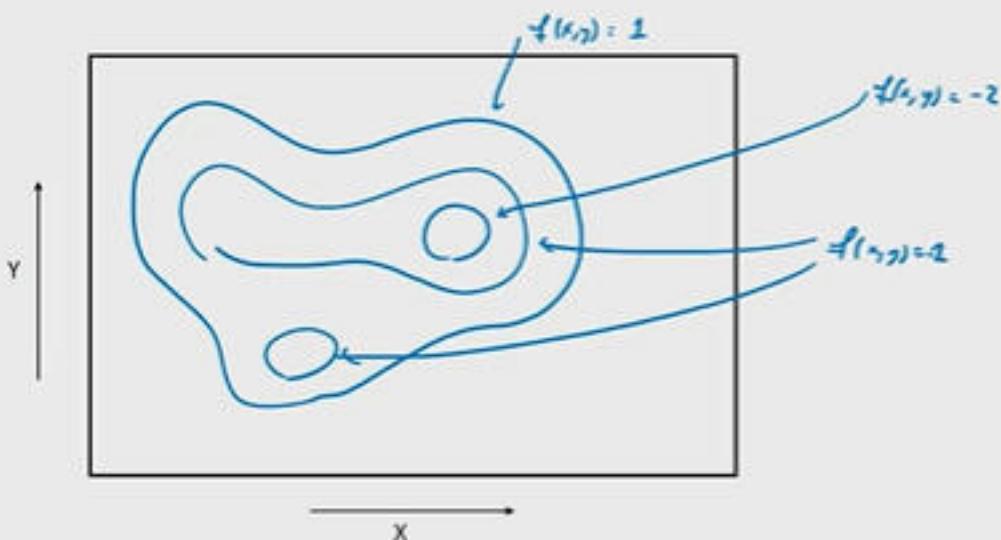
Pictorially: Level Sets

aws training and certification



Pictorially: Level Sets

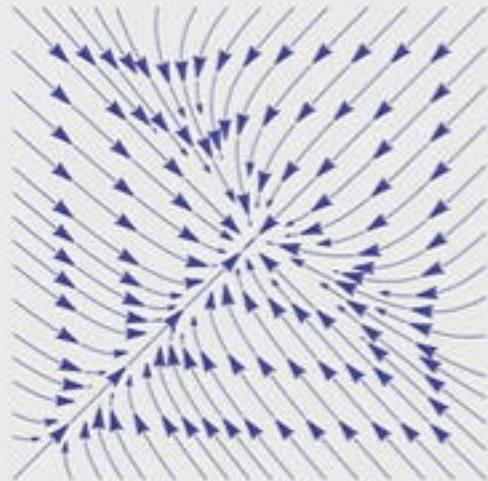
aws training and certification



Example



$$f(x, y) = (x - y)^2 + e^{x+y} - x - y$$



© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

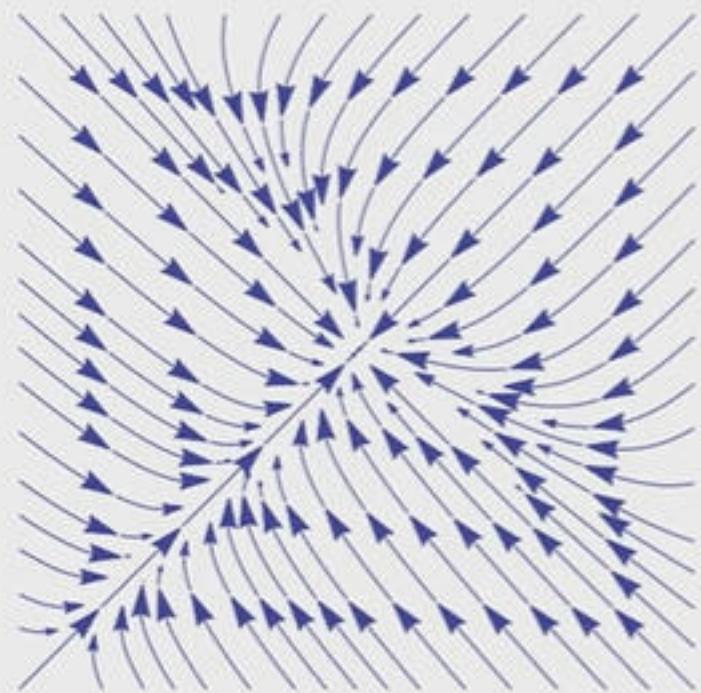


-1:44



Example

$$f(x, y) = (x - y)^2 + e^{x+y} - x - y$$

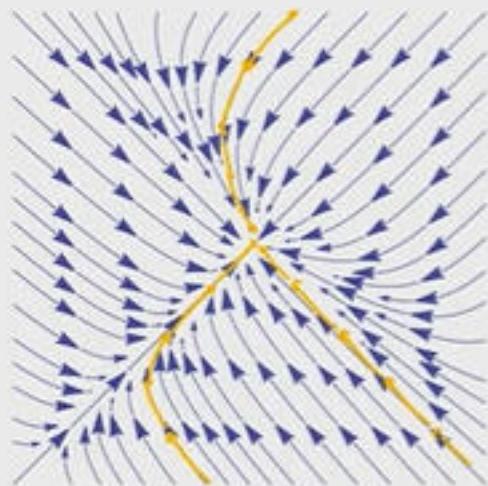


$$\frac{\partial f}{\partial x} = 2(x-y) + e^{x+y} - 1$$

$$\frac{\partial f}{\partial y} = -2(x-y) + e^{x+y} - 1$$

Example

$$f(x, y) = (x - y)^2 + e^{x+y} - x - y$$



$$\frac{\partial f}{\partial x} = 2(x-y) + e^{x+y} - 1$$

$$\frac{\partial f}{\partial y} = -2(x-y) + e^{x+y} - 1$$



Example 1

$$\frac{\partial}{\partial \vec{x}} \left(\vec{\beta} \vec{x} \right) = \underbrace{\quad}_{\text{column}}$$

$$\frac{\partial}{\partial x_i} (\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \dots + \beta_n x_n)$$

