CS 540 Introduction to Artificial Intelligence
Statistics & Linear Algebra Review

University of Wisconsin-Madison

Spring 2025

Announcements

- **HW 1 released**:
  - Writing assignment---nothing too stressful
  - Deadline **Tuesday Feb. 4th 11:59PM**
- Class roadmap:

Statistics & Linear Algebra
Linear Algebra & PCA
Logic
NLP

Mostly Foundations

Review: Bayesian Inference

- Conditional Probability & Bayes Rule:

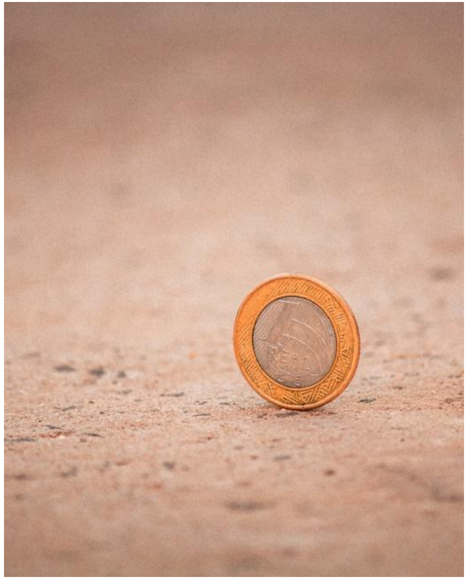
$$P(H|E) = \frac{P(E|H)P(H)}{P(E)}$$

- Evidence *E*: what we can observe
- Hypothesis *H*: what we'd like to infer from evidence
  – Need to plug in prior, likelihood, etc.
- Usually do not know these probabilities. How to estimate?

Samples and Estimation

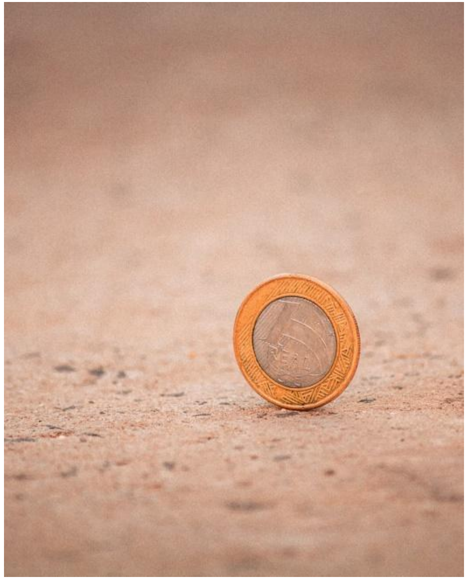
- Usually, we don't know the distribution P
  - Instead, we see a bunch of samples

- Typical statistics problem: **estimate distribution** from samples
  - Estimate probabilities $P(H)$, $P(E)$, $P(E|H)$
  - Estimate the mean $E[X]$
  - Estimate parameters $P_\theta(X)$



Samples and Estimation

- Estimate probability $P(H)$, $P(E)$, $P(E|H)$
- Estimate the mean $E[X]$
- Estimate parameters $P_\theta(X)$



- Example: Bernoulli with parameter *p* *(i.e., a weighted coin flip)*
  - $P(X = 1) = p$
  - Mean $E[X]$ is *p*

Examples: Sample Mean

- Bernoulli with parameter *p*
- See samples $x_1, x_2, \ldots, x_n$
  - Estimate mean with **sample mean**

$$\hat{\mathbb{E}}[X] = \frac{1}{n}\sum_{i=1}^{n} x_i$$

  - That is, counting heads

Break & Quiz

Q 2.1: You see samples of X given by $[0,1,1,2,2,0,1,2]$. Empirically estimate $\mathbb{E}[X^2]$

A. 9/8
B. 15/8
C. 1.5
D. There aren't enough samples to estimate $\mathbb{E}[X^2]$

Break & Quiz

Q 2.1: You see samples of X given by [0,1,1,2,2,0,1,2]. Empirically estimate $\mathbb{E}[X^2]$

A. 9/8
B. 15/8
C. 1.5
D. There aren't enough samples to estimate $\mathbb{E}[X^2]$

Break & Quiz

Q 2.1: You see samples of X given by [0,1,1,2,2,0,1,2]. Empirically estimate $\mathbb{E}[X^2]$

