Introduction to Deep Learning (I2DL)

Exercise 2: Math Recap

Overview

1. Linear Algebra

- Vectors, matrices and tensors
- Dot product and matrix multiplication
- Linear transformations and equation systems
- Eigenvalues and vectors
- Norms

2. Calculus

- Scalar derivates and gradients
- Jacobian Matrix
- Chain Rule

3. Probability and Information Theory

- Probability measure
- PMF, PDF, CDF
- Conditional Probability
- Bayes Rule
- Entropy, Cross-Entropy, KL-Divergence

1. Linear Algebra

Overview

Vector

Matrix

Tensor

Linear Transformation

Affine Transformation

Eigenvalues and Eigenvectors

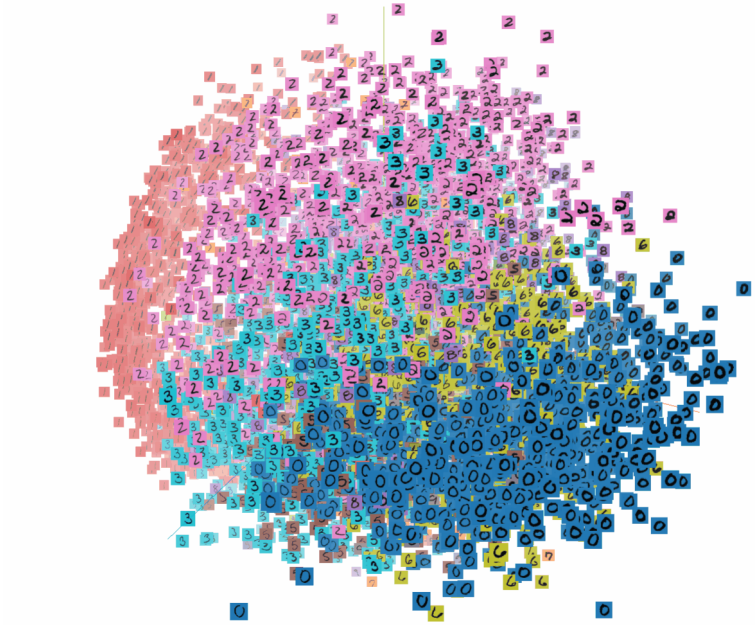
Norm

Vector

A n-dimensional vector describes a point in a n-dimensional space

$$v = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix} \in \mathbb{R}^n$$

Compnents

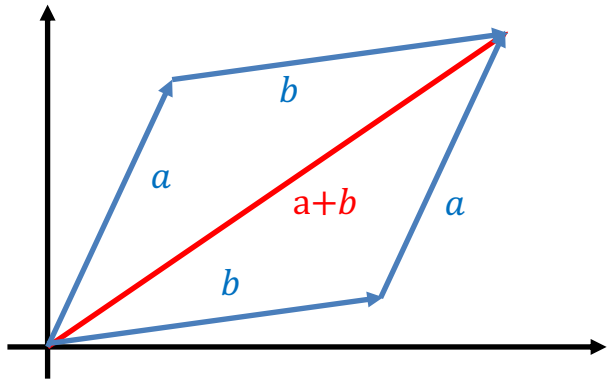


Source: https://projector.tensorflow.org/

Vector

Vector Operations: **Addition** | Subtraction | Scalar Multiplication | Dot Product

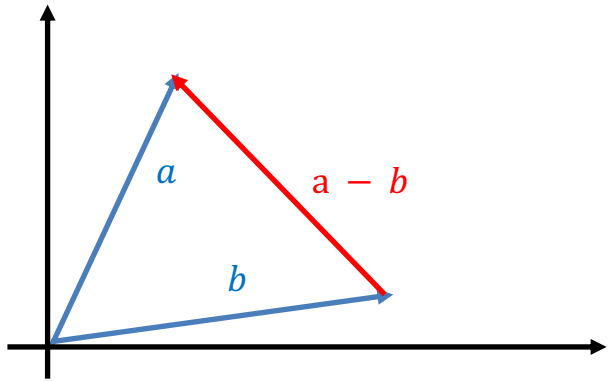
$$a, b \in \mathbb{R}^n, \qquad a + b = \begin{pmatrix} a_1 + b_1 \\ a_2 + b_2 \\ \vdots \\ a_n + b_n \end{pmatrix} \in \mathbb{R}^n$$



Vector

Vector Operations: | Addition | **Subtraction** | Scalar Multiplication | Dot Product |

$$a, b \in \mathbb{R}^n, \qquad a - b = \begin{pmatrix} a_1 - b_1 \\ a_2 - b_2 \\ \vdots \\ a_n - b_n \end{pmatrix} \in \mathbb{R}^n$$

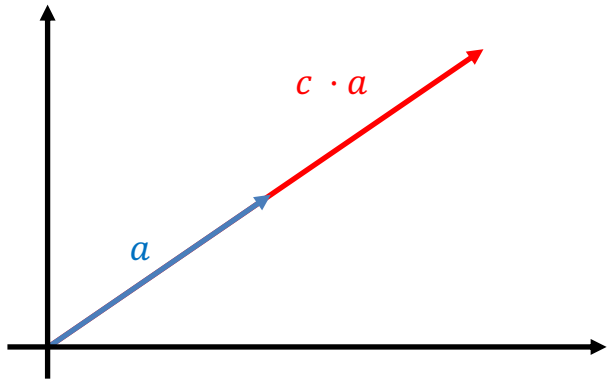


Vector

Vector Operations:

Addition	Subtraction	Scalar Multiplication	Dot Product

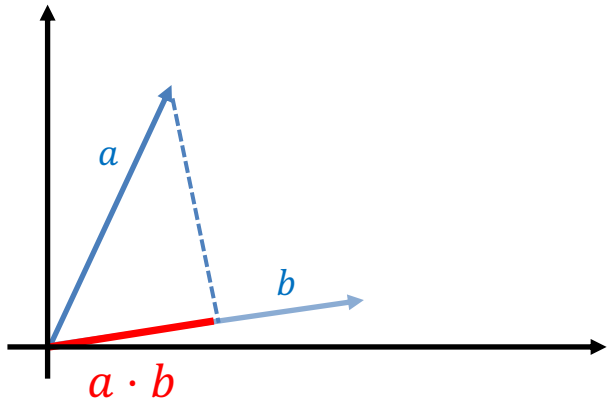
$$a \in \mathbb{R}^n, c \in \mathbb{R}, \qquad ca = \begin{pmatrix} ca_1 \\ ca_2 \\ \vdots \\ ca_n \end{pmatrix} \in \mathbb{R}^n$$



Vector

Vector Operations: | Addition | Subtraction | Scalar Multiplication | **Dot Product**

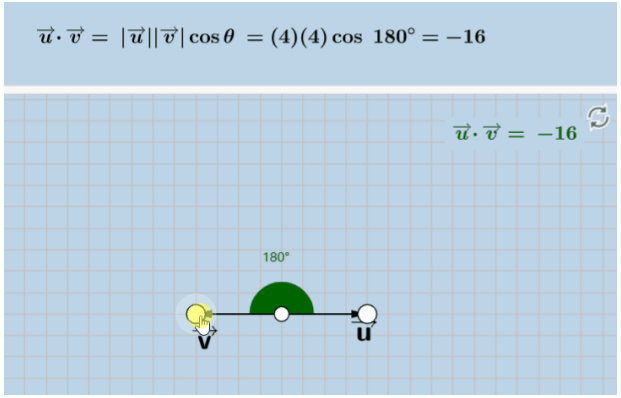
$$a, b \in \mathbb{R}^n, \qquad a \cdot b = \sum_i a_i b_i \in \mathbb{R}$$
$$= a_1 b_1 + a_2 b_2 + \cdots + a_n b_n$$



Vector

Vector Operations: | Addition | Subtraction | Scalar Multiplication | **Dot Product**

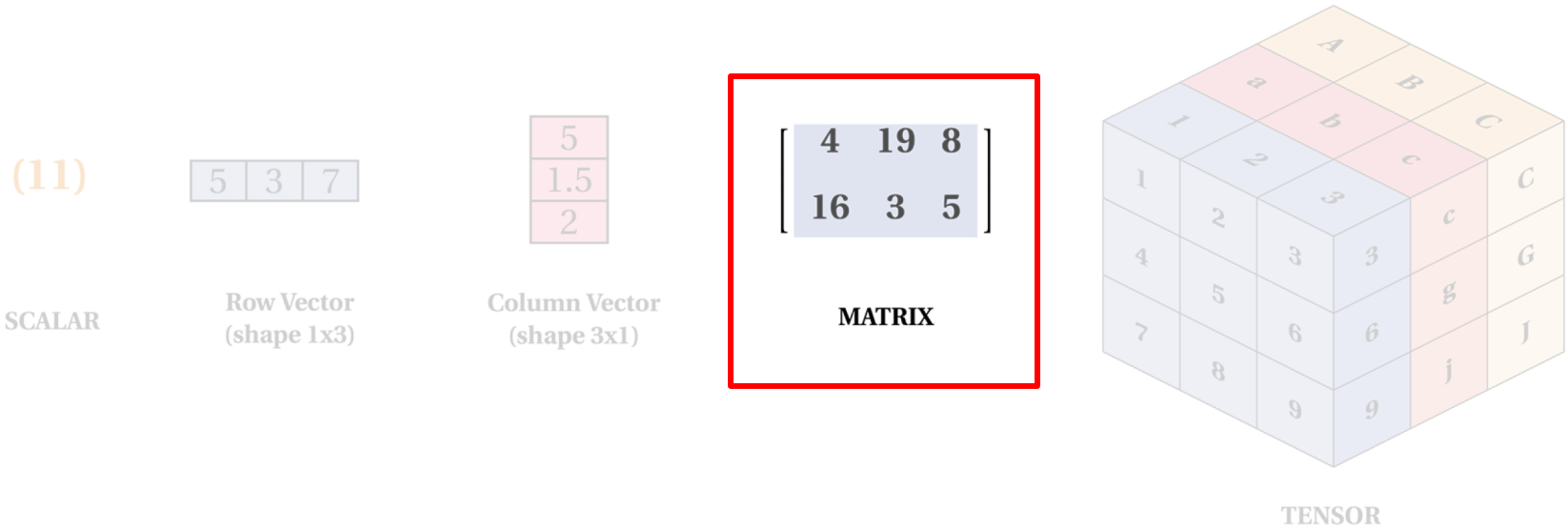
- Properties:
  - Non-negative : $a \cdot a = 0 \Leftrightarrow a = 0$
  - Geometric Interpretation:
    - $a, b \in \mathbb{R}^n, \quad a \cdot b = \|a\| \|b\| \cos \vartheta$
  - Orthogonality: $\theta = 90^\circ \Leftrightarrow a \cdot b = 0$
  - Norm: $\|a\|_2 = \sqrt{a \cdot a}$
  - Metric: $d(a, b) = \|a - b\|_2 = \sqrt{(a - b) \cdot (a - b)}$



Source: https://www.geogebra.org/m/Yu6869By

Matrix

- A matrix lives in the space: $A \in \mathbb{R}^{m \times n}$
- A Matrix has rows and columns (second order tensor)



https://dev.to/mmithrakumar/scalars-vectors-matrices-and-tensors-with-tensorflow-2-0-1f66

Matrix

Vector Operations: **Matrix-Vector Multiplication** | Matrix-Matrix Multiplication | Hadamard Product | Determinant

- Matrix $A \in \mathbb{R}^{m \times n}$, vector $b \in \mathbb{R}^n$
- Operation: $A \cdot b = c \in \mathbb{R}^m$
- Note: $A(m \times n) \cdot b(n \times 1) = c(m \times 1)$ dimensions must match!
- Computation of each element:
  - $c_i = a_{i1}b_1 + a_{i2}b_2 + \cdots + a_{in}b_n = \sum_{k=1}^{n} a_{ik}b_k$

$$A = \begin{pmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{pmatrix} \in \mathbb{R}^{3 \times 2}, b = \begin{pmatrix} 3 \\ 1 \end{pmatrix} \in \mathbb{R}^2, c = Ab = \begin{pmatrix} 7 \\ 11 \\ 1 \end{pmatrix} \in \mathbb{R}^3$$

Matrix

Vector Operations: | Matrix-Vector Multiplication | **Matrix-Matrix Multiplication** | Hadamard Product | Determinant

- Matrix $A \in \mathbb{R}^{m \times n}$, $B \in \mathbb{R}^{n \times p}$
- Matrix-matrix multiplication: $AB = C \in \mathbb{R}^{m \times p}$
- For each element $c_{ij} = a_{i1}b_{1j} + a_{i2}b_{2j} + \cdots + a_{in}b_{nj} = \sum_{k=1}^{n} a_{ik}b_{kj}$

$$\begin{matrix} - \\ + \end{matrix} \begin{bmatrix} 1 & 2 & 1 \\ 0 & 1 & 0 \\ 2 & 3 & 4 \end{bmatrix} \times \begin{bmatrix} 2 & 5 \\ 6 & 7 \\ 1 & 8 \end{bmatrix} \begin{matrix} - \\ + \end{matrix}$$

$$\begin{matrix} - & + \end{matrix} \qquad \begin{matrix} - & + \end{matrix}$$

Source: http://matrixmultiplication.xyz/

Matrix

Vector Operations: | Matrix-Vector Multiplication | Matrix-Matrix Multiplication | **Hadamard Product** | Determinant |

- Matrices $A \in \mathbb{R}^{m \times n}, B \in \mathbb{R}^{m \times n}$
- Hadamard product(Element-wise Product): $A \circ B = C \in \mathbb{R}^{m \times n}$

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

$$C = A \circ B = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \circ \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix} = \begin{pmatrix} a_{11}b_{11} & a_{12}b_{12} \\ a_{21}b_{21} & a_{22}b_{22} \end{pmatrix}$$