$$\frac{\partial}{\partial \vec{x}} \left(\vec{\beta} \vec{x} \right) = \begin{pmatrix} \frac{\partial \vec{\beta} \vec{x}}{\partial x_1} \\ \vdots \\ \frac{\partial \vec{\beta} \vec{x}}{\partial x_n} \end{pmatrix}$$

Example 1

$$\frac{\partial}{\partial \vec{x}} \left(\vec{\beta} \vec{x} \right) = \underbrace{\vec{\beta}}_{\text{column}}$$

$$\begin{aligned} & \frac{\partial}{\partial x_i} (\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \dots + \beta_n x_n) \\ &= \beta_i \end{aligned}$$

$$\frac{\partial}{\partial \vec{x}} \left(\vec{\beta} \vec{x} \right) = \begin{pmatrix} \frac{\partial \vec{\beta} \vec{x}}{\partial x_1} \\ \vdots \\ \frac{\partial \vec{\beta} \vec{x}}{\partial x_n} \end{pmatrix}$$

Example 3

$$\frac{\partial}{\partial \vec{x}} (\vec{x}^T A \vec{x}) = \frac{\partial}{\partial \vec{x}} \left(\sum_{i,j} x_i a_{ij} x_j \right)$$

$$\frac{\partial}{\partial x_k} \left(\sum_{i,j} x_i a_{ij} x_j \right) = \frac{\partial}{\partial x_k} \left(\sum_{\substack{i,j \\ i \neq k \\ j \neq k}} x_i a_{ij} x_j + \sum_{\substack{i,j \\ i=k \\ j \neq k}} x_i a_{ij} x_j + \sum_{\substack{i,j \\ i \neq k \\ j=k}} x_i a_{ij} x_j \right)$$