A. 9/8
B. 15/8
C. 1.5
D. There aren't enough samples to estimate $\mathbb{E}[X^2]$

$$E[X^2] \approx \frac{1}{n}\sum_i X_i^2$$
$$= \frac{1}{8}(0^2 + 1 + 1 + 4 + 4 + 0 + 1 + 4) \ = 15/8$$

Estimating Multinomial Parameters

- k-sized die (special case: k=2 coin)
- Face i has probability p_i, for $i=1\ldots k$
- In n rolls, we observe face i showing up n_i times

$$\sum_{i=1}^{k} n_i = n$$

- Estimate $(p_{1,\ldots,}\ p_k)$ from this data $(n_{1,\ldots,}\ n_k)$

Maximum Likelihood Estimate (MLE)

- The MLE of multinomial parameters $(\widehat{p_1}, \ldots, \widehat{p_k})$

$$\widehat{p_i} = \frac{n_i}{n}$$

- Estimate using frequencies

Break & Quiz

Q 2.2: You are empirically estimating *P(X)* for some random variable *X* that takes on 100 values. You see 50 samples. How many of your *P(X=a)* estimates might be 0?

A. None.
B. Between 5 and 50, exclusive.
C. Between 50 and 100, inclusive.
D. Between 50 and 99, inclusive.

Break & Quiz

Q 2.2: You are empirically estimating *P(X)* for some random variable *X* that takes on 100 values. You see 50 samples. How many of your *P(X=a)* estimates might be 0?

For each a, your estimate is $\mathrm{P}(X = a) = \frac{\#\text{samples taking value } a}{50}$

A. None.
B. Between 5 and 50, exclusive.
C. Between 50 and 100, inclusive.
D. Between 50 and 99, inclusive.

Break & Quiz

Q 2.2: You are empirically estimating *P(X)* for some random variable *X* that takes on 100 values. You see 50 samples. How many of your *P(X=a)* estimates might be 0?

For each a, your estimate is $\mathrm{P}(X = a) = \frac{\#\text{samples taking value } a}{50}$

A. None.
B. Between 5 and 50, exclusive.
C. Between 50 and 100, inclusive.
D. Between 50 and 99, inclusive.

Break & Quiz

Q 2.2: You are empirically estimating *P(X)* for some random variable *X* that takes on 100 values. You see 50 samples. How many of your *P(X=a)* estimates might be 0?

For each a, your estimate is $\mathrm{P}(X = a) = \frac{\#\text{samples taking value } a}{50}$

A. None.
B. Between 5 and 50, exclusive.
C. Between 50 and 100, inclusive.
D. Between 50 and 99, inclusive.

If you don't see a number at all in the 50 samples then the estimated probability of that number is 0.

You can see up to 50 different values in 50 samples. On the other hand, all 50 samples might have the same value in which case 99 values were never seen.

Regularized Estimate

- Hyperparameter $\epsilon > 0$

$$\widehat{p_i} = \frac{n_i + \epsilon}{n + k\epsilon}$$

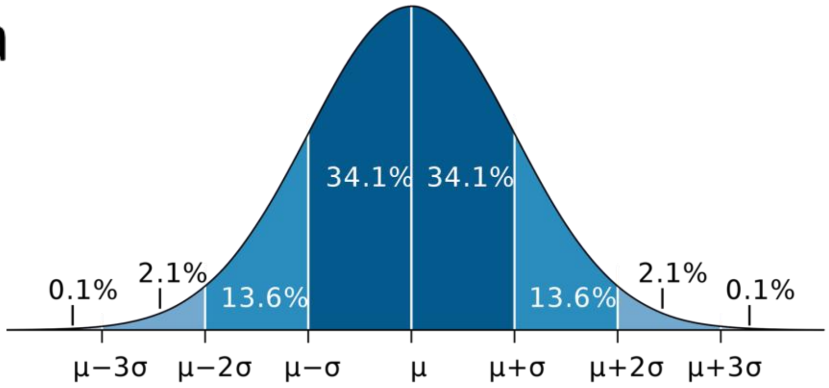
- Avoids zero when *n* is small
- Biased, but has smaller variance
- Equivalent to a specific Maximum A Posteriori (MAP) estimate, or smoothing