Matrix

Vector Operations: Matrix-Vector Multiplication | Matrix-Matrix Multiplication | Hadamard Product | Determinant

- Determinant encodes certain properties of the linear transformation described by the matrix
- Geometrically, it is volumne scaling factor of the linear transformation
- $$\det(A) = \sum_{(m_1, m_2, \ldots, m_n) \in permutation\ n} \big(sign(m_1, \ldots, m_n)\big) A_{m_1,1} \cdots A_{m_n,n}$$

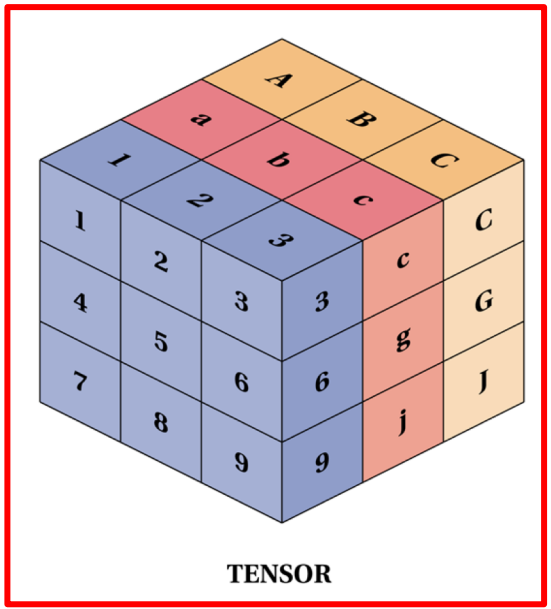
$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathbb{R}^{2\times 2}, \qquad \det(A) = ad - bc$$

Matrix

- Rank: A matrix $A \in \mathbb{R}^{m \times n}$
- Rank: the rank of a matrix is the dimension of the vector space spanned by its columns.
- Rank corresponds to the maximal independent columns of the matrix.
- $0 \leq \text{rank}(A) \leq \min\{m, n\}$

Tensor

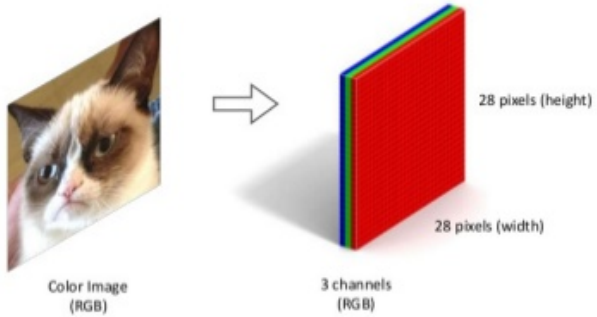
- Definition: a tensor is a generalization of a scalar and a vector.

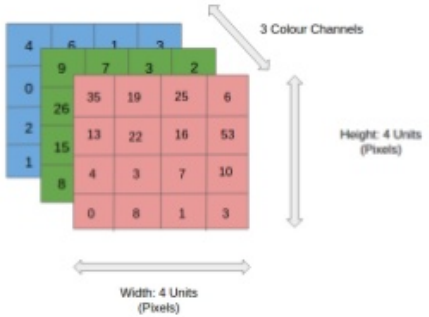


https://dev.to/mmithrakumar/scalars-vectors-matrices-and-tensors-with-tensorflow-2-0-1f66

Tensor

color image is 3rd-order tensor

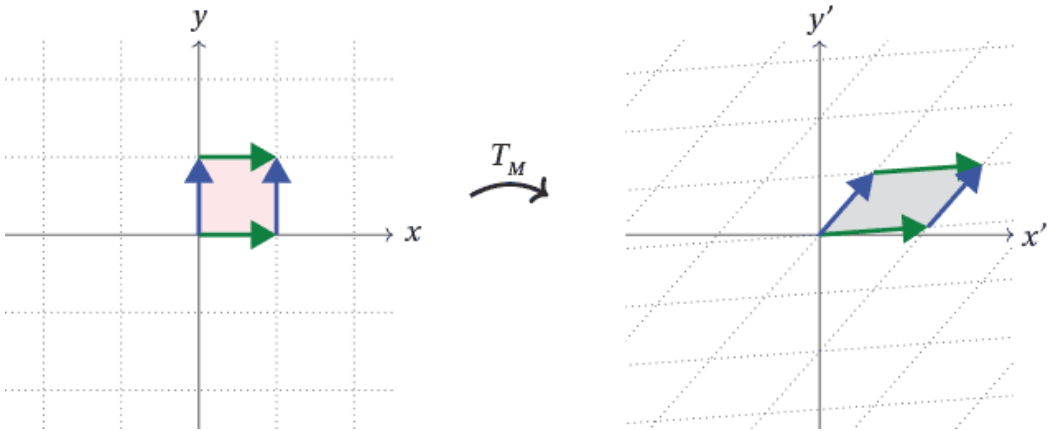




Source: https://www.slideshare.net/BertonEarnshaw/a-brief-survey-of-tensors

Linear Transformation

- A linear transformation T maps v to $T(v)$.
- Linearity $T(cv + dw) = cT(v) + dT(w), \qquad c, d \in \mathbb{R}, \qquad v, w \in \mathbb{R}^n$
- Note: $T(0) = 0$.
  - Transformations with a 'shift' are not linear => affine transformations (see later)



Source:
https://amsi.org.au/ESA_Senior_Years/SeniorTopic8/8a/8a_2content_3.html

Linear Transformation

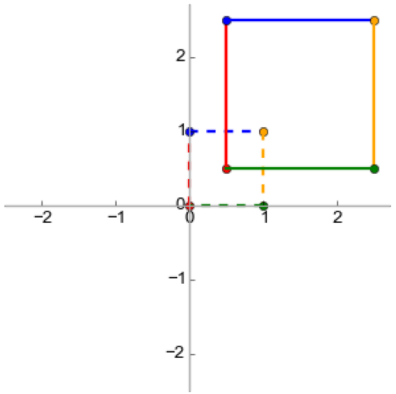
- Linear transformation: $T: \mathbb{R}^n \rightarrow \mathbb{R}^m$, $T(x) = Ax$
  - $A \in \mathbb{R}^{m \times n}, x \in \mathbb{R}^n$
- $x = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, A = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}$
- $Ax = \begin{pmatrix} 3 \\ 3 \end{pmatrix}$
- The combination of two linear transformations is still linear.

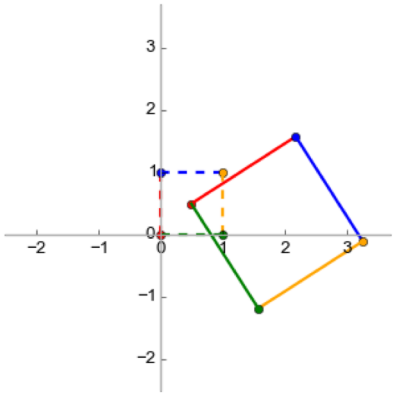
$$B = \begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix}$$

$$B \cdot (Ax) = \begin{pmatrix} 12 \\ 12 \end{pmatrix} \Leftrightarrow (B \cdot A)x = \begin{pmatrix} 5 & 7 \\ 7 & 5 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 12 \\ 12 \end{pmatrix}$$

Affine Transformation

- An affine transformation T maps v to $T(v) = Av + u$
- Affine Layer in Neural Network!
  - matrix A as weight matrix and the vector u as bias





Source: https://eli.thegreenplace.net/2018/affine-transformations/

System of Linear Equations

- Which $\boldsymbol{x}$ do solve the equation $\boldsymbol{Ax} = \boldsymbol{b}$?

System of Linear Equations

- Linear equation in matrix form:

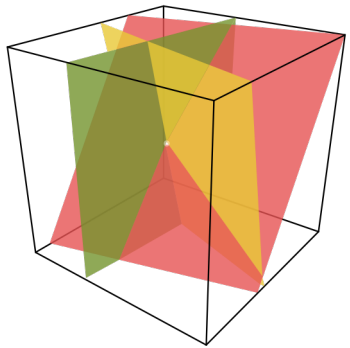
$$\begin{pmatrix} 3 & 2 & -1 \\ 2 & -2 & 4 \\ -1 & 0.5 & -1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ -2 \\ 0 \end{pmatrix}$$

- System of linear equations is of the form:

$$\begin{aligned} 3x + 2y - z &= 1 \\ 2x - 2y + 4z &= -2 \\ -x + 0.5y - z &= 0 \end{aligned}$$

- Solutions in intersect of the three planes:

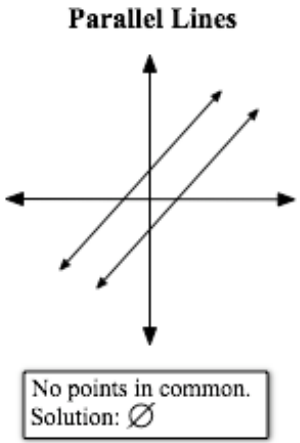
$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ -2 \\ -2 \end{pmatrix}$$

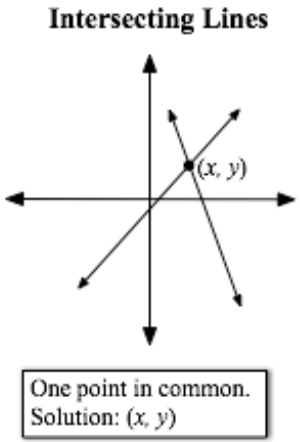


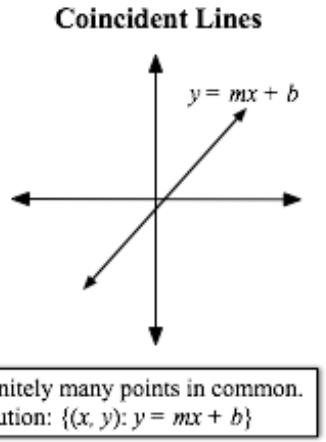
Source: https://en.wikipedia.org/wiki/System_of_linear_equations

Solution of Linear Equations

No unique solution	One unique solution	Infinitely many solutions







Source: http://www.jennifershannonmd.com/WebLectures/8.1-SSLEUG/SSLEUG3.html

Solution of Linear Equations

- Requirement for uniqueness of solution:
  - Matrix $A \in \mathbb{R}^{m \times n}$ is square matrix ($m = n$) and invertible
  - Matrix A has full rank $\Leftrightarrow$ $\det(A) \,! = 0$

$$\begin{pmatrix} 3 & 2 & -1 \\ 2 & -2 & 4 \\ -1 & 0.5 & -1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ -2 \\ 0 \end{pmatrix} \qquad \mathrm{A} = \begin{pmatrix} 3 & 2 & -1 \\ 2 & -2 & 4 \\ -1 & 0.5 & -1 \end{pmatrix}$$

$$\det(A) = -3$$
$$rank(A) = 3$$

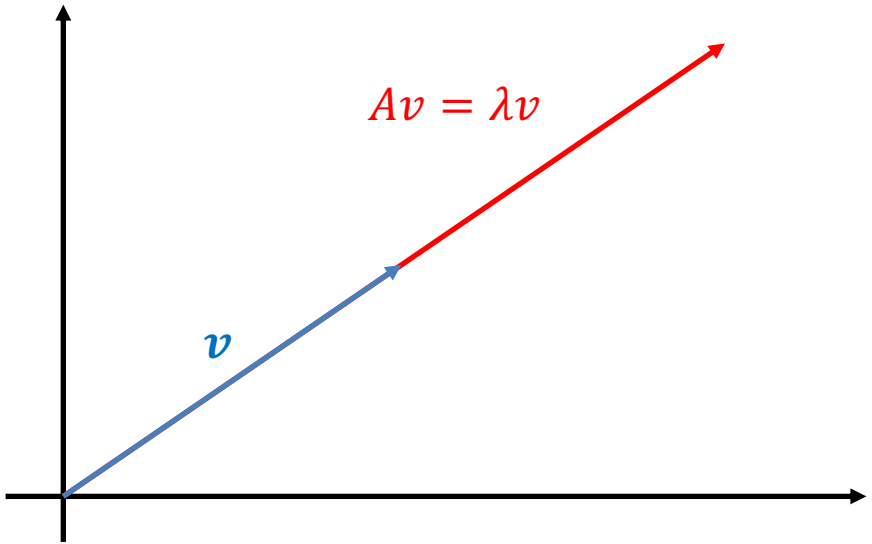
$\Rightarrow$That system has an unique solution

Eigenvalues and Eigenvectors

- Let a vector $v \in \mathbb{C}^n$ for matrix $A \in \mathbb{R}^{n \times n}$, is called eigenvector of A if

$$Av = \lambda v, \qquad v \neq 0$$

For some $\lambda \in \mathbb{C}$. The λ is then called eigenvalue of A.