-13:42



Example 3

$$\frac{\partial}{\partial \vec{x}} (\vec{x}^T A \vec{x}) = \frac{\partial}{\partial x_i} \left(\sum_{i,j} x_i a_{ij} x_j \right)$$

$$\begin{aligned} \frac{\partial}{\partial x_k} \left(\sum_{i,j} x_i a_{ij} x_j \right) &= \frac{\partial}{\partial x_k} \left(\sum_{\substack{i,j \\ i \neq k \\ j \neq k}} x_i a_{ij} x_j + \sum_{i \neq k} x_i a_{ik} x_{ik} + \sum_{j \neq k} x_k a_{kj} x_j + x_k a_{kk} x_k \right) \\ &= \left(0 + \sum_{i \neq k} x_i a_{ik} + \sum_{j \neq k} a_{kj} \right) \end{aligned}$$



-11:58



Example 3



$$\begin{aligned}\frac{\partial}{\partial \vec{x}} (\vec{x}^T A \vec{x}) &= \frac{\partial}{\partial \vec{x}} \left(\sum_{i,j} x_i a_{ij} x_j \right) \\ \frac{\partial}{\partial x_k} \left(\sum_{i,j} x_i a_{ij} x_j \right) &= \frac{\partial}{\partial x_k} \left(\sum_{\substack{i,j \\ i \neq k \\ j \neq k}} x_i a_{ij} x_i + \sum_{i \neq k} x_i a_{ik} x_k + \sum_{j \neq k} x_k a_{kj} x_j + x_k a_{kk} x_k \right) \\ &= \left(0 + \underbrace{\sum_{i \neq k} x_i a_{ik}}_{\text{Term 1}} + \underbrace{\sum_{j \neq k} a_{kj} x_j}_{\text{Term 2}} + \underbrace{2 a_{kk} x_k}_{\text{Term 3}} \right) =\end{aligned}$$



Example 3



$$\begin{aligned}\frac{\partial}{\partial \vec{x}} (\vec{x}^T A \vec{x}) &= \frac{\partial}{\partial \vec{x}} \left(\sum_{i,j} x_i a_{ij} x_j \right) \\ \rightarrow \frac{\partial}{\partial x_k} \left(\sum_{i,j} x_i a_{ij} x_j \right) &= \frac{\partial}{\partial x_k} \left(\sum_{\substack{i,j \\ i \neq k \\ j \neq k}} x_i a_{ij} x_i + \sum_{i \neq k} x_i a_{ik} x_k + \sum_{j \neq k} x_k a_{kj} x_j + x_k a_{kk} x_k \right) \\ &= \left(0 + \underbrace{\sum_{i \neq k} x_i a_{ik} x_i}_{\text{underbrace}} + \underbrace{\sum_{j \neq k} a_{kj} x_j}_{\text{underbrace}} + \cancel{2 a_{kk} x_k} \right) = \sum_i x_i a_{ik} + \sum_j a_{kj} x_j \\ &= \sum_i (\underbrace{a_{ik} + a_{ki}}_{\text{underbrace}}) x_i \quad \downarrow \quad \boxed{\frac{\partial}{\partial \vec{x}} (\vec{x}^T A \vec{x}) = [A^T + A] \vec{x}} \\ &= \sum_i [(A^T + A)]_{ki} x_i = \boxed{[(A^T + A) \vec{x}]_k} \quad \frac{d}{dx} (\epsilon a x) = (\alpha + \epsilon) x\end{aligned}$$



Example 4

$$\frac{\partial}{\partial X} \| CX \|^2 = \frac{\partial}{\partial X} \sum_{i,j} \left(\sum_k c_{ik} x_{kj} \right)^2$$

$$\frac{\partial}{\partial x_{ab}} \left(\sum_{i,j} \left(\sum_k c_{ik} x_{kj} \right)^2 \right) = \sum_{i,j} 2 \left(\sum_k c_{ik} x_{kj} \right) \cdot c$$



-5:07



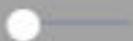
Example 4

$$\frac{\partial}{\partial X} \| CX \|^2 = \frac{\partial}{\partial X} \sum_{i,j} \left(\sum_k c_{ik} x_{kj} \right)^2$$
$$\frac{\partial}{\partial x_{ab}} \left(\sum_{i,j} \left(\sum_k c_{ik} x_{kj} \right)^2 \right) = \sum_{i,j} 2 \left(\sum_k c_{ik} x_{kj} \right) \cdot \underbrace{\frac{\partial}{\partial x_{ab}} \left(\sum_k c_{ik} x_{kj} \right)}_{= \sum_i 2 \left(\sum_k c_{ik} x_{kb} \right) \frac{\partial}{\partial x_{ab}} \left(\sum_k c_{ik} x_{kb} \right)} = \sum_i 2 \sum_k \left(\sum_k c_{ik} x_{kb} \right) c_{ia}$$