Estimating 1D Gaussian Parameters

- Gaussian (aka Normal) distribution $N(\mu, \sigma^2)$
  - True mean μ, true variance σ^2
- Observe *n* data points from this distribution

$$x_1, \ldots, x_n$$

- Estimate μ, σ^2 from this data



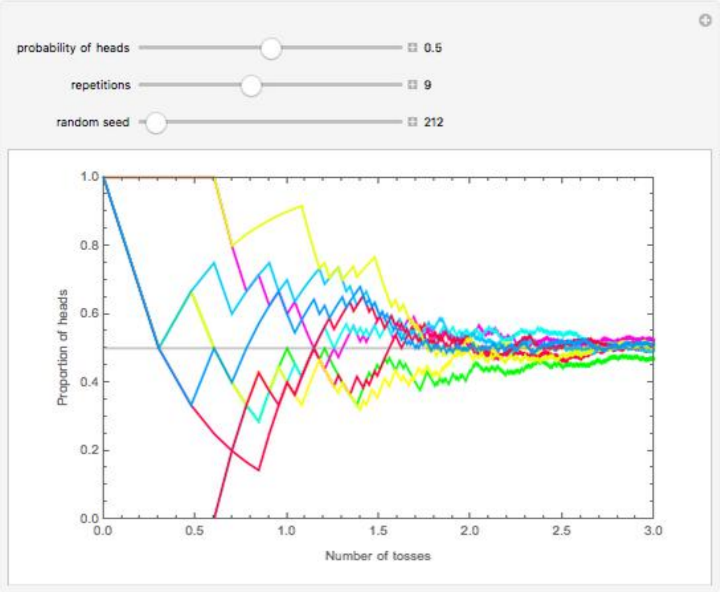
Wikipedia: Normal distribution

Estimating 1D Gaussian Parameters

- Mean estimate $\hat{\mu} = \frac{x_1 + \cdots + x_n}{n}$
- Variance estimates
  - Unbiased $s^2 = \frac{\sum_{i=1}^{n}(x_i - \hat{\mu})^2}{n-1}$
  - MLE $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(x_i - \hat{\mu})^2}{n}$

Estimation Theory

- Is the sample mean a good estimate of the true mean?
  - Law of large numbers
  - Central limit theorems