Eigenvalues and Eigenvectors

- Computing eigenvalues and eigenvectors using characteristic polynomial:

$$p_A(\lambda) = det(\lambda I - A) = 0$$

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

  - Find eigenvalues from the characteristic polynomial:

$$p_A(\lambda) = \begin{vmatrix} \lambda - 1 & 1 \\ 0 & \lambda - 2 \end{vmatrix} = (\lambda - 1)(\lambda - 2) = 0$$

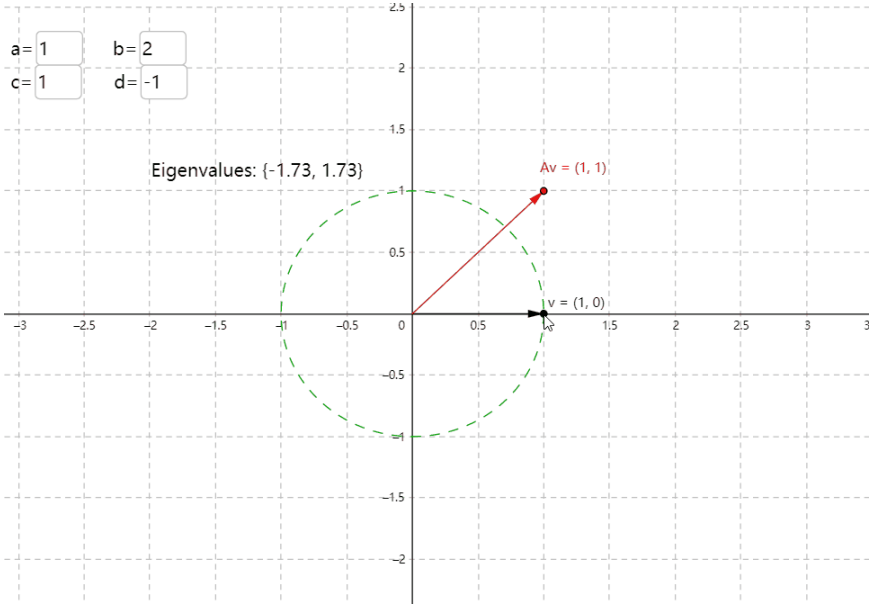
$$\lambda_1 = 1, \lambda_2 = 2$$

  - For every $\boldsymbol{\lambda}$, find its corresponding eigenvector:

$$v_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, v_2 = \begin{pmatrix} -1 \\ 1 \end{pmatrix}$$

Eigenvalues and Eigenvectors

- Geometric interpretation of eigenvalues and eigenvectors:



Source: https://www.geogebra.org/m/JP2XZpzV

Eigenvalues and Eigenvectors

- **Trace** of a matrix is the sum of its eigenvalues

$$\text{tr}(A) = \sum_i \lambda_i$$

- **Determinant** of a matrix is the product of its eigenvalues

$$\det(A) = \prod_i \lambda_i$$

Norms

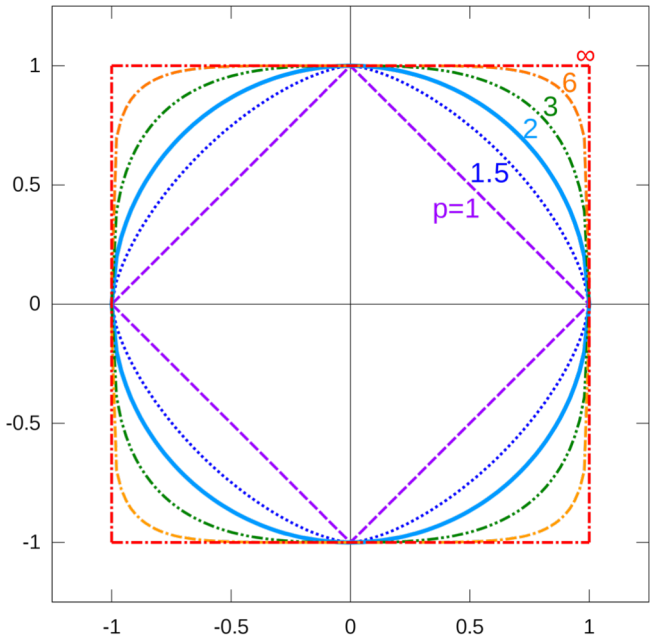
- The **norm** $\|A\|$ is a function that maps A to a real positive number that satisfies certain properties:
  - Positivity: $\|A\| \geq 0, \|A\| = 0 \Leftrightarrow A = 0$
  - Homogeneity: $\|aA\| = |a|\|A\|$
  - Triangle inequality: $\|A + B\| \leq \|A\| + \|B\|$

Norms

- For vectors: p-norm

$$x \in \mathbb{R}^n,$$

$$\|x\|_p = \left(\sum_i |x_i|^p \right)^{\frac{1}{P}}$$



Source: https://en.wikipedia.org/wiki/Lp_space

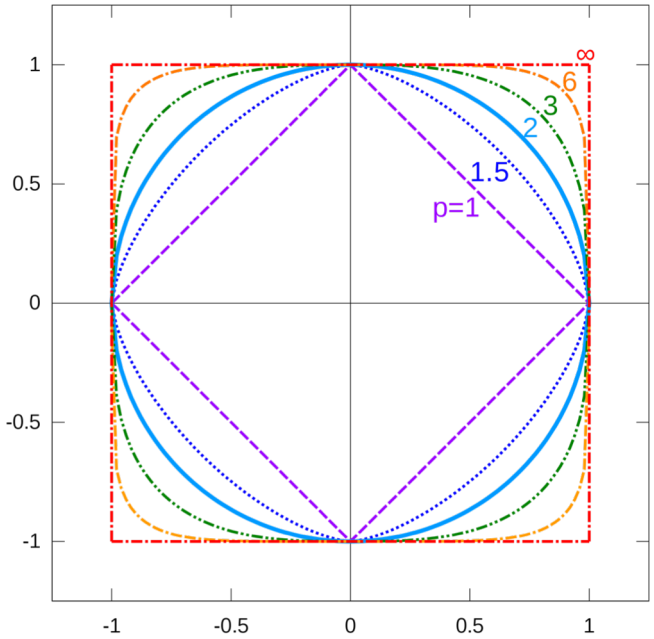
Norms

- L1:

$$x \in \mathbb{R}^n, \|x\|_1 = \sum_i |x_i|$$

$$x = \begin{pmatrix} 1 \\ 3 \\ 2 \end{pmatrix} \in \mathbb{R}^3,$$

$$\|x\|_1 = |1| + |3| + |2| = 6$$



Source: https://en.wikipedia.org/wiki/Lp_space

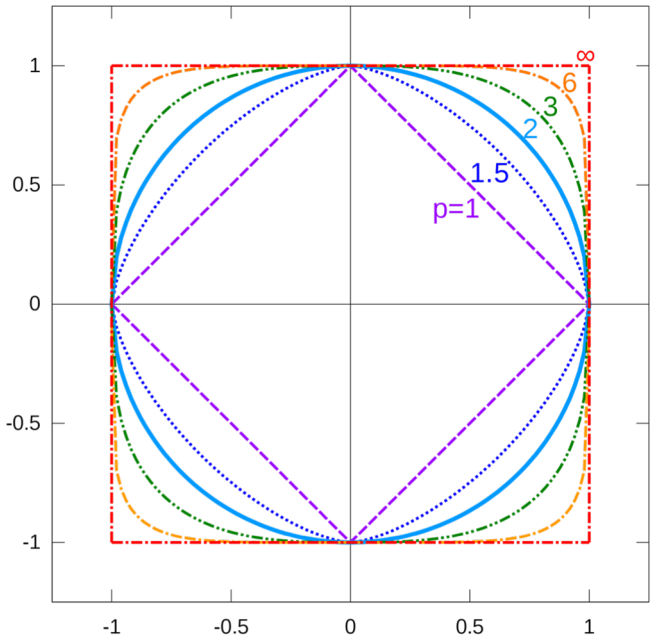
Norms

- L2: Euclidean norm

$$x \in \mathbb{R}^n, \|x\|_2 = \left(\sum_i \left(x_i^2\right)\right)^{\frac{1}{2}}$$

$$x = \begin{pmatrix} 1 \\ 3 \\ 2 \end{pmatrix} \in \mathbb{R}^3,$$

$$\|x\|_2 = \sqrt{1^2 + 3^2 + 2^2} = \sqrt{14}$$



Source: https://en.wikipedia.org/wiki/Lp_space

Norms

- Matrix Norm:
  - Frobenius Norm: $\|A\|_F = \sqrt{\mathrm{tr}(A^T A)} = \sqrt{\sum_{i=1}^{r} \lambda_i}$

$$A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 2 \end{pmatrix},$$

$$\|A\|_F = \sqrt{1 + 3 + 2} = \sqrt{6}$$

References

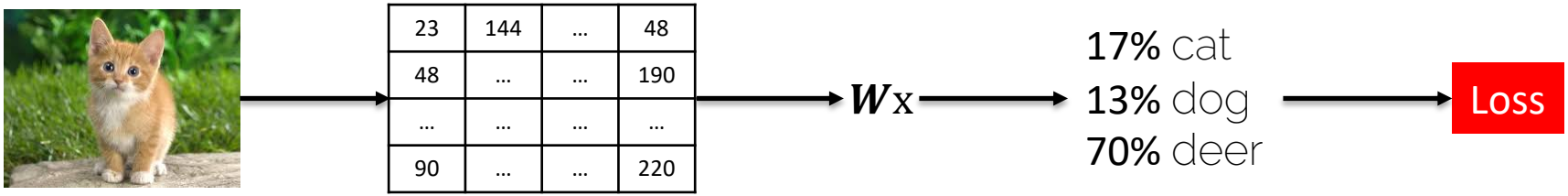
- https://math.mit.edu/~gs/linearalgebra/
- Gilbert Strang "Introduction to Linear Algebra"– **recommended!**

2. Calculus

Overview

Critical Points	Scalar derivative	Gradient	Jacobian matrix
Single variable chain rule	Total derivative chain rule		

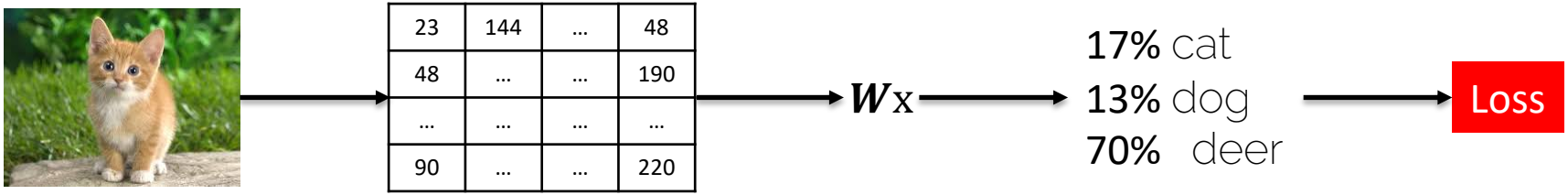
Why do we need calculus in DL?



How can we get an accurate $\boldsymbol{W}$ and decrease the Loss?

(image from: http://www.isfpga.org/fpga2017/slides/D2_S1_02.pdf)

Why do we need calculus in DL?



Adjusting weights of neural network by gradient descent

(image from: http://www.isfpga.org/fpga2017/slides/D2_S1_02.pdf)

Gradient Descent

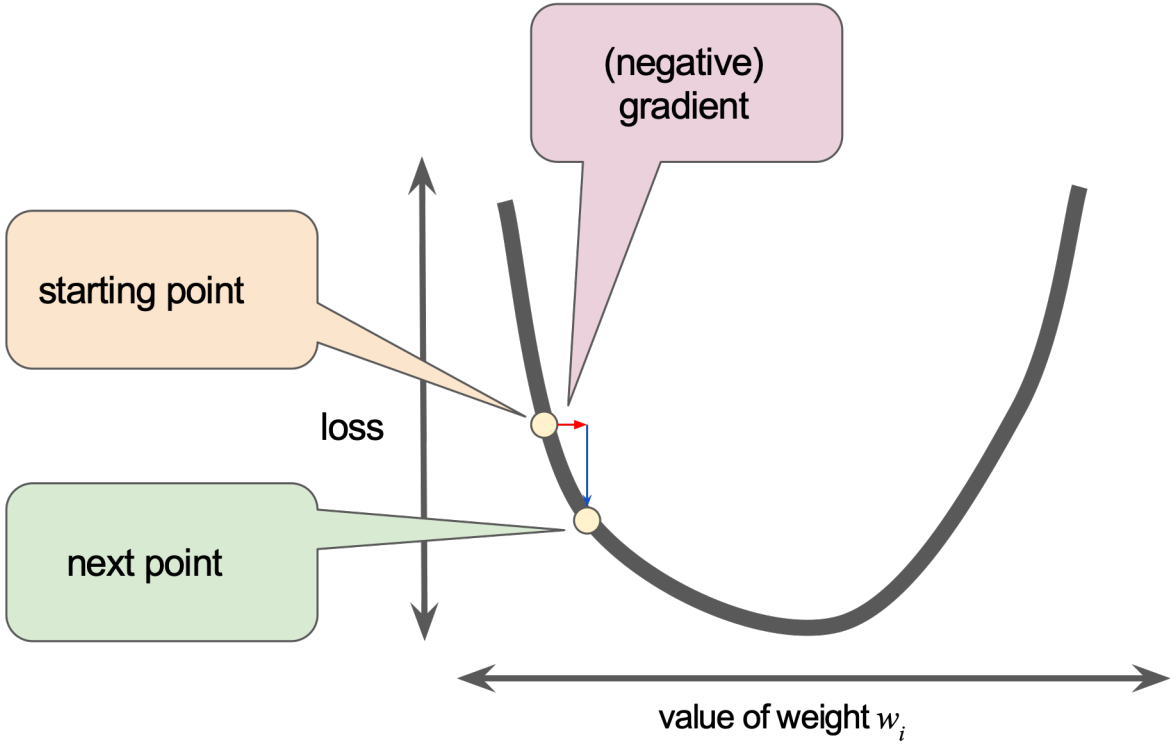
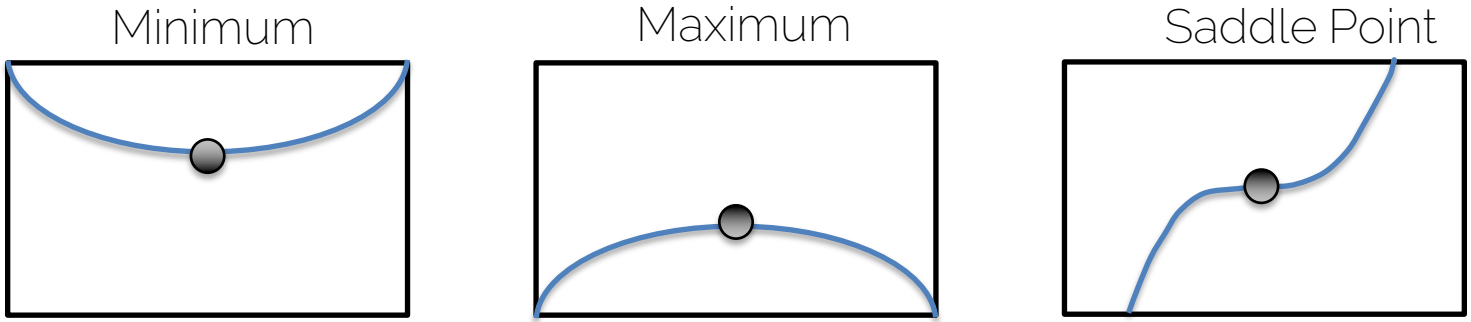


