Neural Networks and Deep Learning

Dr. Jerome J. Braun

This Lecture:
Math Preliminaries (Review)

Course: Neural Networks and Deep Learning
IE 7615

This Lecture

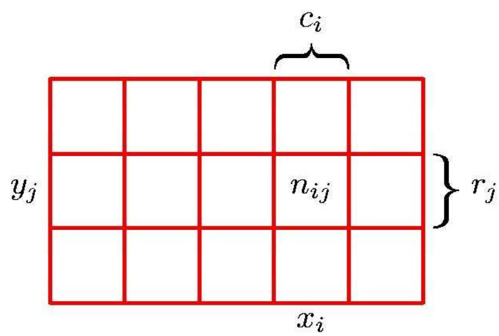
Review selected basics of

- **Probability theory**
- **Linear algebra**

Probability Mass Function

- **Domain of P — set of all possible states of x**
- **Range** $\forall x \in \mathrm{x}, 0 \leq P(x) \leq 1$
  - ✦ **Impossible event: P=0**
  - ✦ **Guaranteed event: P=1**
- **Normalization** $\sum_{x \in \mathrm{x}} P(x) = 1$

Probabilities: Joint, Marginal, Conditional



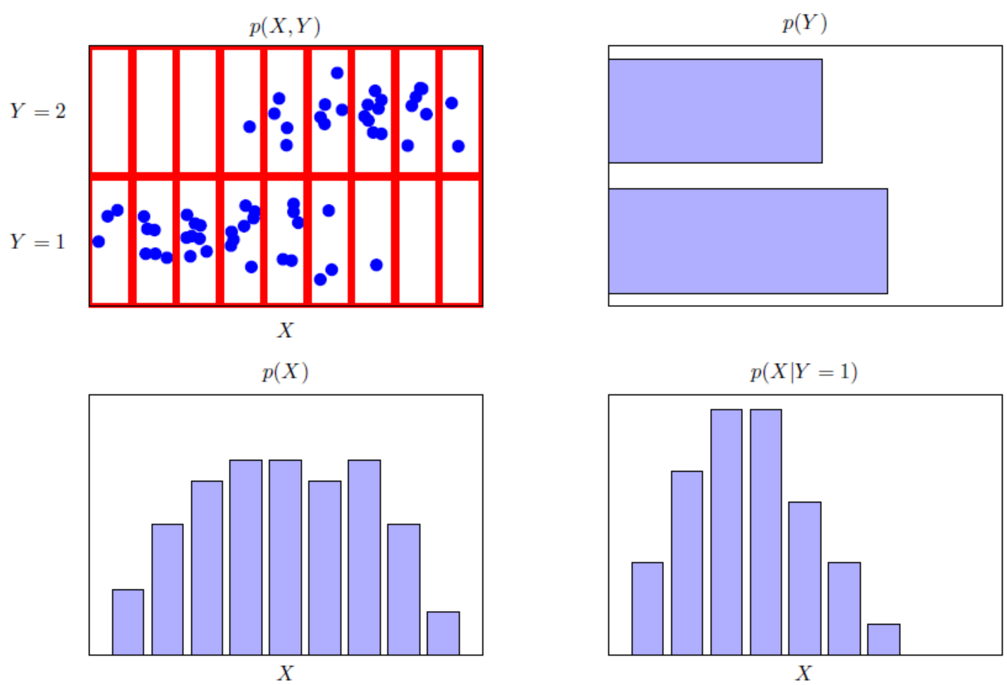
[Bishop, Ch. 1.2]

Marginal probability $\quad p(X = x_i) = \frac{c_i}{N}$

Joint probability $\quad p(X = x_i, Y = y_j) = \frac{n_{ij}}{N}$

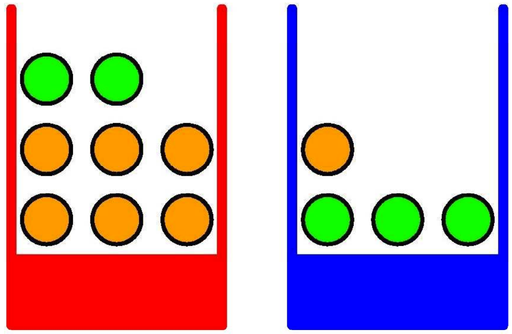
Conditional probability $\quad p(Y = y_j | X = x_i) = \frac{n_{ij}}{c_i}$