Example 4

$$\begin{aligned}\frac{\partial}{\partial X} \|CX\|^2 &= \frac{\partial}{\partial X} \sum_{i,j} \left(\sum_k c_{ik} x_{kj} \right)^2 \\ \frac{\partial}{\partial x_{ab}} \left(\sum_{i,j} \left(\sum_k c_{ik} x_{kj} \right)^2 \right) &= \sum_{i,j} 2 \left(\sum_k c_{ik} x_{kj} \right) \cdot \underbrace{\frac{\partial}{\partial x_{ab}} \left(\sum_k c_{ik} x_{kj} \right)}_{[CX]_{i,b}} = \sum_i 2 \left(\sum_k c_{ik} x_{kb} \right) \frac{\partial}{\partial x_{ab}} \left(\sum_k c_{ik} x_{kb} \right) \\ &= 2 \sum_i \underbrace{\left(\sum_k c_{ik} x_{kb} \right)}_{[CX]_{i,b}} c_{ia} = 2 \sum_i [CX]_{i,b} c_{ia} = 2 \sum_i [C^T]\end{aligned}$$



Example 4

$$\begin{aligned} \frac{\partial}{\partial X} \|CX\|^2 &= \frac{\partial}{\partial X} \sum_{i,j} \left(\sum_k c_{ik} x_{kj} \right)^2 \\ \frac{\partial}{\partial x_{ab}} \left(\sum_{i,j} \left(\sum_k c_{ik} x_{kj} \right)^2 \right) &= \sum_{i,j} 2 \left(\sum_k c_{ik} x_{kj} \right) \cdot \underbrace{\frac{\partial}{\partial x_{ab}} \left(\sum_k c_{ik} x_{kj} \right)}_{= 2 \sum_i \left(\sum_k c_{ik} x_{kb} \right) c_{ia}} = \sum_i 2 \left(\sum_k c_{ik} x_{kb} \right) \frac{\partial}{\partial x_{ab}} \left(\sum_k c_{ik} x_{kb} \right) \\ &= 2 \sum_i \underbrace{\left(\sum_k c_{ik} x_{kb} \right)}_{[CX]_{i,b}} c_{ia} = 2 \sum_i [CX]_{i,b} c_{ia} = 2 \sum_i [C^T]_{a,i} [CX]_{i,b} \\ &= [2 C^T C X]_{a,b} \Rightarrow \boxed{\frac{\partial}{\partial X} \|CX\|^2 = 2 C^T C X} \\ &\quad \text{d } (cx)^2 \end{aligned}$$



Extension of Newton's Method



You may extend Newton's method to higher dimensions with the **Hessian**. We'll skip the derivation, but the result is

$$x_{n+1} = x_n - [Hf(x_n)]^{-1} \nabla f(x_n)$$



Extension of Newton's Method



You may extend Newton's method to higher dimensions with the **Hessian**. We'll skip the derivation, but the result is

$$x_{n+1} = x_n - [Hf(x_n)]^{-1} \nabla f(x_n)$$



Extension of Newton's Method



You may extend Newton's method to higher dimensions with the **Hessian**. We'll skip the derivation, but the result is

$$x_{n+1} = x_n - [Hf(x_n)]^{-1} \nabla f(x_n)$$

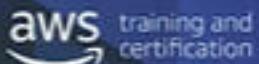
$$x_{n+1} = x_n - \frac{1}{\nabla^2 f(x_n)}.$$



-2:24



An Issue



The computational complexity of inverting an $n \times n$ matrix is not actually known, but the best known algorithm is $O(n^{2.373})$.

For high dimensional data sets, anything past linear time in the dimensions is often impractical, so Newton's Method is reserved for a few hundred dimensions at most.