Image from: https://developers.google.com/machine-learning/crash-course/reducing-loss/gradient-descent

Critical Points



	Minimum	Maximum	Saddle Point
$\frac{df}{dx^-}$	< 0	> 0	>0 (or < 0)
$\frac{df}{dx}$	= 0	= 0	= 0
$\frac{df}{dx^+}$	> 0	< 0	>0 (or < 0)

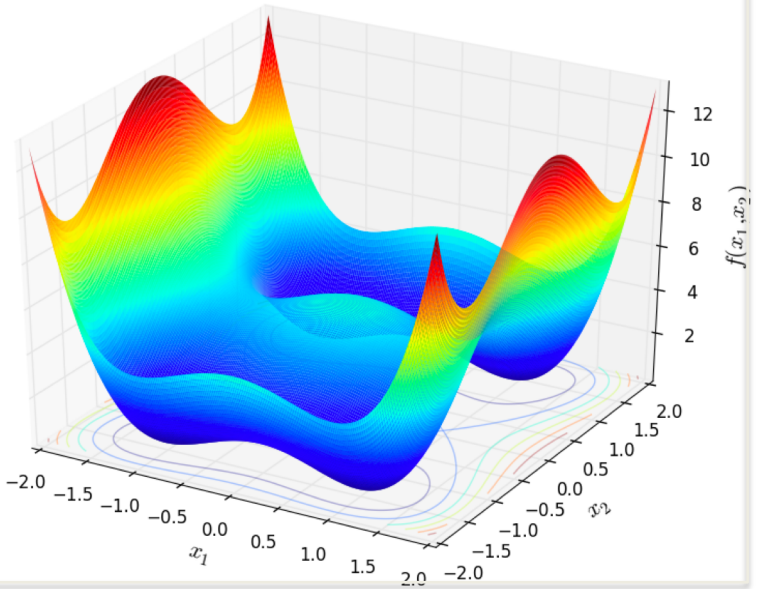
Scalar derivative

Rule	$f(x)$	Scalar derivative notation with respect to x
Constant	c	0
Multiplication by constant	cf	$c\frac{df}{dx}$
Power Rule	x^n	nx^{n-1}
Sum Rule	$f+g$	$\frac{df}{\mathrm{d}x}+\frac{dg}{\mathrm{d}x}$
Difference Rule	$f-g$	$\frac{df}{dx}-\frac{dg}{dx}$
Product Rule	fg	$f\frac{dg}{dx}+\frac{df}{dx}g$
Chain Rule	$f(g(x))$	$\frac{df(u)}{du}\frac{du}{dx}$, let $u=g(x)$

Multivariate Functions

Multivariate Function

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



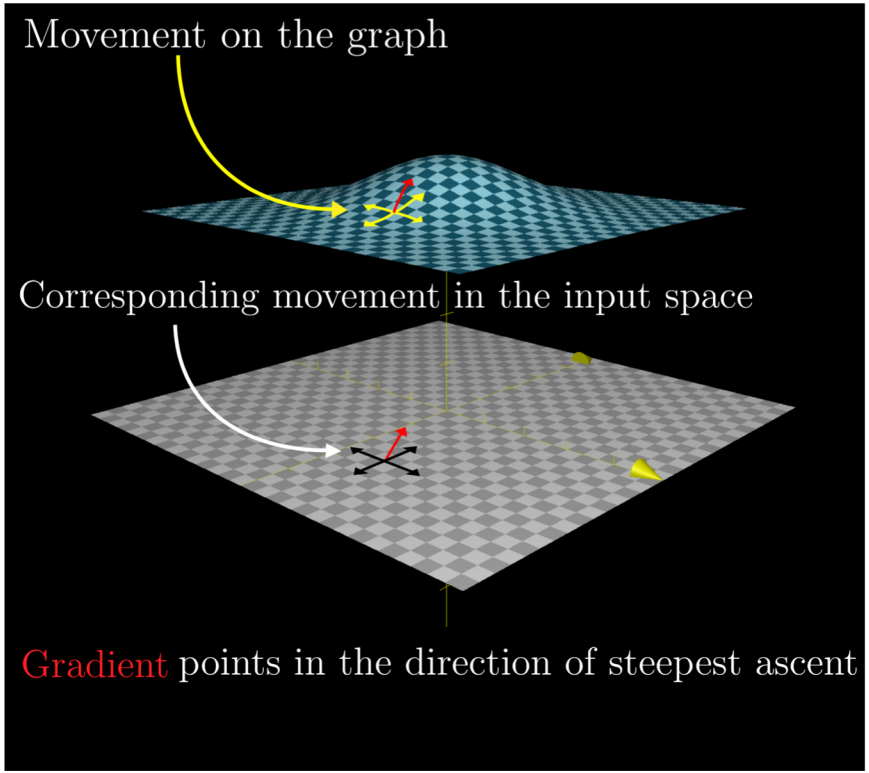
Gradient

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

Partial derivative

$$\mathbf{x} \rightarrow \nabla f(\mathbf{x}) = \begin{pmatrix} \frac{\partial f(\mathbf{x})}{\partial x_1} \\ \frac{\partial f(\mathbf{x})}{\partial x_2} \\ \dots \\ \frac{\partial f(\mathbf{x})}{\partial x_n} \end{pmatrix}$$

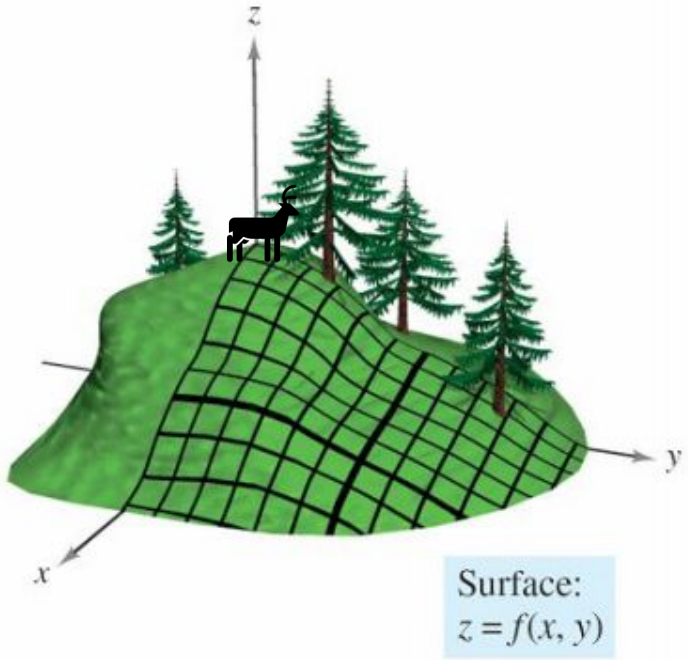
Geometric interpretation of gradient



Images from: https://www.khanacademy.org/math/multivariable-calculus/multivariable-derivatives/partial-derivative-and-gradient-articles/a/the-gradient

Gradient

Gradient – Example 1



$$f(x,y) = 3x^2y \qquad \nabla f(x,y) = \left[\frac{\partial f(x,y)}{\partial x}, \frac{\partial f(x,y)}{\partial y}\right]$$

$$\frac{\partial}{\partial x} 3yx^2 = 3y\frac{\partial}{\partial x}x^2 = 3y2x = 6yx$$

$$\frac{\partial}{\partial y} 3x^2y = 3x^2\frac{\partial}{\partial y}y = 3x^2\frac{\partial y}{\partial y} = 3x^2 \times 1 = 3x^2$$

$$\nabla f(x,y) = \left[\frac{\partial f(x,y)}{\partial x}, \frac{\partial f(x,y)}{\partial y}\right] = [6yx, 3x^2]$$

Images from: https://www.zhihu.com/question/36301367

Gradient – Example 2

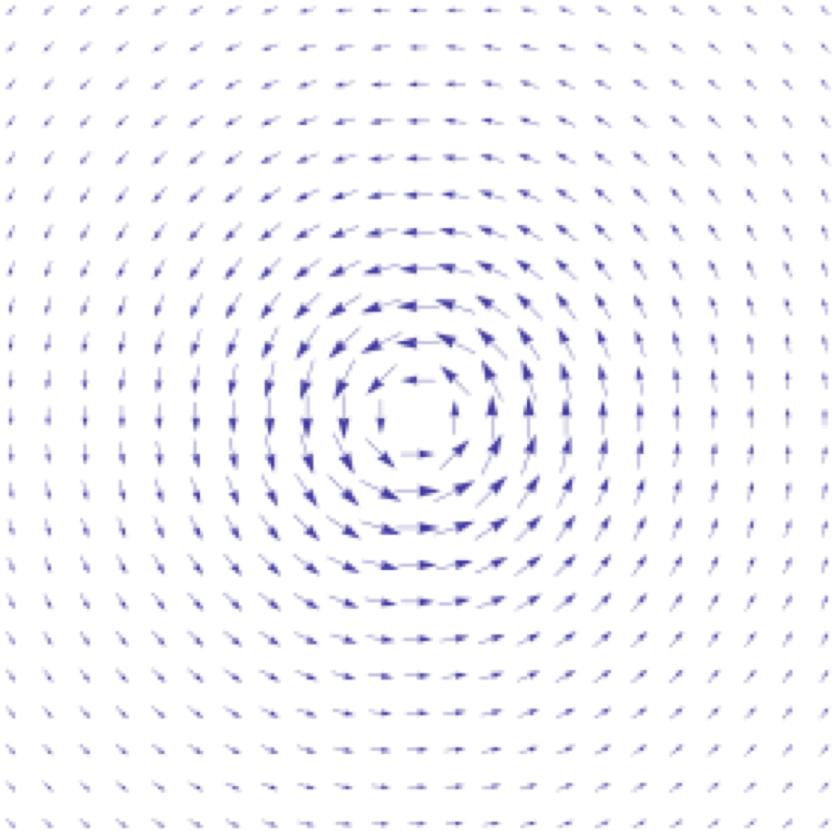
$$g(x, y) = 2x + y^8$$

$$\frac{\partial g(x,y)}{\partial x} = \frac{\partial 2x}{\partial x} + \frac{\partial y^8}{\partial x} = 2\frac{\partial x}{\partial x} + 0 = 2 \times 1 = 2$$

$$\frac{\partial g(x,y)}{\partial y} = \frac{\partial 2x}{\partial y} + \frac{\partial y^8}{\partial \mathrm{y}} = 0 + 8y^7 = 8y^7$$

$$\nabla g(x, y) = [2, 8y^7]$$

Vector-Valued Functions



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

Jacobian Matrix

Vector-Valued function

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

Jacobian Matrix

$$\mathbf{J}_f: \mathbb{R}^n \rightarrow \mathbb{R}^m$$

$$\mathbf{x} \rightarrow \mathbf{J}_f(\mathbf{x}) = \begin{pmatrix} \frac{\partial f_1(\mathbf{x})}{\partial x_1} & \cdots & \frac{\partial f_1(\mathbf{x})}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m(\mathbf{x})}{\partial x_1} & \cdots & \frac{\partial f_m(\mathbf{x})}{\partial x_n} \end{pmatrix}$$

Jacobian Matrix – Example 3

$$f(x, y) = 3x^2 y$$

$$g(x, y) = 2x + y^8$$

Calculate Jacobian Matrix:

$$\boldsymbol{J} = \begin{bmatrix} \nabla f(x,y) \\ \nabla g(x,y) \end{bmatrix} = \begin{bmatrix} \frac{\partial f(x,y)}{\partial x} & \frac{\partial f(x,y)}{\partial y} \\ \frac{\partial g(x,y)}{\partial x} & \frac{\partial g(x,y)}{\partial y} \end{bmatrix} = \begin{bmatrix} 6yx & 3x^2 \\ 2 & 8y^7 \end{bmatrix}$$

Summary

		Variable: Scalar x	Variable: Vector $\mathbf{x} \in \mathbb{R}^n$
Function	Scalar f	Derivative $\frac{\partial f}{\partial x}$	Gradient $\left[\frac{\partial f}{\partial x_1}, \dots, \frac{\partial f}{\partial x_n}\right]^{\mathrm{T}}$
Function	Vector $\mathbf{f} \in \mathbb{R}^m$	Partial Derivative $\left[\frac{\partial f_1(x)}{\partial x}, \dots, \frac{\partial f_m(x)}{\partial x}\right]^{\mathrm{T}}$	Jacobian Matrix $\begin{bmatrix} \frac{\partial}{\partial x_1} f_1(\mathbf{x}) & \dots & \frac{\partial}{\partial x_n} f_1(\mathbf{x}) \\ & \dots & \\ \frac{\partial}{\partial x_1} f_m(\mathbf{x}) & \dots & \frac{\partial}{\partial x_n} f_m(\mathbf{x}) \end{bmatrix}$