Example



[Bishop]

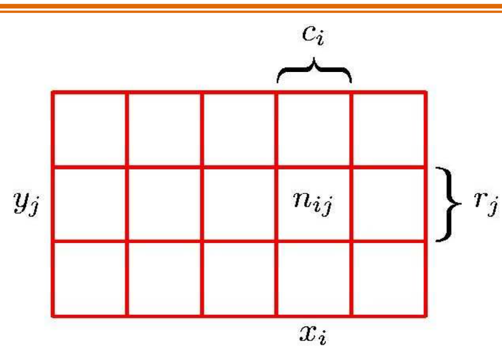
Basic Probability — Example



[Bishop, Ch. 1.2]

- **Experiment**
  - **Randomly pick box (red or blue)**
  - **Then from that box randomly pick item (fruit — apple or orange)**
  - **Observe result, then replace item back**
  - **Repeat many times**
- **Question: What is overall probability of picking apple?**

Sum Rule

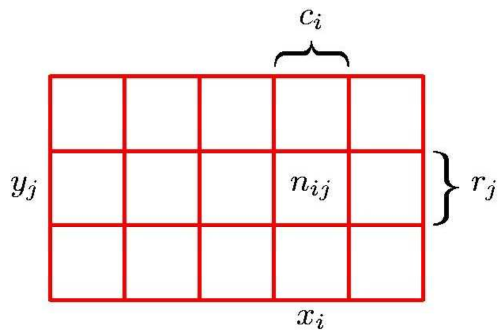


[Bishop, Ch. 1.2]

$$p(X = x_i) = \frac{c_i}{N} = \frac{1}{N} \sum_{j=1}^{L} n_{ij}$$

$$= \sum_{j=1}^{L} p(X = x_i, Y = y_j)$$

Product Rule



[Bishop, Ch. 1.2]

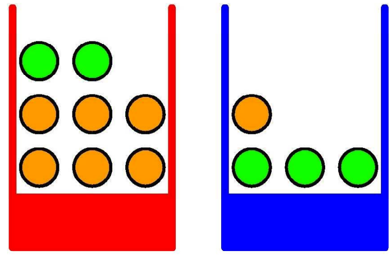
$$
\begin{aligned}
p(X = x_i, Y = y_j) &= \frac{n_{ij}}{N} = \frac{n_{ij}}{c_i} \cdot \frac{c_i}{N} \\
&= p(Y = y_j | X = x_i) p(X = x_i)
\end{aligned}
$$

Rules of Probability

- **Sum rule** $$p(X) = \sum_Y p(X, Y)$$
- **Product rule** $$p(X, Y) = p(Y|X)p(X)$$

Basic Probability— Example

- Pick box at random
  - Turns out B=b (blue)
- What is probability of picking apple?
  - How about probability of picking orange?



[Bishop, Ch. 1.2]

$$\begin{aligned} p(F=a|B=r) &= 1/4 \\ p(F=o|B=r) &= 3/4 \\ p(F=a|B=b) &= 3/4 \\ p(F=o|B=b) &= 1/4. \end{aligned}$$

$$\begin{aligned} p(B=r) &= 4/10 \\ p(B=b) &= 6/10 \end{aligned}$$

Normalization

$$p(F=a|B=r)+p(F=o|B=r)=1$$
$$p(F=a|B=b)+p(F=o|B=b)=1$$

Apply sum and product rules

$$p(F=a) \quad = \quad p(F=a|B=r)p(B=r)+p(F=a|B=b)p(B=b) \quad = \quad \frac{1}{4}\times\frac{4}{10}+\frac{3}{4}\times\frac{6}{10}=\frac{11}{20}$$

By sum rule

$$p(F=o)=1-11/20=9/20$$

Bayes' Theorem

Likelihood

Prior

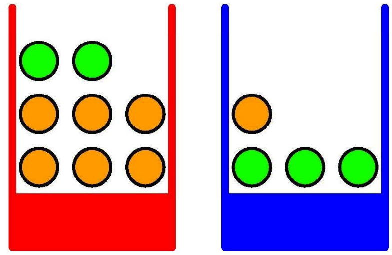
$$p(Y|X)=\frac{p(X|Y)p(Y)}{p(X)}$$

$$p(X)=\sum_Y p(X|Y)p(Y)$$

$$\boxed{\text{Posterior} \propto \text{Likelihood} \times \text{Prior}}$$

Using Bayes' Theorem — Example

- **Told that fruit picked was orange**
- **From which box did it come?**
  - **How about probability of picking orange?**