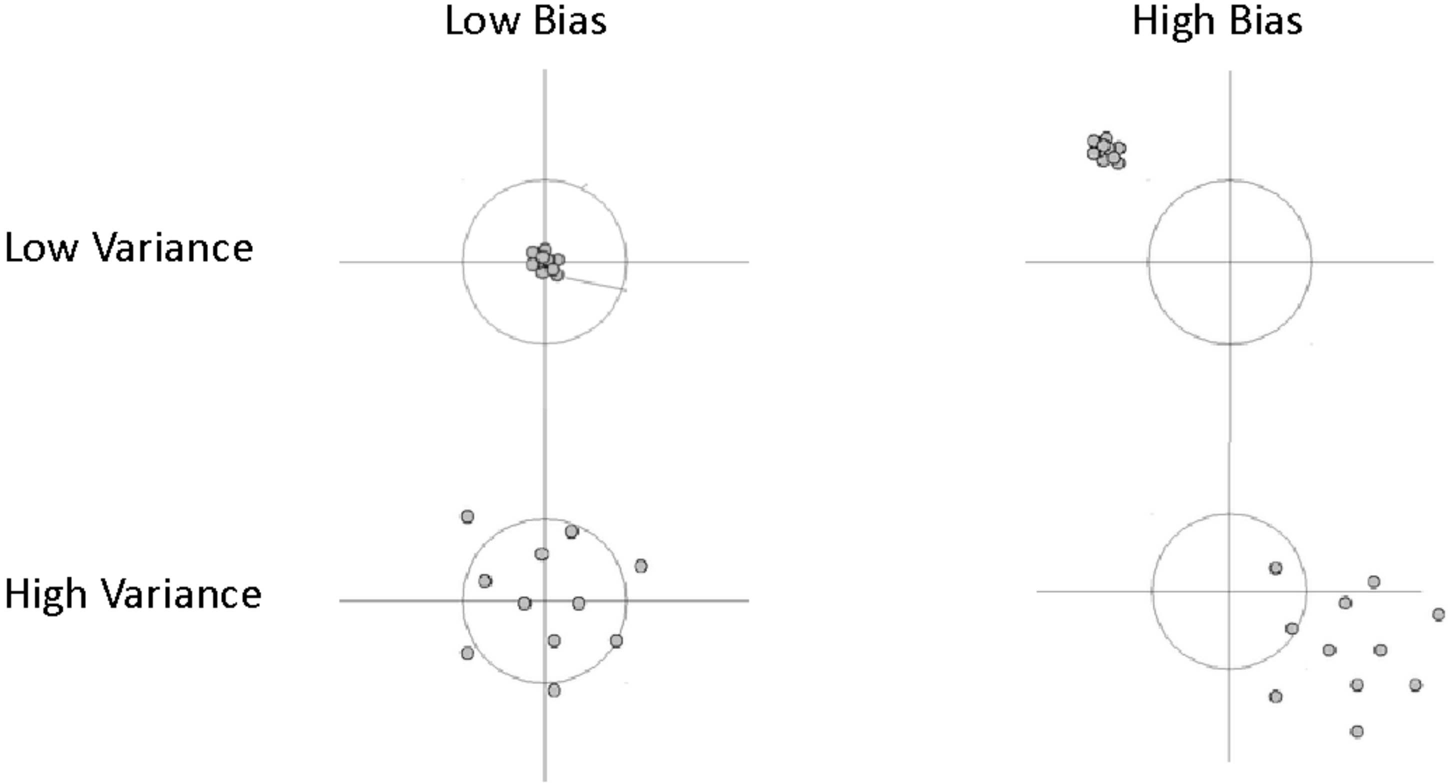
Wolfram Demo

Estimation Errors

- With finite samples, likely error in the estimate.
- Mean squared error
  - $\mathrm{MSE}[\hat{\theta}] = \mathbb{E}\left[\left(\hat{\theta} - \theta\right)^2\right]$
- Bias / Variance Decomposition
  - $\mathrm{MSE}[\hat{\theta}] = \mathbb{E}\left[\left(\hat{\theta} - E[\hat{\theta}]\right)^2\right] + \left(\mathbb{E}[\hat{\theta}] - \theta\right)^2$