Single Variable Chain Rule

1. Introduce the intermidiate variable

$$y = f(g(x)) \qquad u = g(x)$$

2. Compute derivatives.

$$\frac{dy}{du} \qquad \frac{du}{dx}$$

3. Combine

$$\frac{dy}{dx} = \frac{dy}{du}\frac{du}{dx}$$

4. Substitute intermediate variables back

Single Variable Chain Rule – Example 4

$$y = f\big(g(x)\big) = \sin(x^2)$$

1. Introduce the intermidiate variable

$u = x^2$ (Relative to definition $f\big(g(x)\big), g(x) = x^2$)

$y = \sin(u)$ $(y = f(u) = \sin(u))$

2. Compute derivatives.

$\frac{du}{dx} = 2x$ (Take derivative with respect to $\boldsymbol{x}$)

$\frac{dy}{du} = \cos(u)$ (Take derivative with respect to $\boldsymbol{u}$ not $\boldsymbol{x}$)

Single Variable Chain Rule – Example 4

$$y = f\big(g(x)\big) = \sin(x^2)$$

3. Combine.

$$\frac{du}{dx} = 2x \qquad \text{(Take derivative with respect to } \boldsymbol{x}\text{)}$$

$$\frac{dy}{du} = \cos(u) \qquad \text{(Take derivative with respect to } \boldsymbol{u} \text{ not } \boldsymbol{x}\text{)}$$

$$\frac{dy}{dx} = \frac{dy}{du}\frac{du}{dx} = \cos(u)\, 2x$$

4. Substitute.

$$\frac{dy}{dx} = \frac{dy}{du}\frac{du}{dx} = \cos(x^2)\, 2x = 2xcos(x^2)$$

Total Derivative Chain Rule

- General Formalism:

$$\frac{\partial f(x, u_1, \dots, u_n)}{\partial x} = \frac{\partial f}{\partial x} + \frac{\partial f}{\partial u_1}\frac{\partial u_1}{\partial x} + \frac{\partial f}{\partial u_2}\frac{\partial u_2}{\partial x} + \cdots + \frac{\partial f}{\partial u_n}\frac{\partial u_n}{\partial x} = \frac{\partial f}{\partial x} + \sum_{i=1}^{n} \frac{\partial f}{\partial u_i}\frac{\partial u_i}{\partial x}$$

References

- https://en.wikipedia.org/wiki/Matrix_calculus
- http://parrt.cs.usfca.edu/doc/matrix-calculus/index.html
- https://arxiv.org/pdf/1802.01528.pdf
- https://www.khanacademy.org/math/multivariable-calculus/multivariable-derivatives
- https://explained.ai/matrix-calculus/
- http://www.deeplearningbook.org/contents/part_basics.html
- https://towardsdatascience.com/calculating-gradient-descent-manually-6d9bee09aa0b

3 Probability Theory and Information Theory

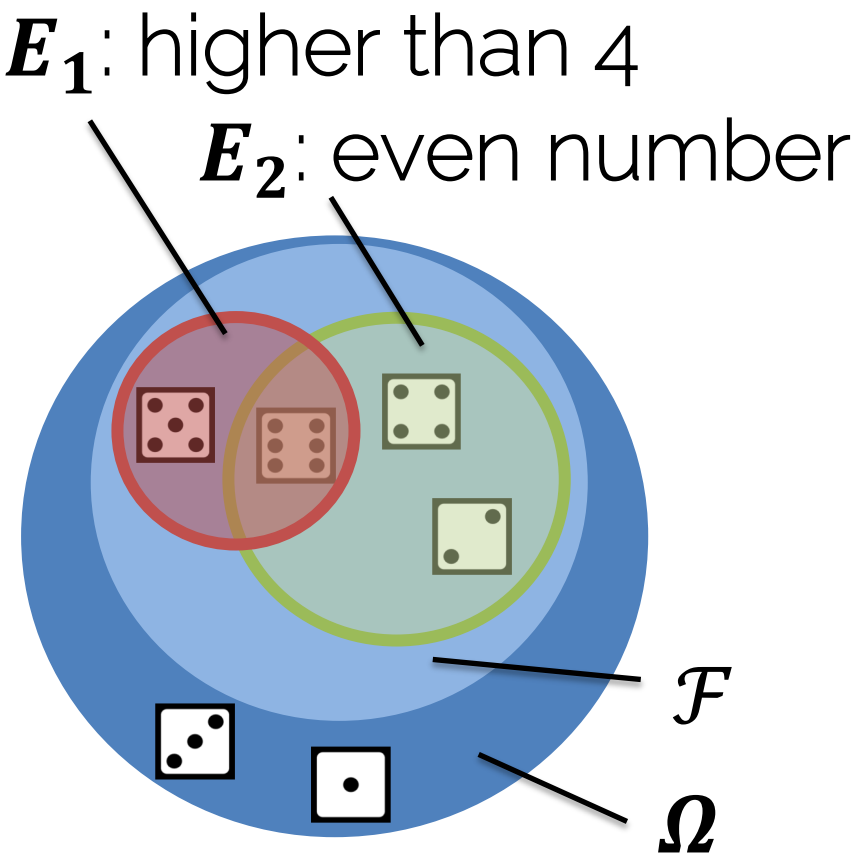
Overview

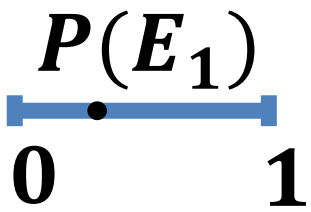
Elements of Probability	Independence	Conditional Probabilities	PMF, CDF, PDF
Moments	Common Probability Distributions	Performance Measures	Quick Information Theory Recap

3a Probability Theory

Elements of Probability

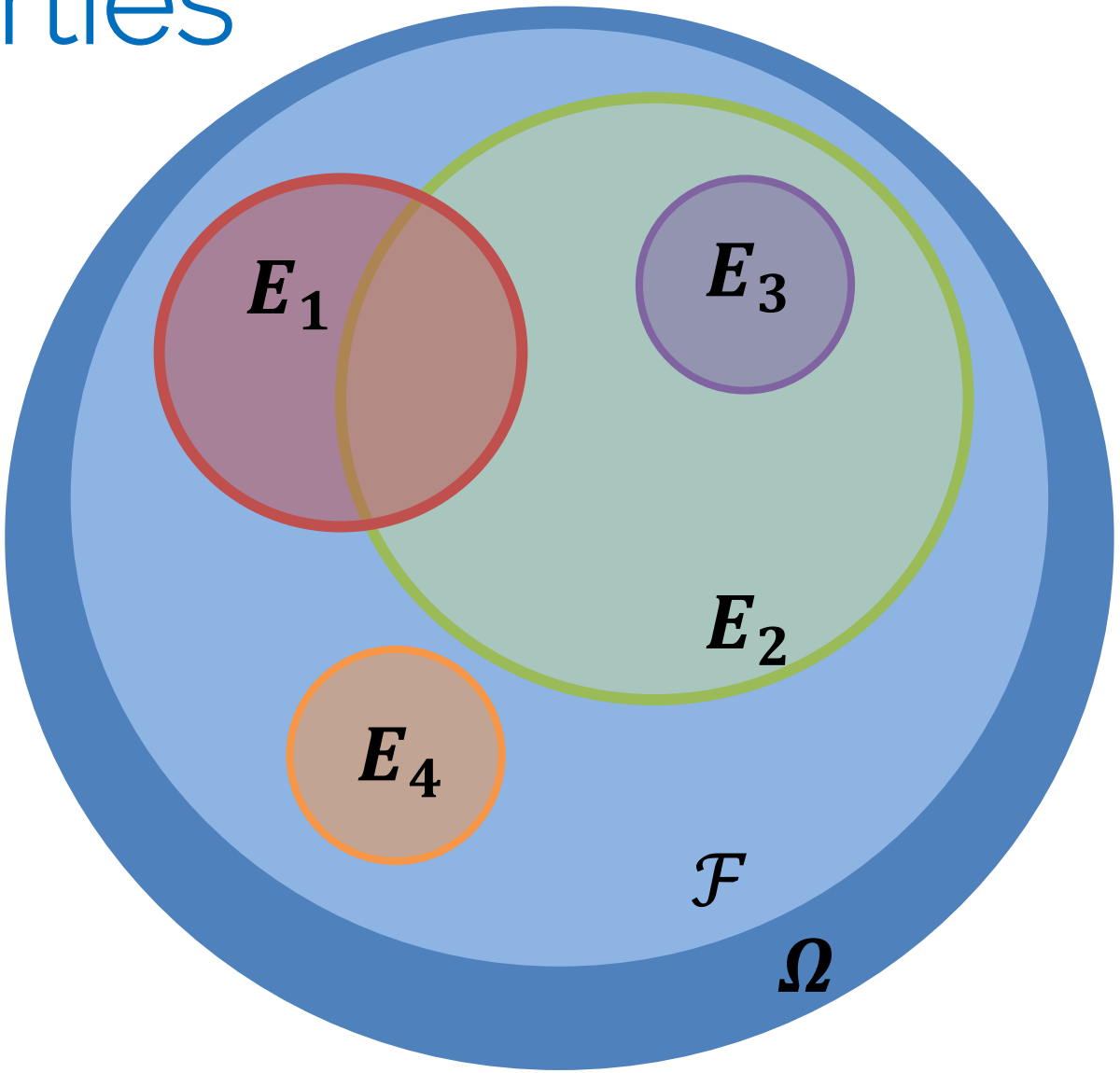
- Sample Space $\boldsymbol{\Omega}$: the set of all outcomes of a random experiment
- Event Space $\mathcal{F}$: set of events $E \in \mathcal{F}$, which are subsets of $\boldsymbol{\Omega}$ i.e. $E \subseteq \boldsymbol{\Omega}$
- Probability Measure: A function $P: \mathcal{F} \rightarrow \mathbb{R}$ that satisfies the following properties:
  - $0 \leq P(E) \leq 1 \; \forall \; E \in \mathcal{F}$
  - $P(\Omega) = 1$
  - $P(\bigcup_{i=1}^{n} E_i) = \sum_{i=1}^{n} P(E_i)$ for disjoint events $E_1, \ldots, E_n$



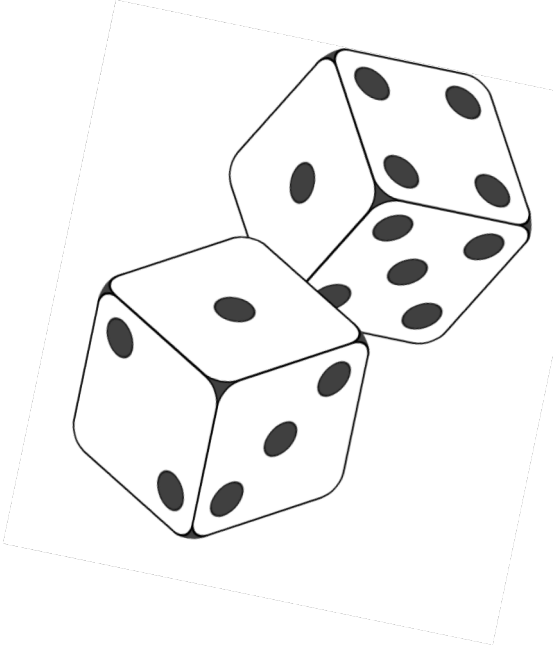


Properties

- $A \subseteq B \Rightarrow P(A) \leq P(B)$
- $P(A \cap B) \leq \min\big(P(A), P(B)\big)$
- $P(A \cup B) \leq P(A) + P(B)$
- $P(\Omega \backslash \mathrm{A}) = 1 - P(A)$



Example: Rolling 2 Dice



Sample Space

$$\Omega = \{(1,1), (1,2), (1,3), \dots, (6,5), (6,6)\}$$

Event 1: "first number is even"

$$E_1 = \{(2,1), (2,2), (2,3), \dots (4,1), (4,2), \dots (6,5), (6,6)\}$$

Event 2: "the 2nd number is a five"

$$E_2 = \{(1,5), (2,5), (3,5), (4,5), (5,5), (6,5)\}$$

-> Event Space $\mathcal{F} = \{E_1, E_2\}$

[image: https://upload.wikimedia.org/wikipedia/commons/c/c4/2-Dice-Icon.svg]

Independence

Two events A and B are independent if and only if:

$$P(A \cap B) = P(A)P(B)$$

Two events A and B are conditionally independent given a third event C if and only if:

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

Independence - Example

Annual income	University A	University B	TOTAL
Under $20,000	36	24	60
$20,000 to 39,999	109	56	165
$40,000 and over	35	40	75
TOTAL	180	120	300

a) Are the events "income is $40,000 and over" and "attended University B" independent?
b) Are the events "income is under $20,000" and "attended University B" independent?

[Source: https://www.khanacademy.org/math/ap-statistics/probability-ap/stats-conditional-probability/a/check-independence-conditional-probability]

Independence - Example

Annual income	University A	University B	TOTAL
Under $20,000	36	24	60
$20,000 to 39,999	109	56	165
$40,000 and over	35	40	75
TOTAL	180	120	300

a)

P(>40K) = 75/300 = 0.25

P(Uni B) = 120 / 300 = 0.4

= 0.1

P(>40K ∩ Uni B) = 40/120 = 1/3

-> not independent!