[Bishop, Ch. 1.2]

$$
\begin{aligned}
p(B=r) &= 4/10 \\
p(B=b) &= 6/10
\end{aligned}
$$

Reversing conditional probability and using Bayes' Theorem

$$p(B=r|F=o) = \frac{p(F=o|B=r)p(B=r)}{p(F=o)} = \frac{3}{4} \times \frac{4}{10} \times \frac{20}{9} = \frac{2}{3}$$

Probability Densities

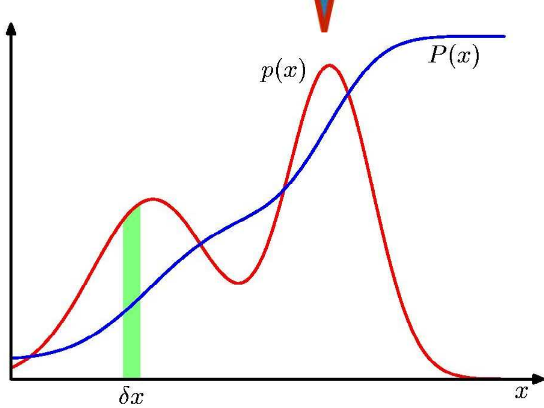
$$p(x) \geqslant 0 \qquad p(x \in (a,b)) = \int_a^b p(x)\,\mathrm{d}x$$

Probability density function (pdf)

$$\int_{-\infty}^{\infty} p(x)\,\mathrm{d}x = 1$$

$$P(z) = \int_{-\infty}^{z} p(x)\,\mathrm{d}x$$

[Bishop, Ch. 1]



Cumulative distribution function (cdf)

Multivariate Probability Densities

Continuous variables

$$x_1, \ldots, x_D$$

Joint probability density

$$p(\mathbf{x}) = p(x_1, \ldots, x_D)$$

Infinitesimal volume

$$p(\mathbf{x})\delta\mathbf{x}$$

p(x) must satisfy

$$\begin{cases} p(\mathbf{x}) & \geqslant & 0 \\ \int p(\mathbf{x})\,\mathrm{d}\mathbf{x} & = & 1 \end{cases}$$

over entire D-dim space

Continuous Densities

- **For continuous variables**
- **Sum and product rules apply also in case of continuous densities**
- **For instance if x and y are real variables:**

$$p(x) = \int p(x, y)\,\mathrm{d}y$$

$$p(x, y) = p(y|x)p(x)$$

- **Bayes' theorem also applies**

Expectations

Discrete

$$\mathbb{E}[f] = \sum_x p(x) f(x)$$

Continuous

$$\mathbb{E}[f] = \int p(x) f(x)\, \mathrm{d}x$$

Conditional (discrete)

$$\mathbb{E}_x[f|y] = \sum_x p(x|y) f(x)$$

Approximate

$$\mathbb{E}[f] \simeq \frac{1}{N} \sum_{n=1}^{N} f(x_n)$$

Linearity of expectations

$$\mathbb{E}_{\mathrm{X}}[\alpha f(x) + \beta g(x)] = \alpha \mathbb{E}_{\mathrm{X}} f(x) + \beta \mathbb{E}_{\mathrm{X}} g(x)$$

Variance and Covariance

Variance

$$\mathrm{var}[f] = \mathbb{E}\left[\left(f(x) - \mathbb{E}[f(x)] \right)^2 \right] = \mathbb{E}[f(x)^2] - \mathbb{E}[f(x)]^2$$

Covariance

$$\mathrm{Cov}(f(x), g(y)) = \mathbb{E}\left[\left(f(x) - \mathbb{E}\left[f(x) \right] \right) \left(g(y) - \mathbb{E}\left[g(y) \right] \right) \right]$$

Covariance (scalar x, y)

$$\begin{aligned} \mathrm{cov}[x, y] &= \mathbb{E}_{x,y}\left[\{x - \mathbb{E}[x]\} \{y - \mathbb{E}[y]\} \right] \\ &= \mathbb{E}_{x,y}[xy] - \mathbb{E}[x]\mathbb{E}[y] \end{aligned}$$

Covariance (vector x, y)