An Issue



The computational complexity of inverting an $n \times n$ matrix is not actually known, but the best known algorithm is $O(n^{2.373})$.

For high dimensional data sets, anything past linear time in the dimensions is often impractical, so Newton's Method is reserved for a few hundred dimensions at most.



An Issue

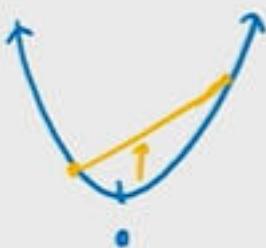


The computational complexity of inverting an $n \times n$ matrix is not actually known, but the best known algorithm is $O(n^{2.373})$.

For high dimensional data sets, anything past linear time in the dimensions is often impractical, so Newton's Method is reserved for a few hundred dimensions at most.



$$f(x) = x^2$$



$$f(x) = e^x - x$$



⇒ convexity

The Benefits

- ONE UNIQUER MINIMUM
- NO LO



-3:38



The Benefits

- ONE UNIQUE MINIMUM
- NO LO



-3:38



The Benefits

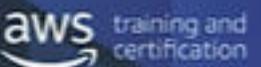
- ONE UNIQUE MINIMUM
- NO LAG



-3:38



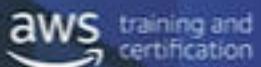
A Warning



- *Neural Networks*



A Warning



- *Neural Networks*

⇒ many local minima & many saddle points.



Topics



- Note
- Definition
 - Hessian
- 2+D Intuition
- Example: Not Always So Simple
- An Extra Definition
 - Trace
- Classification of Critical Points in 2D
- Example



Note



For a function from $\mathbb{R}^n \rightarrow \mathbb{R}$, there are now n^2 many derivatives.

$$f_{x_i x_j} = \frac{\partial^2 f}{\partial x_i \partial x_j}$$



Note



For a function from $\mathbb{R}^n \rightarrow \mathbb{R}$, there are now n^2 many derivatives.

$$f_{x_i x_j} = \underbrace{\frac{\partial^2 f}{\partial x_i \partial x_j}}_{h^2 \text{ new terms}}$$



Definition

Hessian

$$f: \mathbb{R}^n \rightarrow \mathbb{R}$$
$$H_f = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_n} \end{bmatrix}$$

2+D Intuition



$$f(x,y) = x^2 + y^2$$



$$f(x,y) = x^2 - y^2$$



$$f(x,y) = -x^2 - y^2$$

If the matrix is diagonal, a positive entry is a direction where it curves up, and a negative entry is a direction where it curves down.

2+D Intuition



$$f(x,y) = x^2 + y^2$$



$$f(x,y) = x^2 - y^2$$



$$f(x,y) = -x^2 - y^2$$



If the matrix is diagonal, a positive entry is a direction where it curves up, and a negative entry is a direction where it curves down.

2⁺D Intuition

aws training and certification

$$f(x,y) = x^2 + y^2$$



$$\frac{\partial f}{\partial x} = 2x$$
$$\frac{\partial f}{\partial y} = 2y$$

$$Hf = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$

$$f(x,y) = x^2 - y^2$$



$$f(x,y) = -x^2 - y^2$$



If the matrix is diagonal, a positive entry is a direction where it curves up, and a negative entry is a direction where it curves down.

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.



2+D Intuition

$$f(x, y) = x^2 + y^2$$



$$\frac{\partial f}{\partial x} = 2x$$

$$\frac{\partial f}{\partial y} = 2y$$

$$Hf = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$

$$f(x, y) = x^2 - y^2$$



$$\frac{\partial f}{\partial x} = 2x$$

$$\frac{\partial f}{\partial y} = -2y$$

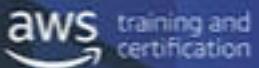
$$f(x, y) = -x^2 - y^2$$



If the matrix is diagonal, a positive entry is a direction where it curves up, and a negative entry is a direction where it curves down.