Variance Bias

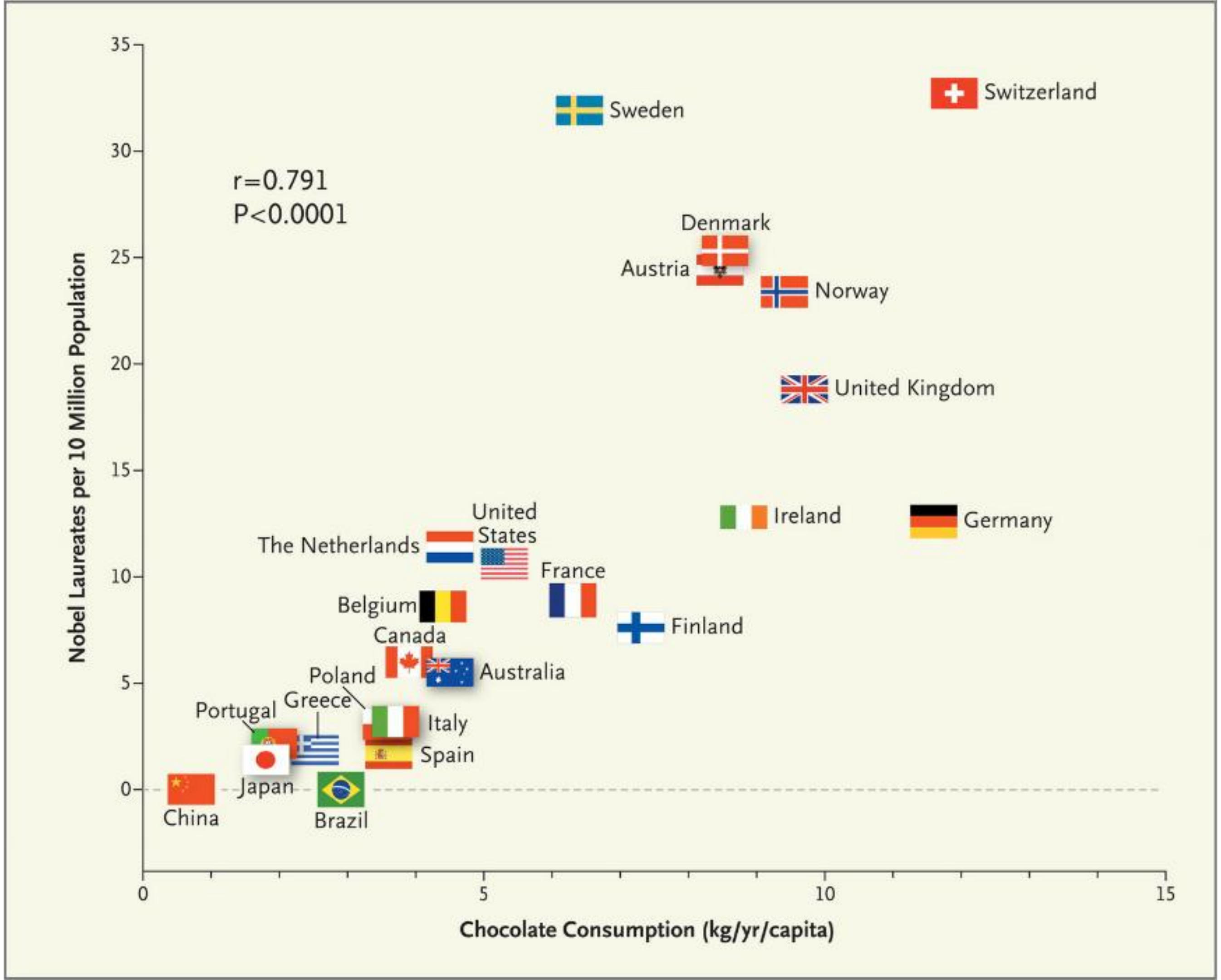
Bias / Variance



Wikipedia: Bias-variance tradeoff

Correlation vs. Causation

- Conditional probabilities only define correlation (aka association)
- $P(Y|X)$ “large” does not mean X causes Y
- Example: X=yellow finger, Y=lung cancer
- Common cause: smoking

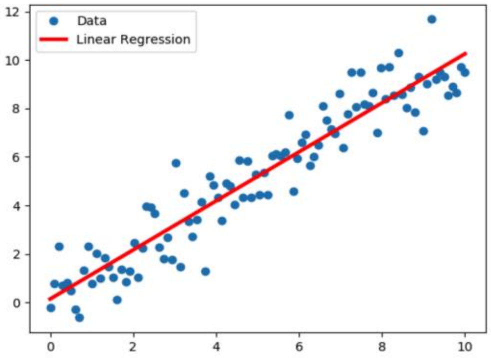


https://www.nejm.org/doi/full/10.1056/NEJMon1211064

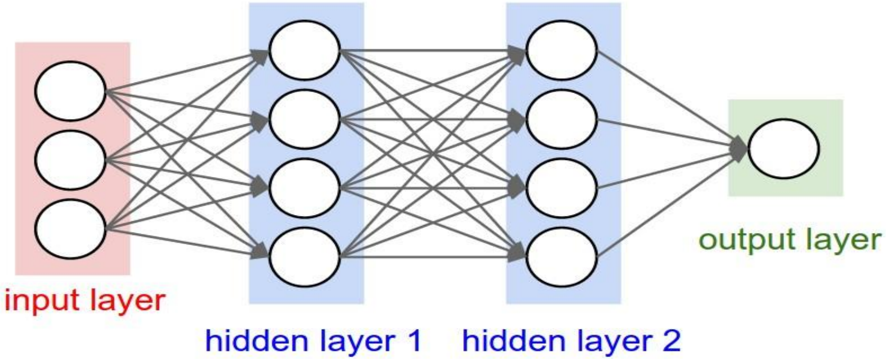
Linear Algebra

Linear Algebra: What is it good for?