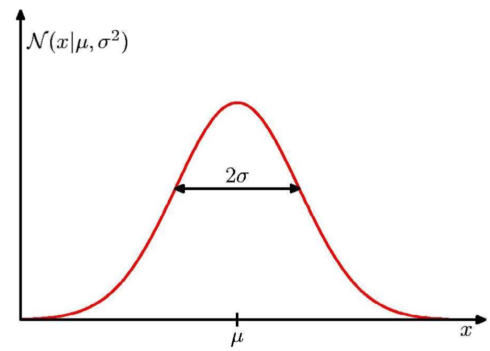
$$\begin{aligned} \mathrm{cov}[\mathbf{x}, \mathbf{y}] &= \mathbb{E}_{\mathbf{x},\mathbf{y}}\left[\{\mathbf{x} - \mathbb{E}[\mathbf{x}]\} \{\mathbf{y}^{\mathrm{T}} - \mathbb{E}[\mathbf{y}^{\mathrm{T}}]\} \right] \\ &= \mathbb{E}_{\mathbf{x},\mathbf{y}}[\mathbf{x}\mathbf{y}^{\mathrm{T}}] - \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{y}^{\mathrm{T}}] \end{aligned}$$

Univariate Gaussian Distribution

$$\mathcal{N}\left(x|\mu,\sigma^2\right) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

Mean

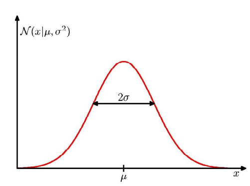
Variance



$$\mathcal{N}(x|\mu,\sigma^2) > 0$$

$$\int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu,\sigma^2\right) \mathrm{d}x = 1$$

Gaussian Mean and Variance



$$\mathbb{E}[x] = \int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu,\sigma^2\right) x \,\mathrm{d}x = \mu$$

$$\mathbb{E}[x^2] = \int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu,\sigma^2\right) x^2 \,\mathrm{d}x = \mu^2 + \sigma^2$$

$$\mathrm{var}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2 = \sigma^2$$

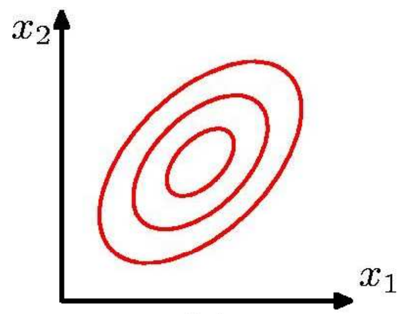
Multivariate Gaussian Distribution

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

Mean vector

Covariance matrix

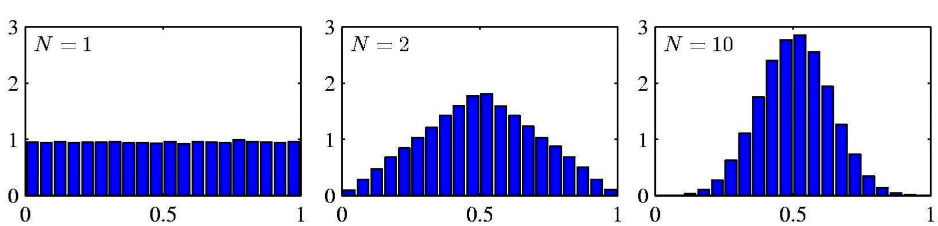
Space dimensionality



Central Limit Theorem (Laplace)

- Sum of N random variables
  - Becomes increasingly Gaussian as N increases
    - (under some mild conditions)

- Example — N uniform [0,1] random variables:

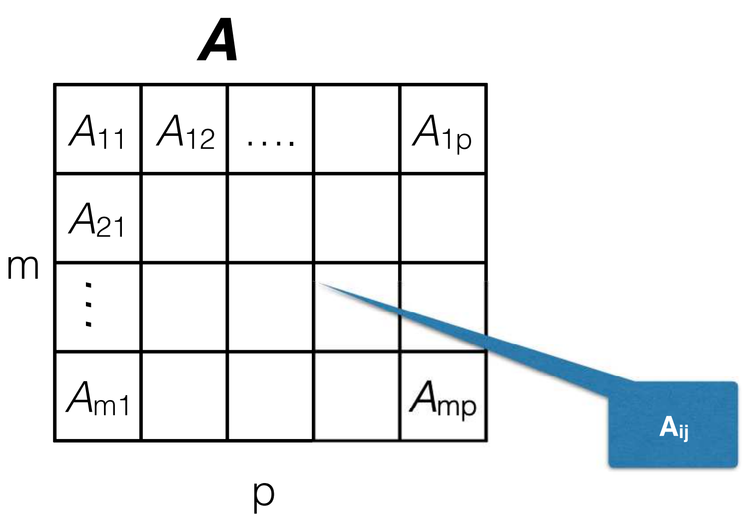


This Lecture

Review selected basics of

- **Probability theory**
- **Linear algebra**
  - **Matrices and vectors**
  - **Matrix-vector multiplication**
  - **Matrix-matrix multiplication**
  - **Matrix multiplication properties**
  - **Matrix inverse and transpose**
  - **Eigenvalues and eigenvectors**

Matrix



$$\begin{bmatrix} 17 & 242 & 22 \\ -12 & 6 & 54 \\ 80 & 5 & -85 \\ 35 & 2 & 4 \end{bmatrix}$$

2D matrix:
m rows, p columns

Vector: Nx1 Matrix

$$V = \begin{bmatrix} 17 \\ -12 \\ 80 \\ 35 \end{bmatrix}$$

$$v = \begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ \dots \\ \dots \\ \dots \\ v_n \end{bmatrix} \qquad v = \begin{bmatrix} v_0 \\ v_1 \\ v_2 \\ \dots \\ \dots \\ \dots \\ v_{n-1} \end{bmatrix}$$

1-indexed or 0-indexed

Matrix Addition

$$\begin{bmatrix} 4 & 10 \\ -3 & 20 \\ 10 & 2 \\ 1 & 8 \end{bmatrix} + \begin{bmatrix} 3 & 12 \\ -5 & 1 \\ 0 & 6 \\ 5 & -3 \end{bmatrix} =$$

$$\begin{bmatrix} 17 & 242 & 22 \\ -12 & 6 & 54 \\ 80 & 5 & -85 \\ 35 & 2 & 4 \end{bmatrix} + \begin{bmatrix} 3 & -5 \\ 12 & 2 \end{bmatrix} =$$

Multiplying/Dividing by Scalar

$$\begin{bmatrix} 7 & 2 \\ -2 & 8 \\ 10 & -4 \end{bmatrix} \times 2 =$$

$$3 \times \begin{bmatrix} 7 & 2 \\ -2 & 8 \\ 10 & -4 \end{bmatrix} =$$

$$\begin{bmatrix} 8 & 20 \\ -20 & 4 \\ 40 & 0 \end{bmatrix} /4 =$$

Combining Operations

$$\begin{bmatrix} 8 & 16 \\ -4 & 4 \\ 12 & -16 \end{bmatrix} /4 \; + \begin{bmatrix} 2 & 0 \\ 0 & 4 \\ -1 & 0 \end{bmatrix} \times 2 \; + \begin{bmatrix} 1 & 1 \\ 2 & 0 \\ -2 & 1 \end{bmatrix} =$$

Today

Review selected basics of

- **Probability theory**
- **Linear algebra**
  - **Matrices and vectors**
  - **Matrix-vector multiplication**
  - **Matrix-matrix multiplication**
  - **Matrix multiplication properties**
  - **Matrix inverse and transpose**
  - **Eigenvalues and eigenvectors**

Matrix-Vector Multiplication

$$A \times \vec{x} = \vec{y}$$

$$y_i = \sum_{k=1}^{n} A_{ik} x_k$$

$$\underset{m \times n}{\begin{bmatrix} \quad & \quad \end{bmatrix}} \times \underset{n \times 1}{\begin{bmatrix} \quad \\ \quad \end{bmatrix}} = \underset{m \times 1}{\begin{bmatrix} \quad \end{bmatrix}}$$

Matrix-Vector Multiplication Examples

$$\begin{bmatrix} 2 & 3 \\ -1 & 1 \\ 0 & 4 \end{bmatrix} \cdot \begin{bmatrix} 2 \\ -1 \end{bmatrix} =$$

$$\begin{bmatrix} 2 & 1 & 3 & 5 \\ 1 & -1 & 2 & 3 \\ 4 & 0 & -1 & 2 \end{bmatrix} \cdot \begin{bmatrix} 3 \\ -2 \\ 1 \\ -2 \end{bmatrix} =$$