[Source: https://www.khanacademy.org/math/ap-statistics/probability-ap/stats-conditional-probability/a/check-independence-conditional-probability]

Independence - Example

Annual income	University A	University B	TOTAL
Under $20,000	36	24	60
$20,000 to 39,999	109	56	165
$40,000 and over	35	40	75
TOTAL	180	120	300

b)

P(<20K) = 60/300 = 0.2

P(Uni B) = 120 / 300 = 0.4

= 0.08

P(<20K ∩ Uni B) = 24/300 = 0.08

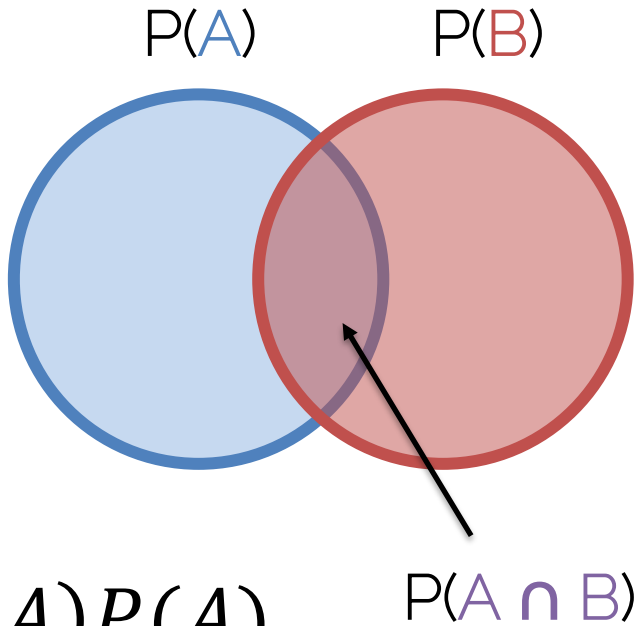
-> independent!

[Source: https://www.khanacademy.org/math/ap-statistics/probability-ap/stats-conditional-probability/a/check-independence-conditional-probability]

Conditional Probabilities

Let B be an event with non-zero probability. The conditional probability of an event A given B is defined as

$$P(A|B) \widehat{=} \frac{P(A \cap B)}{P(B)}$$



This leads to the chain rule:

$$P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$$

Bayes' Rule

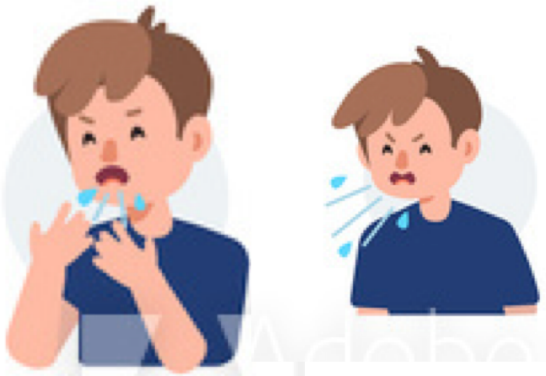
Using the previous results, we can derive Bayes' Rule:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Extended Form:

$$P(A_k|B) = \frac{P(B|A_k)P(A_k)}{\sum_{i=1}^{n} P(B|A_i)P(A_i)}$$

Example: COVID-20



Population infected:
1%

Correct Diagnosis:
95%

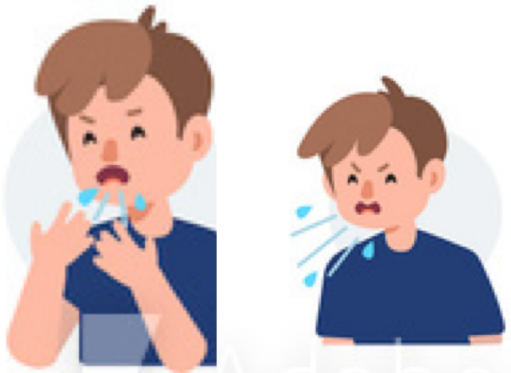
?

You've been chosen at random to get tested.

a) What is the probability that the test is positive?
b) If so, what is the probability you have the disease?

[images: Freepik.com]

Example: COVID-20



Population infected:
1%

Correct Diagnosis:
95%

?

a) P(T)
b) P(D|T)

a) $p(D) = 0.01$
$p(T|D) = 0.95, p(\neg T|\neg D) = 0.95$

$$p(T) = p(T, D) + p(T, \neg D)$$
$$= p(T|D)p(D) + p(T|\neg D)p(\neg D)$$
$$= 0.95 * 0.01 + 0.05 * 0.99 = 5.9\%$$

b) $$p(D|T) = \frac{p(T|D) * p(D)}{p(T)}$$

$$= \frac{0.95 * 0.01}{0.059}$$

$$= 16.1\%$$

[images: Freepik.com]

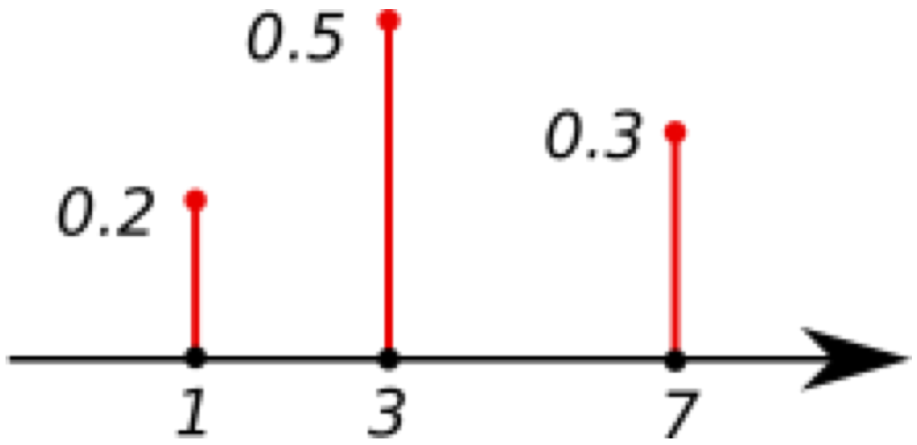
Probability Mass Function

A probability mass function is the probability distribution of a discrete random variable and provides the possible values and their associated probabilities. It is the function $p: \mathbb{R} \rightarrow [0,1]$ defined by

$$p_X(x_i) = P(X = x_i)$$

Properties:

- $0 \leq p_X(x) \leq 1$
- $\sum_x p_X(x) = 1$



A sample PMF

[Figure: https://en.wikipedia.org/wiki/Probability_mass_function]

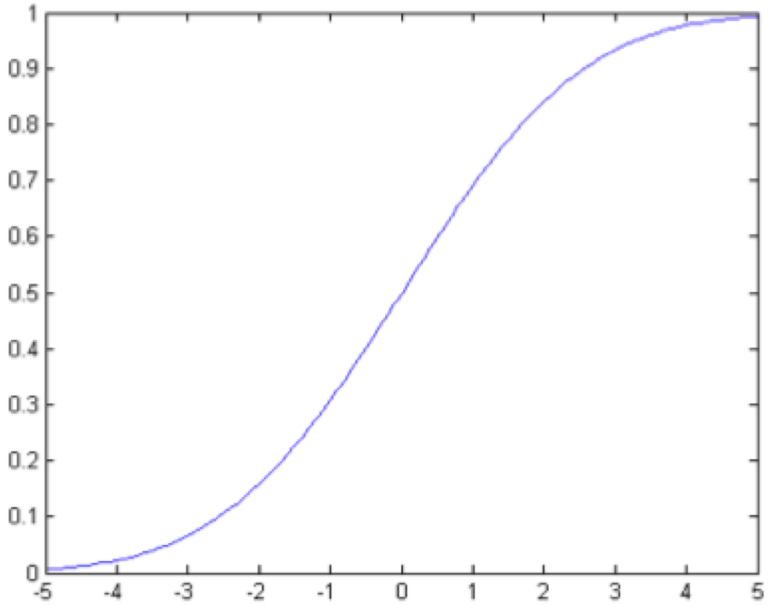
Cumulative Distribution Function

A cumulative distribution function (CDF) of a random variable X is a function $F_X: \mathbb{R} \rightarrow [0,1]$

$$F_X(x) = P(X \leq x)$$

Properties:

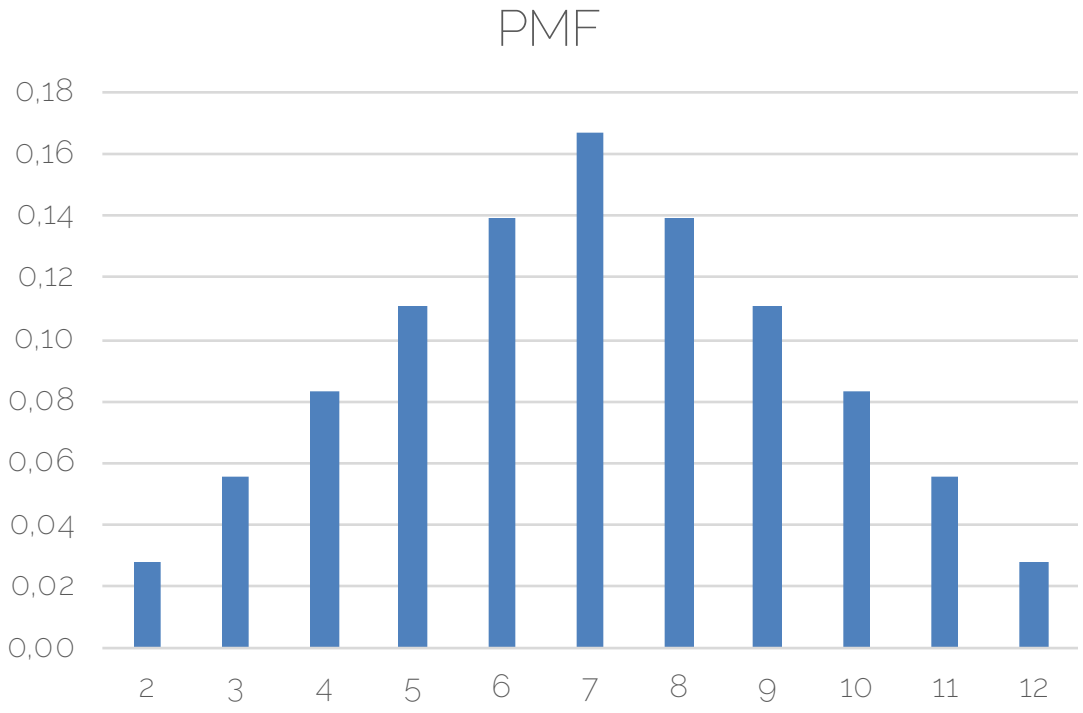
- $0 \leq F_X(x) \leq 1$
- $\lim_{x \rightarrow -\infty} F_x(x) = 0$
- $\lim_{x \rightarrow \infty} F_X(x) = 1$
- $x \leq y \rightarrow F_x(x) \leq F_x(y)$

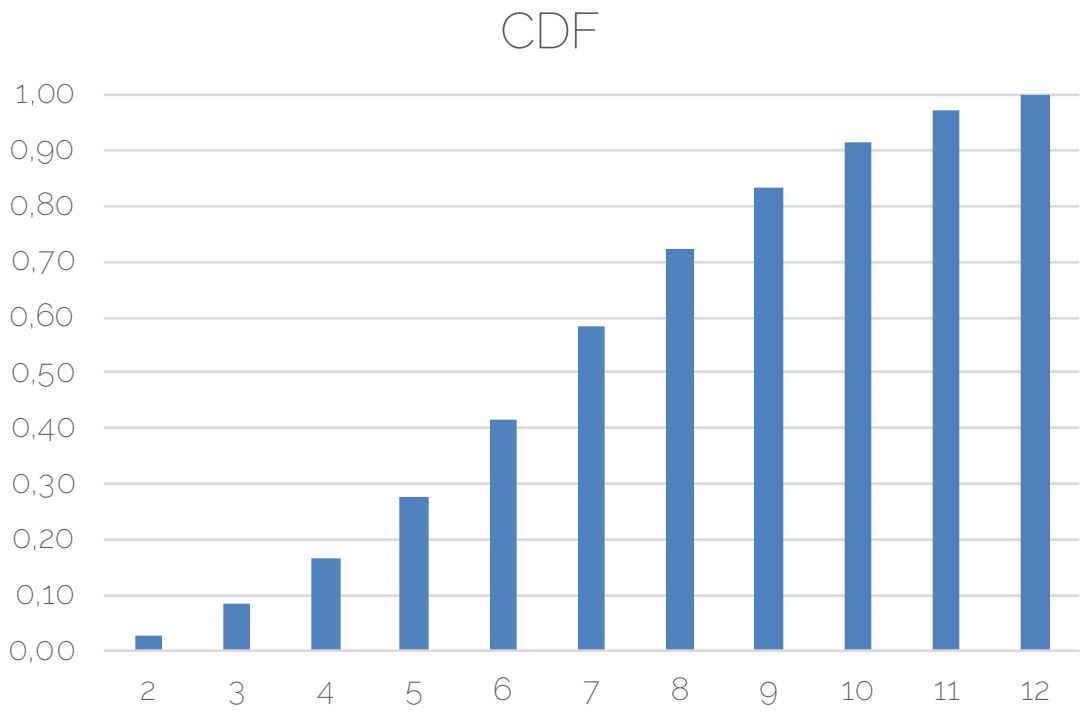


A sample CDF

[Figure: http://cs229.stanford.edu/section/cs229-prob.pdf]

Discrete Example: Sum of 2 Dice Rolls





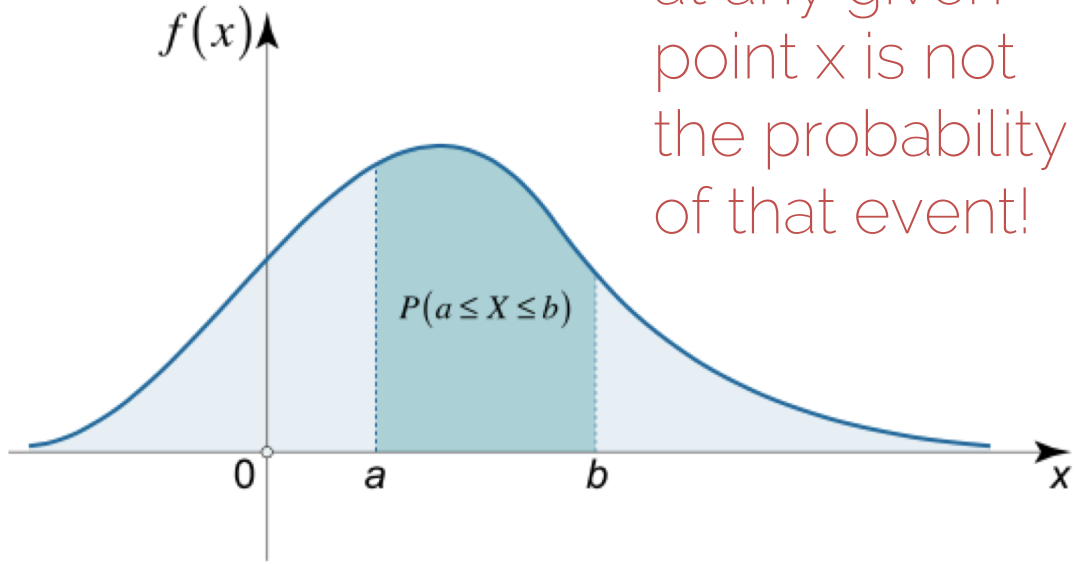
Probability Density Function

For some continuous random variables, the CDF $F_X(x)$ is differentiable everywhere. In these cases, the probability density function (PDF) is defined as the derivative of the CDF, i.e.:

$$f_x(x) = \frac{dF_X(x)}{dx}$$

Properties:

- $f_X(x) \geq 0$
- $\int_{-\infty}^{\infty} f_X(x) = 1$
- $\int_{-\infty}^{a} f_X(x)dx = F_X(a)$
- $\int_{a}^{b} f_X(x)dx = P(a \leq X \leq b)$



Note: the value of a PDF at any given point x is not the probability of that event!