Example: Not Always So Simple



For

$$f(x, y) = (x - 2y)^2 - (2x + y)^2 + (x - 1)^3$$

we can compute that

$$Hf = \begin{bmatrix} 6x - 12 & -8 \\ -8 & 6 \end{bmatrix}$$



Example: Not Always So Simple



For

$$f(x, y) = (x - 2y)^2 - (2x + y)^2 + (x - 1)^3$$

we can compute that

$$Hf = \begin{bmatrix} 6x - 12 & -8 \\ -8 & 6 \end{bmatrix}$$

*I hope ... your point
not zero off*



An Extra Definition



Trace

$$A = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix}$$

$\text{tr}(A)$ = sum of diagonal terms
= $\sum_i a_{ii}$
= $a_{11} + \dots + a_{nn}$

Classification of Critical Points in 2D



$$\boxed{\nabla f = \vec{0}}$$

• $\det(H_f) < 0$ \Rightarrow Saddle Point

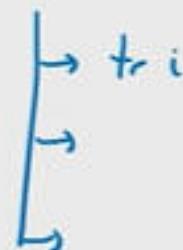
Classification of Critical Points in 2D



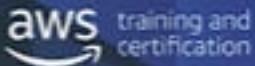
$$\boxed{\nabla f = \vec{0}}$$

• $\det(H_f) < 0 \Rightarrow \text{SADDLE POINT}$

• $\det(H_f) > 0$



Classification of Critical Points in 2D



$$\boxed{\nabla f = 0}$$

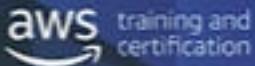
• $\det(H_f) < 0 \Rightarrow \text{Saddle Point}$

• $\det(H_f) > 0$

$$\begin{cases} \rightarrow \text{tr}(H_f) > 0 \\ \rightarrow \text{tr}(H_f) < 0 \\ \rightarrow \text{tr}(H_f) = 0 \end{cases}$$



Classification of Critical Points in 2D



$$\boxed{\nabla f = \vec{0}}$$

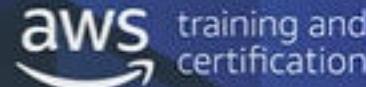
• $\det(H_f) < 0 \Rightarrow \text{Saddle Point}$

• $\det(H_f) > 0$

$$\begin{cases} \rightarrow \text{tr}(H_f) > 0 \\ \rightarrow \text{tr}(H_f) < 0 \\ \rightarrow \text{tr}(H_f) = 0 \end{cases}$$



Classification of Critical Points in 2D



$$\boxed{\nabla f = \vec{0}}$$

• $\det(H_f) < 0 \Rightarrow \text{SADDLE POINT}$

• $\det(H_f) > 0$

$\rightarrow \text{tr}(H_f) > 0 \Rightarrow \text{LOCAL MINIM}$

$\rightarrow \text{tr}(H_f) < 0 \Rightarrow \text{LOCAL MAXIM}$

$\rightarrow \text{tr}(H_f) = 0 \Rightarrow \text{DEG. OR FLAT}$

• $\det(H_f) = 0 \Rightarrow \text{UNCL}$

Example



For

$$f(x, y) = (x - 2y)^2 - (2x + y)^2 + (x - 1)^3$$

we found that

$$Hf = \begin{bmatrix} 6x - 12 & -8 \\ -8 & 6 \end{bmatrix}$$

and thus that $\det(HF) = 4(9x - 39)$.



Example



$$f(x, y) = (x - 2y)^2 - (2x + y)^2 + (x - 1)^3 \Rightarrow Hf = \begin{bmatrix} 6x - 12 & -8 \\ -8 & 6 \end{bmatrix}$$

and thus that $\det(HF) = 4(9x - 39)$.

Some points are extrema and some points are saddles

- No maximum
- Minimum at $\frac{1}{9}(34 + 5\sqrt{43}), \frac{4}{27}(34 + 5\sqrt{43})$
- Saddle point at $\frac{1}{9}(34 - 5\sqrt{43}), \frac{4}{27}(34 - 5\sqrt{43})$





Certificate of Completion
Hem Bahadur Gurung

Has successfully completed
Math for Machine Learning

A handwritten signature in black ink that reads "Maureen Lohrman".

Director, Training and Certification

8 hours

2 September, 2021

Duration

Completion Date