- Study of Linear functions: simple, tractable
- In AI/ML: building blocks for **all models**
  - e.g., linear regression; part of neural networks



Hieu Tran

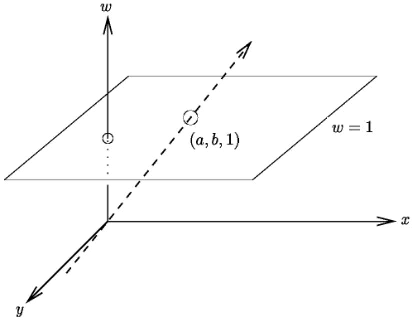


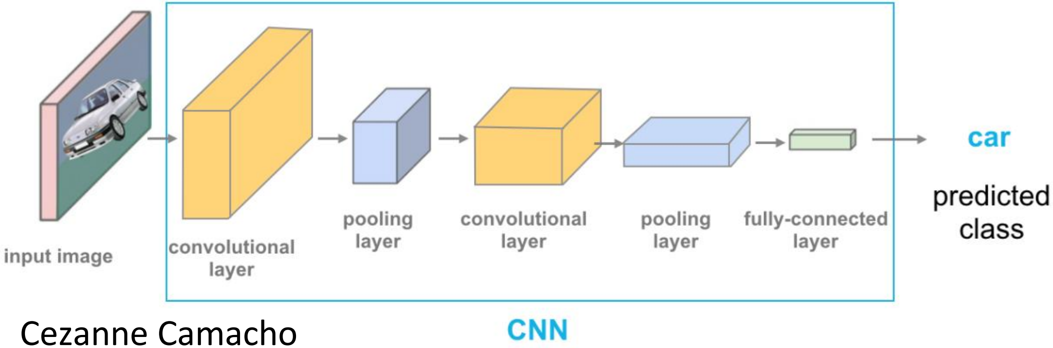
Stanford CS231n

Basics: **Vectors**

$$x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} \in \mathbb{R}^5$$

- Many interpretations
  - List of values (represents information)
  - Point in a space
- Dimension: number of values: $x \in \mathbb{R}^d$
- AI/ML: often use **very high dimensions**:
  - Ex: images!





Cezanne Camacho

Basics: **Matrices**

- Many interpretations
  - Table of values; list of vectors
  - Represent linear transformations
  - Apply to a vector, get another vector

$$A = \begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{33} & A_{33} \\ A_{41} & A_{43} & A_{43} \end{bmatrix}$$

- Dimensions: #rows $\times$ #columns, $A \in \mathbb{R}^{m \times n}$
  - Indexing!

Basics: **Transposition**

- Transposes: flip rows and columns
  - Vector: standard is a column. Transpose: row vector
  - Matrix: go from $m \times n$ to $n \times m$

$$x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \quad x^T = \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix}$$

$$A = \begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \end{bmatrix} \quad A^T = \begin{bmatrix} A_{11} & A_{21} \\ A_{12} & A_{22} \\ A_{13} & A_{23} \end{bmatrix}$$

Matrix & Vector **Operations**

- **Vectors**
  - **Addition:** component-wise
    - Commutative: $x + y = y + x$
    - Associative: $(x + y) + z = x + (y + z)$

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

  - **Scalar Multiplication**
    - Uniform stretch / scaling

$$cx = \begin{bmatrix} cx_1 \\ cx_2 \\ cx_3 \end{bmatrix}$$

Matrix & Vector **Operations**

- **Vector products**
  - **Inner product** (e.g., dot product)

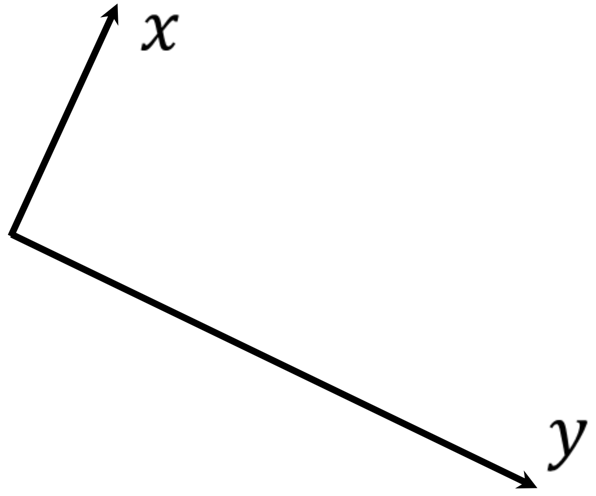
$$< x, y > := x^T y = \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = x_1 y_1 + x_2 y_2 + x_3 y_3$$

  - **Outer product**

$$xy^T = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \begin{bmatrix} y_1 & y_2 & y_3 \end{bmatrix} = \begin{bmatrix} x_1 y_1 & x_1 y_2 & x_1 y_3 \\ x_2 y_1 & x_2 y_2 & x_2 y_3 \\ x_3 y_1 & x_3 y_2 & x_3 y_3 \end{bmatrix}$$