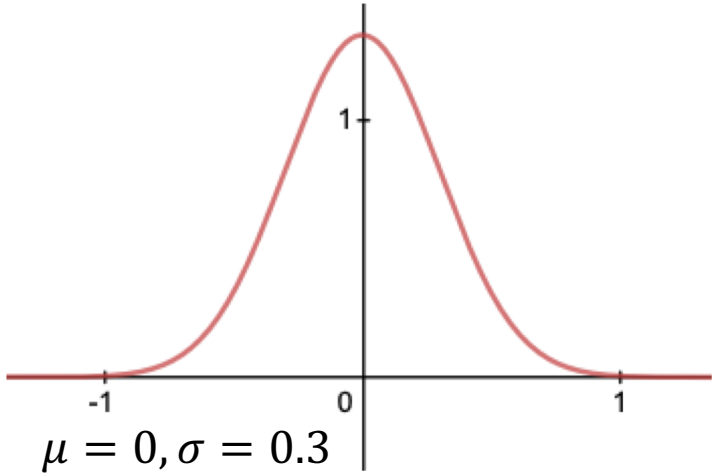
[Figure:https://www.math24.net/probability-density-function/]

Moments

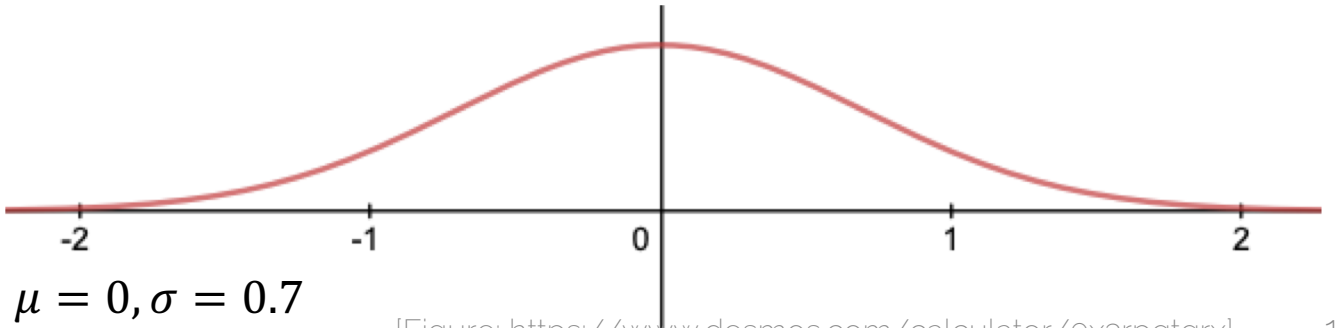
- **mean** of a random variable X is a weighted average of the possible values that the random variable can take: $\mathbb{E}[X]$
- **variance** of a random variable X is a measure of how concentrated the distribution of a random variable X is around its mean.

$$\mathbf{Var(X)} = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$$

- Standard Deviation: $\sigma = \sqrt{Var(X)}$



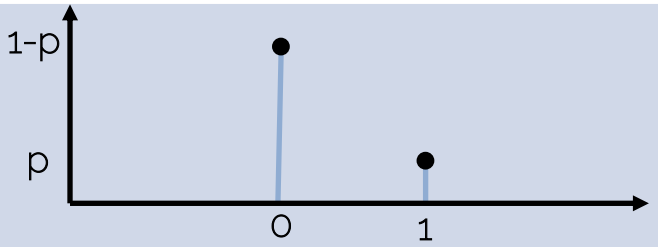
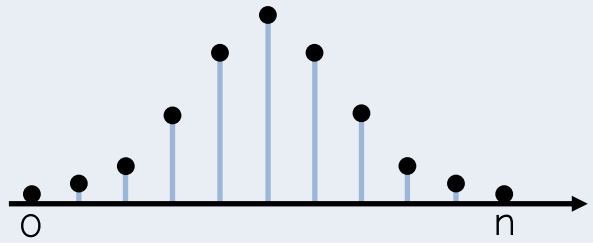
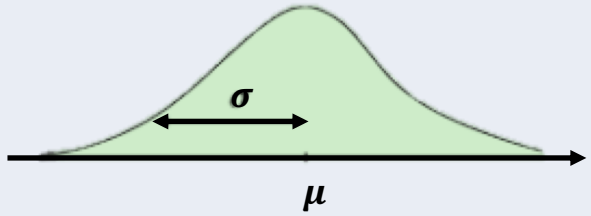
$\mu = 0, \sigma = 0.3$



$\mu = 0, \sigma = 0.7$

[Figure: https://www.desmos.com/calculator/0x3rpqtgrx]

Standard Probability Distributions

Distribution	PDF or PMF	Mean	Variance	Illustration
$Ber(p)$	$\begin{cases} p & if\ x = 1 \\ 1-p & if\ x = 0 \end{cases}$	p	$p(1-p)$	
$\mathcal{B}(n,p)$	$\binom{n}{k} p^k (1-p)^{n-k}$	np	npq	
$\mathcal{U}(a,b)$	$\frac{1}{b-a}\ \forall x \in (a,b)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	
$\mathcal{N}(\mu,\sigma^2)$	$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	μ	σ^2	

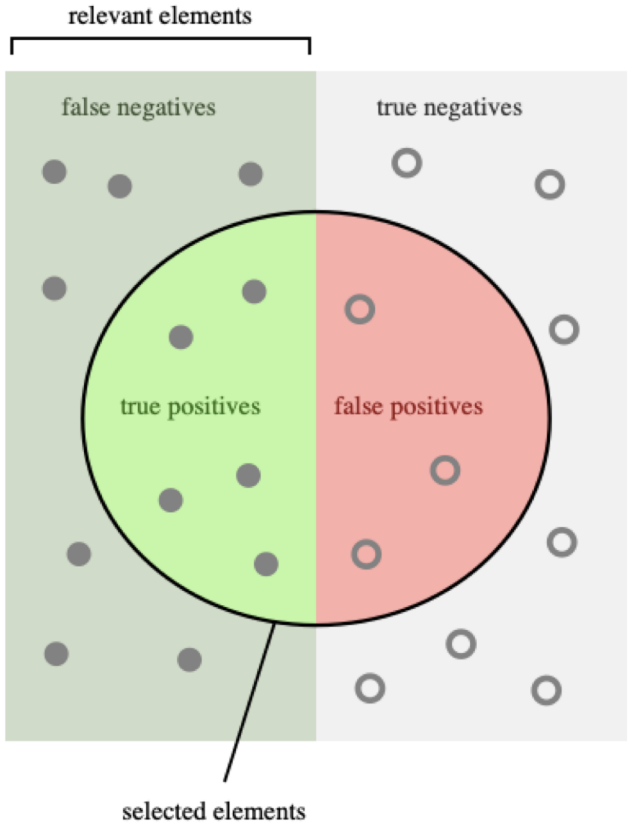
Performance Measures

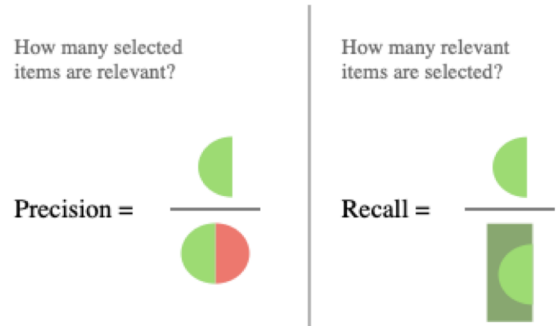
Accuracy: $acc = \frac{TP+TN}{TP+TN+FP+FN}$

Precision: $prec = \frac{TP}{TP+FP}$

Recall: $rec = \frac{TP}{TP+FN}$

F1 Score: $f1 = \frac{2 \cdot prec \cdot rec}{prec + rec}$





[Figure: https://en.wikipedia.org/wiki/Sensitivity_and_specificity]

Performance Measures

a) To support your marketing team, you train a classifier to detect whether a person is in your company's target group.

b) You train a Neural Network to predict from medical values whether a person might have COVID-20.

Which measure is more important:
Precision or Recall? – Explain!

[images: Freepik.com]

3b Information Theory

Overview

p Target distribution
q Estimated distribution

$$H(p) = -\sum_{i=1}^{n} p(x_i) \log p(x_i)$$

Entropy: expression of the disorder, or randomness of a system

$$H(p,q) = -\sum_{i=1}^{n} p(x_i) \log q(x_i)$$

Cross Entropy: uncertainty introduced by assuming estimated distribution

$$D_{KL}(p||q) = -\sum_{i=1}^{n} p(x_i) \log \frac{q(x_i)}{p(x_i)}$$

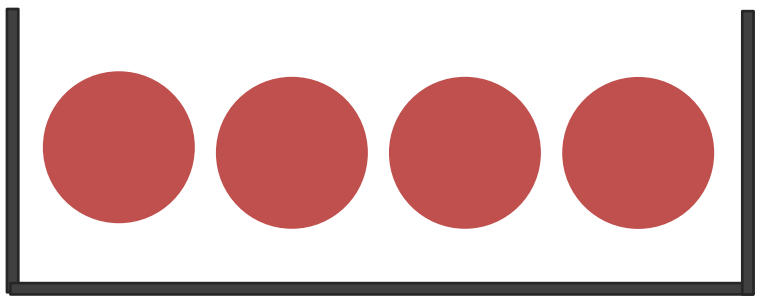
Kullback–Leibler divergence: measure for probability distributions; information gain achieved if q is used instead of p

$$D_{KL}(p||q) = H(p,q) - H(p)$$

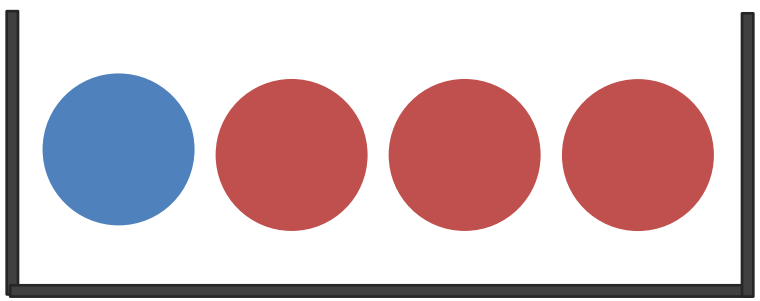
Relation between KL-Divergence, Entropy, and Cross Entropy

Entropy

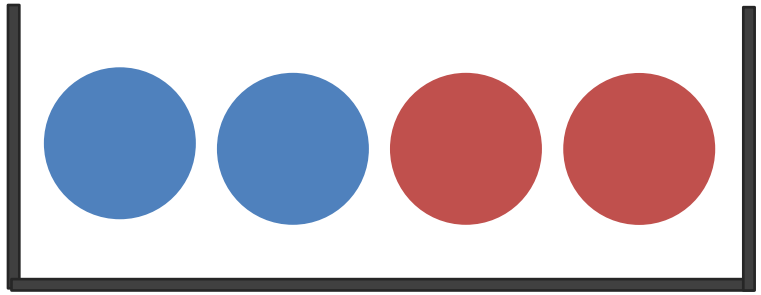
$$H(p) = -\sum_{i=1}^{n} p(x_i) \log p(x_i)$$



Bucket 1



Bucket 2

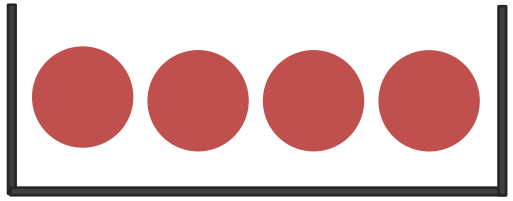


Bucket 3

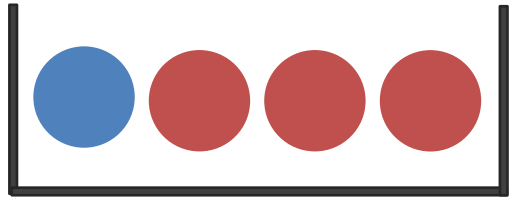
Entropy: expression of the disorder, or randomness of a system