Matrix-Vector Product and Prediction

Hypothesis as function of variable x, and parameters w

$$h(x) = w_0 + w_1 x$$

Hypothesis (prediction for given x)

$$\boxed{h(x) = -5 + 0.1x}$$

Different values of x

h

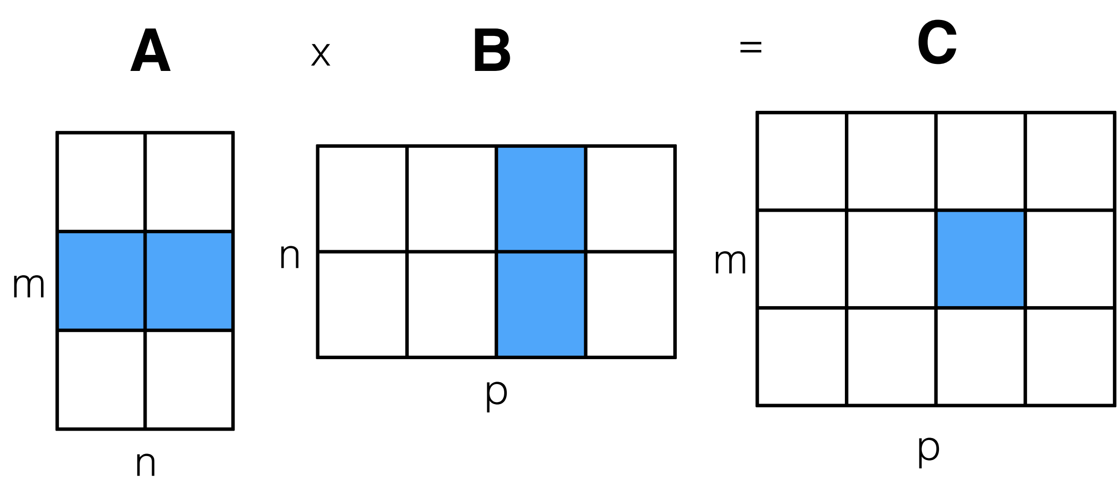
$$\begin{bmatrix} 1 & 10 \\ 1 & 20 \\ 1 & 30 \\ 1 & 40 \end{bmatrix} * \begin{bmatrix} -5 \\ 0.1 \end{bmatrix} = \begin{bmatrix} -4 \\ -3 \\ -2 \\ -1 \end{bmatrix}$$

This Lecture

Review selected basics of

- **Probability theory**
- **Linear algebra**
  - **Matrices and vectors**
  - **Matrix-vector multiplication**
  - **Matrix-matrix multiplication**
  - **Matrix multiplication properties**
  - **Matrix inverse and transpose**
  - **Eigenvalues and eigenvectors**

Matrix-Matrix Multiplication

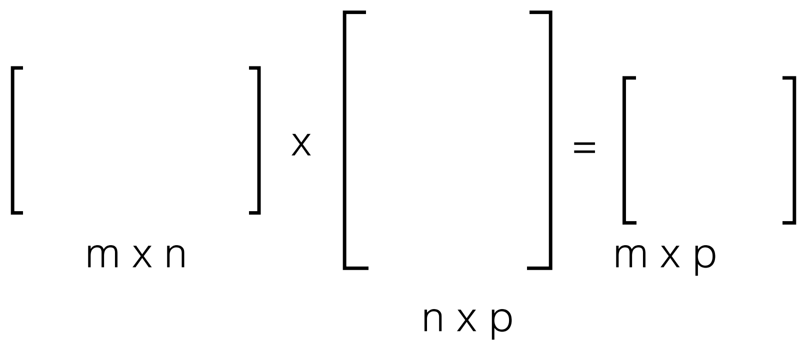


Note matching dimension *n*

Matrix-Matrix Multiplication

$$A \times B = C$$

$$C_{ij} = \sum_{k=1}^{n} A_{ik} B_{kj}$$



Matrix Multiplication Examples

$$\begin{bmatrix} 2 & 3 & 1 \\ -1 & 1 & 4 \\ 3 & -4 & 1 \end{bmatrix} \cdot \begin{bmatrix} 2 \\ 0 \\ -1 \end{bmatrix} =$$

$$\begin{bmatrix} 1 & 3 & 1 \\ 3 & -4 & 1 \end{bmatrix} \cdot \begin{bmatrix} 2 \\ 1 \\ -1 \end{bmatrix} =$$

$$\begin{bmatrix} 1 & 3 & 1 \\ 3 & -4 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 & 2 \\ -1 & 1 \\ 0 & -1 \end{bmatrix} =$$