Matrix & Vector **Operations**

- x and y are **orthogonal** if $\langle x, y \rangle$=0



- Vector **norms**: "length"

$$\|x\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2}$$

Matrix & Vector **Operations**

- **Matrices:**
  - **Addition:** Component-wise
  - Commutative, Associative

$$A + B = \begin{bmatrix} A_{11} + B_{11} & A_{12} + B_{12} \\ A_{21} + B_{21} & A_{22} + B_{22} \\ A_{31} + B_{31} & A_{32} + B_{32} \end{bmatrix}$$

  - **Scalar Multiplication**
  - "Stretching" the linear transformation

$$cA = \begin{bmatrix} cA_{11} & cA_{12} \\ cA_{21} & cA_{22} \\ cA_{31} & cA_{32} \end{bmatrix}$$

Matrix & Vector **Operations**

- **Matrix-Vector multiplication**
  - Linear transformation; plug in vector, get another vector
  - Each entry in Ax is the inner product of a row of A with x

$$x \in \mathbb{R}^n, A \in \mathbb{R}^{m \times n}$$

$$Ax = \begin{bmatrix} \langle A_{1:}, x \rangle \\ \langle A_{2:}, x \rangle \\ \vdots \\ \langle A_{m:}, x \rangle \end{bmatrix} = \begin{bmatrix} A_{11}x_1 + A_{12}x_2 + \cdots + A_{1n}x_n \\ A_{21}x_1 + A_{22}x_2 + \cdots + A_{2n}x_n \\ \vdots \\ A_{m1}x_1 + A_{m2}x_2 + \cdots + A_{mn}x_n \end{bmatrix}$$

Matrix & Vector **Operations**

Ex: feedforward neural networks. Input *x*.

- Output of layer *k* is

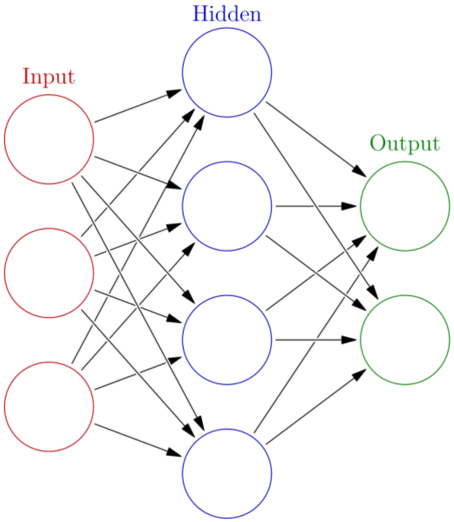
nonlinearity

$$f^{(k)}(x) = \sigma(W_k^T f^{(k-1)}(x)))$$

Output of layer k-1: **vector**

Output of layer k: vector

Weight **matrix** for layer k:
Note: linear transformation!



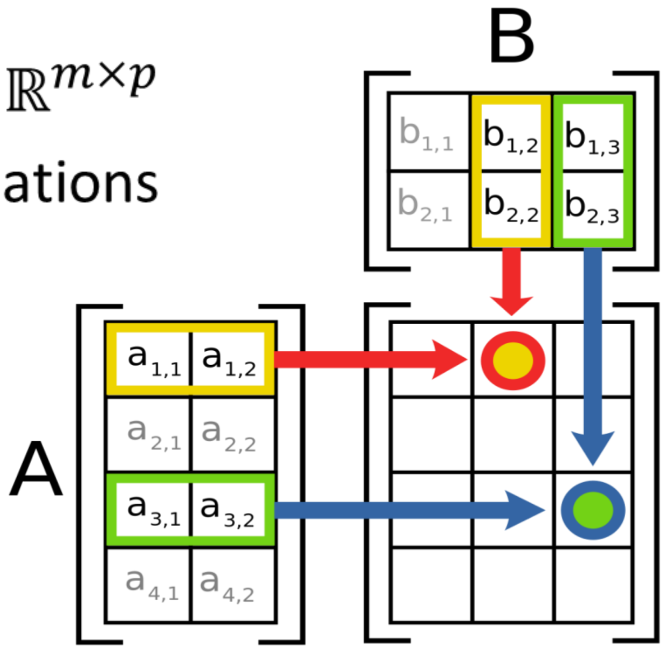
Wikipedia

Matrix & Vector **Operations**

- **Matrix multiplication**
  - $A \in \mathbb{R}^{m \times n}, B \in \mathbb{R}^{n \times p}$, then $AB \in \mathbb{R}^{m \times p}$
  - "Composition" of linear transformations
  - **Not commutative** in general!