Entropy

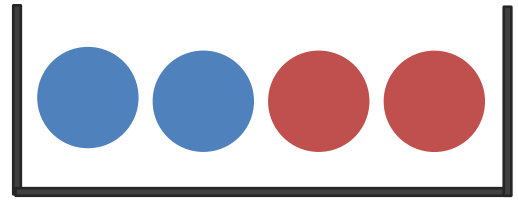
$$H(p) = -\sum_{i=1}^{n} p(x_i) \log p(x_i)$$



Bucket 1



Bucket 2



Bucket 3

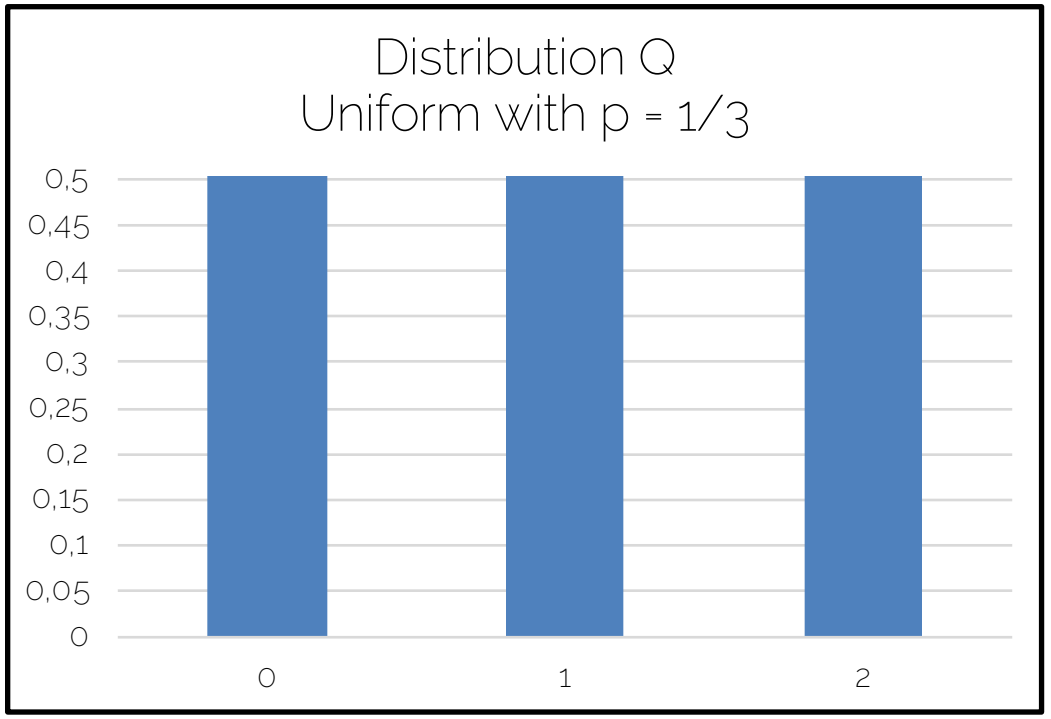
Bucket 1 $1 \cdot \log(1) = 0$

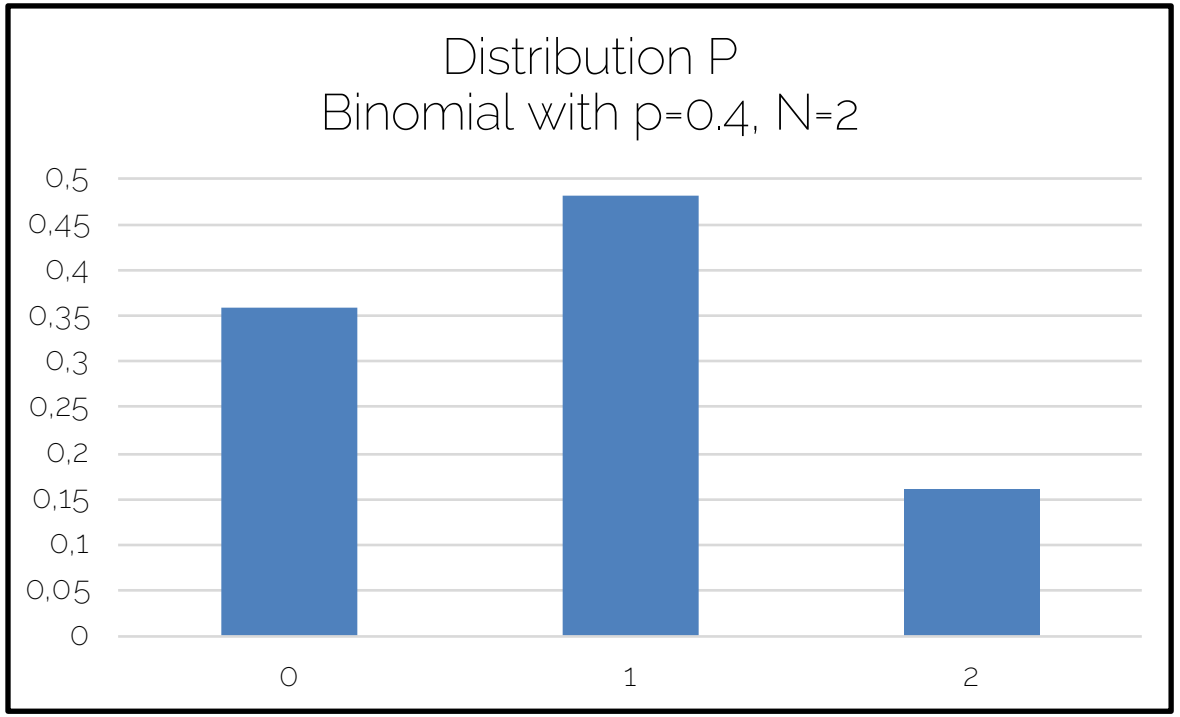
Bucket 2 $-\left(\frac{3}{4} \cdot \log\left(\frac{3}{4}\right) + \frac{1}{4} \cdot \log\left(\frac{1}{4}\right)\right) \approx 0.56$

Bucket 3 $-\left(\frac{1}{2} \cdot \log\left(\frac{1}{2}\right) + \frac{1}{2} \cdot \log\left(\frac{1}{2}\right)\right) \approx 0.69$

KL-Divergence

$$D_{KL}(p||q) = -\sum_{i=1}^{n} p(x_i) \log \frac{q(x_i)}{p(x_i)}$$



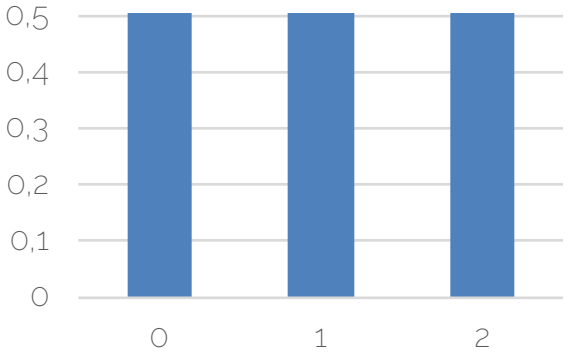


Kullback–Leibler divergence:
measure for probability distributions

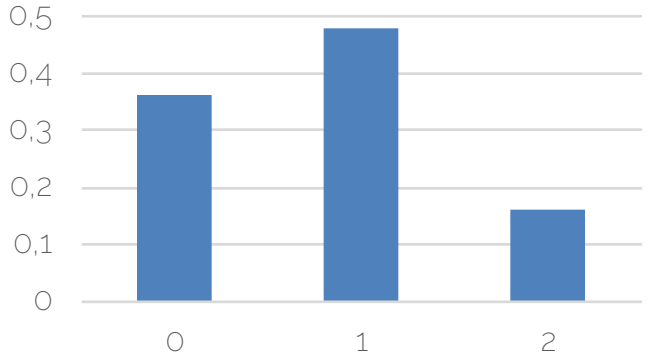
KL-Divergence

$$D_{KL}(p||q) = -\sum_{i=1}^{n} p(x_i) \log \frac{q(x_i)}{p(x_i)}$$

Distribution Q
Uniform with p = 1/3



Distribution P
Binomial with p=0.4, N=2



X	0	1	2
Distribution *P*(x)	0.36	0.48	0.16
Distribution *Q*(x)	0.333	0.333	0.333

$$D_{KL}(P||Q) = 0.36 \log\left(\frac{0.36}{0.333}\right) + 0.48 \log\left(\frac{0.48}{0.333}\right) + 0.16 \log\left(\frac{0.16}{0.333}\right) = 0.085$$

$$D_{KL}(Q||P) = 0.333 \log\left(\frac{0.333}{0.36}\right) + 0.333 \log\left(\frac{0.333}{0.48}\right) + 0.333 \log\left(\frac{0.333}{0.16}\right) = 0.096$$

Note that the KL divergence is **not** symmetric!

Jensen–Shannon Divergence

The Jensen–Shannon divergence (JSD) is a symmetrized and smoothed version of the KL-Divergence.

$$JSD(P||Q) = \frac{1}{2} D_{KL}(P||M) + \frac{1}{2} D_{KL}(Q||M)$$

$$\text{where } M = \frac{1}{2}(P + Q)$$

Cross Entropy

$$H(p,q) = -\sum_{i=1}^{n} p(x_i) \log q(x_i)$$

Classification Example


Dog
[1 0 0 0 0]


Fox
[0 1 0 0 0]


Horse
[0 0 1 0 0]


Eagle
[0 0 0 1 0]


Squirrel
[0 0 0 0 1]

Cross Entropy

$$H(p,q) = -\sum_{i=1}^{n} p(x_i) \log q(x_i)$$

Our first sample is a dog!

Labels:

$P_1(dog) = 1$ $P_1(fox) = 0$ $P_1(horse) = 0$ $P_1(eagle) = 0$ $P_1(squirrel) = 0$

Prediction:

$Q_1(dog) = 0.4$ $Q_1(fox) = 0.3$ $Q_1(horse) = 0.05$ $Q_1(eagle) = 0.05$ $Q_1(squirrel) = 0.2$

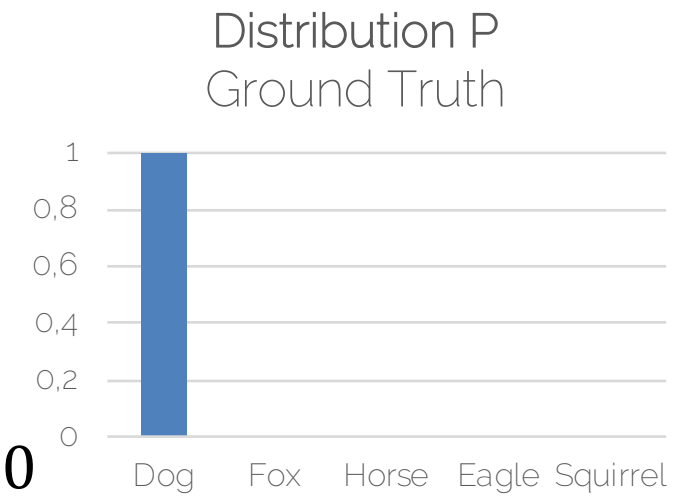
Cross Entropy: uncertainty introduced by assuming estimated distribution

Cross Entropy

$$H(p,q) = -\sum_{i=1}^{n} p(x_i) \log q(x_i)$$

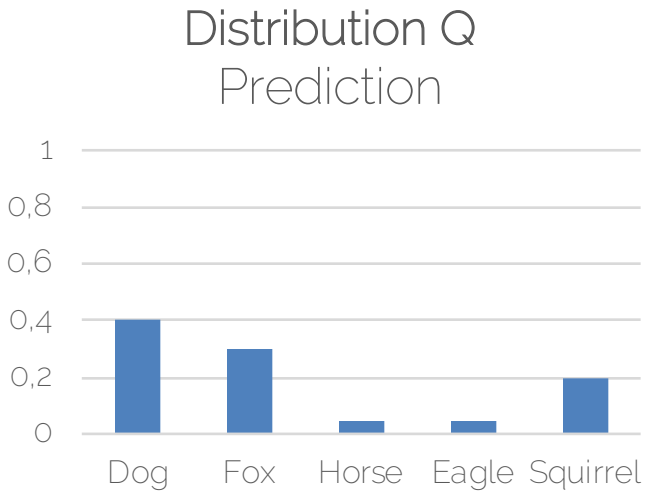
Labels:

$$P_1(dog) = 1$$
$$P_1(fox) = 0$$
$$P_1(horse) = 0$$
$$P_1(eagle) = 0$$
$$P_1(squirrel) = 0$$



Prediction:

$$Q_1(dog) = 0.4$$
$$Q_1(fox) = 0.3$$
$$Q_1(horse) = 0.05$$
$$Q_1(eagle) = 0.05$$
$$Q_1(squirrel) = 0.2$$



$$H(P_i) = 0$$

Entropy of labels is 0!

Cross Entropy "Loss":

$$H(P_1, Q_1) = -\sum_i P_1(i) \log Q_1(i)$$
$$= -\log 0.4$$
$$= 0.916$$

References

- http://cs229.stanford.edu/section/cs229-prob.pdf
  - Comprehensive Probability Review – **recommended!**
- https://stanford.edu/~shervine/teaching/cme-106/cheatsheet-probability
  - Quick Overview
- https://www.deeplearningbook.org/contents/prob.html
  - Another great resource. Also covers information theory basics.