Linear Equations

$$2x - y = 0$$
$$x + y = 27$$

$$\begin{bmatrix} 2 & -1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 0 \\ 27 \end{bmatrix}$$

Multiple Hypotheses

Hypothesis as function of variable x, and parameters w

$$h(x) = w_0 + w_1 x$$

Three different hypothses

$$h^1(x) = -5 + 0.1x$$
$$h^2(x) = 80 + 2x$$
$$h^3(x) = 40 + 15x$$

$x = \ldots$

$$\begin{bmatrix} 1 & 10 \\ 1 & 20 \\ 1 & 30 \\ 1 & 40 \end{bmatrix} * \begin{bmatrix} -5 & 80 & 40 \\ 0.1 & 2 & 15 \end{bmatrix} = \begin{array}{c} \begin{array}{ccc} h^1 & h^2 & h^3 \end{array} \\ \begin{bmatrix} -4 & 100 & 190 \\ -3 & 120 & 340 \\ -2 & 140 & 490 \\ -1 & 160 & 640 \end{bmatrix} \end{array}$$

This Lecture

Review selected basics of

- **Probability theory**
- **Linear algebra**
  - **Matrices and vectors**
  - **Matrix-vector multiplication**
  - **Matrix-matrix multiplication**
  - **Matrix multiplication properties**
  - **Matrix inverse and transpose**
  - **Eigenvalues and eigenvectors**

Matrix Non-Commutativity

- **In general, matrices do not commute**

$$A \times B \neq B \times A$$

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

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

- **In some cases commutativity exists**

Matrix Associativity and Distributivity

- **Matrix multiplication is associative**
  - **(AB)C = A(BC)**
- **Matrix multiplication is distributive**
  - **A(B + C) = AB + AC**
-

Identity Matrix

- **Denoted as $\mathbf{I}$ or $\mathbf{I}_{nxm}$**

$$I_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$$

- **Example:**

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

- **For any matrix A**
  - $$\mathbf{A} \cdot \mathbf{I} = \mathbf{I} \cdot \mathbf{A} = \mathbf{A}$$

This Lecture

Review selected basics of

- Probability theory
- Linear algebra
  - Matrices and vectors
  - Matrix-vector multiplication
  - Matrix-matrix multiplication
  - Matrix multiplication properties
  - Matrix inverse and transpose
  - Eigenvalues and eigenvectors

Matrix Inverse

- If square matrix A has inverse then

$$\mathbf{A} \cdot \mathbf{A}^{-1} = \mathbf{A}^{-1} \cdot \mathbf{A} = \mathbf{I}$$

- Singular or degenerate matrix — inverse does not exist
- Ex: Find inverse of 2x2 matrix
  -

Matrix Invertibility

- **A - square matrix**
- **A invertible iff**
  - **Determinant is non-zero**
- **A invertible iff**
  - **columns linearly independent**
  - **Equivalently — rows independent**
- **Cannot be inverted if has redundant rows/column ("linearly dependent", "low rank")**

Matrix Transpose

- **If A is n x m matrix and B=A^T, then**
  - **B is m x n matrix, and**

$$B_{ij} = A_{ji}$$

- **If y=Ax, then**

$$\mathbf{y}^T = (\mathbf{A}\mathbf{x})^T = \mathbf{x}^T\mathbf{A}^T$$

- $$\begin{bmatrix} 1 & 2 \\ -1 & 1 \\ 0 & -1 \end{bmatrix}^T =$$

This Lecture

Review selected basics of

- **Probability theory**
- **Linear algebra**
  - **Matrices and vectors**
  - **Matrix-vector multiplication**
  - **Matrix-matrix multiplication**
  - **Matrix multiplication properties**
  - **Matrix inverse and transpose**
  - **Eigenvalues and eigenvectors**

Eigenvalues and Eigenvectors

- **Eigenvector x of matrix A — x lies on same line as Ax**
  - **$Ax = \lambda x$**
    - **Scalar λ is eigenvalue**
- **Scalar λ is an eigenvalue of A iff $(A - \lambda I)$ is singular**
- **Characteristic-polynomial equation for eigenvalues**
  - **$\det(A - \lambda I) = 0$**
    - **If A is n x n, then above has n solutions (some may repeats)**
- **Finding eigenvectors**
  - **For each eigenvalue, solve $(A - \lambda I)x=0$**

Questions?