$$AB \neq BA$$

  - Lots of interpretations



Wikipedia

Identity Matrix

- Like “1”
- Multiplying by it gets back the same matrix or vector

- Rows & columns are the “**standard basis vectors**” e_i

$$I = \begin{array}{c} \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} \\ \begin{array}{cccc} \downarrow & \downarrow & & \downarrow \\ e_1 & e_2 & \quad & e_n \end{array} \end{array}$$

Break & Quiz

- **Q 1.1**: What is $\begin{bmatrix} 1 & 2 \\ 3 & 1 \\ 1 & 1 \end{bmatrix} \times \begin{bmatrix} 0 \\ 1 \end{bmatrix}$?
- A. $[-1\ 1\ 1]^T$
- B. $[2\ 1\ 1]^T$
- C. $[1\ 3\ 1]^T$
- D. $[1.5\ 2\ 1]^T$

Break & Quiz

- **Q 1.1**: What is $\begin{bmatrix} 1 & 2 \\ 3 & 1 \\ 1 & 1 \end{bmatrix} \times \begin{bmatrix} 0 \\ 1 \end{bmatrix}$?
- A. $[-1\ 1\ 1]^T$
- **B. $[2\ 1\ 1]^T$**
- C. $[1\ 3\ 1]^T$
- D. $[1.5\ 2\ 1]^T$

Break & Quiz

- **Q 1.1**: What is $\begin{bmatrix} 1 & 2 \\ 3 & 1 \\ 1 & 1 \end{bmatrix} \times \begin{bmatrix} 0 \\ 1 \end{bmatrix}$?

- A. $[-1\ 1\ 1]^T$
- **B. $[2\ 1\ 1]^T$**
- C. $[1\ 3\ 1]^T$
- D. $[1.5\ 2\ 1]^T$

Check dimensions: answer must be 3 x 1 matrix (i.e., column vector).

$$\begin{bmatrix} 1 & 2 \\ 3 & 1 \\ 1 & 1 \end{bmatrix} \times \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 * 1 + 1 * 2 \\ 0 * 3 + 1 * 1 \\ 0 * 1 + 1 * 1 \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \\ 1 \end{bmatrix}$$

Break & Quiz

- **Q 1.2**: Given matrices $A \in \mathbb{R}^{m \times n}, B \in \mathbb{R}^{d \times m}, C \in \mathbb{R}^{p \times n}$
  What are the dimensions of BAC^T

- A. *n x p*
- B. *d x p*
- C. *d x n*
- D. Undefined

Break & Quiz

- **Q 1.2**: Given matrices $A \in \mathbb{R}^{m \times n}, B \in \mathbb{R}^{d \times m}, C \in \mathbb{R}^{p \times n}$

What are the dimensions of BAC^T

- A. *n x p*
- **B. *d x p***
- C. *d x n*
- D. Undefined

Break & Quiz

- **Q 1.2**: Given matrices $A \in \mathbb{R}^{m \times n}, B \in \mathbb{R}^{d \times m}, C \in \mathbb{R}^{p \times n}$

What are the dimensions of BAC^T

- A. *n x p*
- **B. *d x p***
- C. *d x n*
- D. Undefined

To rule out (D), check that for each pair of adjacent matrices XY, the # of columns of X = # of rows of Y

Then, B has d rows so solution must have d rows. C^T has p columns so solution has p columns.

Break & Quiz

- **Q 1.3**: A and B are matrices, neither of which is the identity. Is $AB = BA$?

- A. Never
- B. Always
- C. Sometimes

Break & Quiz

- **Q 1.3**: A and B are matrices, neither of which is the identity. Is $AB = BA$?

- A. Never
- B. Always
- **C. Sometimes**

Break & Quiz

- **Q 1.3**: A and B are matrices, neither of which is the identity. Is $AB = BA$?

- A. Never
- B. Always
- **C. Sometimes**

Matrix multiplication is not necessarily commutative.

Thanks!