# Lecture 2: Maths Review Linear Algebra \& Probability 

COMP 590/776: Computer Vision
Instructor: Soumyadip (Roni) Sengupta
TA: Mykhailo (Misha) Shvets

## Overview

- Linear Algebra Review
- Multivariate Calculus Review
- Probability Review


## Overview

- Linear Algebra Review
- Multivariate Calculus Review
- Probability Review


## Vector and Matrix Products

Inner Product

$$
x^{T} y \in \mathbb{R}=\left[\begin{array}{llll}
x_{1} & x_{2} & \cdots & x_{n}
\end{array}\right]\left[\begin{array}{c}
y_{1} \\
y_{2} \\
\vdots \\
y_{n}
\end{array}\right]=\sum_{i=1}^{n} x_{i} y_{i} .
$$

## Outer Product

$$
x y^{T} \in \mathbb{R}^{m \times n}=\left[\begin{array}{c}
x_{1} \\
x_{2} \\
\vdots \\
x_{m}
\end{array}\right]\left[\begin{array}{llll}
y_{1} & y_{2} & \cdots & y_{n}
\end{array}\right]=\left[\begin{array}{cccc}
x_{1} y_{1} & x_{1} y_{2} & \cdots & x_{1} y_{n} \\
x_{2} y_{1} & x_{2} y_{2} & \cdots & x_{2} y_{n} \\
\vdots & \vdots & \ddots & \vdots \\
x_{m} y_{1} & x_{m} y_{2} & \cdots & x_{m} y_{n}
\end{array}\right]
$$

## Vector and Matrix Products

$$
C=A B=\left[\begin{array}{c}
-a_{1}^{T} \\
-a_{2}^{T} \\
\vdots \\
\vdots \\
-a_{m}^{T}
\end{array}\right]\left[\begin{array}{cccc}
\mid & \mid & & \mid \\
a_{1} & b_{2} & \cdots & b_{p} \\
\mid & \mid & & \mid
\end{array}\right]=\left[\begin{array}{cccc}
a_{1}^{T} b_{1} & a_{1}^{T} b_{2} & \cdots & a_{1}^{T} b_{p} \\
a_{2}^{T} b_{1} & a_{2}^{T} b_{2} & \cdots & a_{2}^{T} b_{p} \\
\vdots & \vdots & \ddots & \vdots \\
a_{m}^{T} b_{1} & a_{m}^{T} & \ddots & \cdots \\
a_{m}^{T} & \cdots & a_{m}^{T} b_{p}
\end{array}\right] .
$$

## Transpose

The transpose of a matrix results from "flipping" the rows and columns. Given a matrix $A \in \mathbb{R}^{m \times n}$, its transpose, written $A^{T} \in \mathbb{R}^{n \times m}$, is the $n \times m$ matrix whose entries are given by

$$
\left(A^{T}\right)_{i j}=A_{j i} .
$$

- $\left(A^{T}\right)^{T}=A$
- $(A B)^{T}=B^{T} A^{T}$
- $(A+B)^{T}=A^{T}+B^{T}$


## Symmetric Matrix

A square matrix $A \in \mathbb{R}^{n \times n}$ is symmetric if $A=A^{T}$. It is anti-symmetric if $A=-A^{T}$. It is easy to show that for any matrix $A \in \mathbb{R}^{n \times n}$, the matrix $A+A^{T}$ is symmetric and the matrix $A-A^{T}$ is anti-symmetric. From this it follows that any square matrix $A \in \mathbb{R}^{n \times n}$ can be represented as a sum of a symmetric matrix and an anti-symmetric matrix, since

$$
A=\frac{1}{2}\left(A+A^{T}\right)+\frac{1}{2}\left(A-A^{T}\right)
$$



## Trace

The trace of a square matrix $A \in \mathbb{R}^{n \times n}$, denoted $\operatorname{tr}(A)$ (or just $\operatorname{tr} A$ if the parentheses are obviously implied), is the sum of diagonal elements in the matrix:

$$
\operatorname{tr} A=\sum_{i=1}^{n} A_{i i} .
$$

- For $A \in \mathbb{R}^{n \times n}, \operatorname{tr} A=\operatorname{tr} A^{T}$.
- For $A, B \in \mathbb{R}^{n \times n}, \operatorname{tr}(A+B)=\operatorname{tr} A+\operatorname{tr} B$.
- For $A \in \mathbb{R}^{n \times n}, t \in \mathbb{R}, \operatorname{tr}(t A)=t \operatorname{tr} A$.
- For $A, B$ such that $A B$ is square, $\operatorname{tr} A B=\operatorname{tr} B A$.
- For $A, B, C$ such that $A B C$ is square, $\operatorname{tr} A B C=\operatorname{tr} B C A=\operatorname{tr} C A B$, and so on for the product of more matrices.


## Norm

A norm of a vector $\|x\|$ is informally a measure of the "length" of the vector. For example, we have the commonly-used Euclidean or $\ell_{2}$ norm,

$$
\|x\|_{2}=\sqrt{\sum_{i=1}^{n} x_{i}^{2}}
$$

Note that $\|x\|_{2}^{2}=x^{T} x$.
More formally, a norm is any function $f: \mathbb{R}^{n} \rightarrow \mathbb{R}$ that satisfies 4 properties:

1. For all $x \in \mathbb{R}^{n}, f(x) \geq 0$ (non-negativity).
2. $f(x)=0$ if and only if $x=0$ (definiteness).
3. For all $x \in \mathbb{R}^{n}, t \in \mathbb{R}, f(t x)=|t| f(x)$ (homogeneity).
4. For all $x, y \in \mathbb{R}^{n}, f(x+y) \leq f(x)+f(y)$ (triangle inequality).

## Norm

Other examples of norms are the $\ell_{1}$ norm,

$$
\|x\|_{1}=\sum_{i=1}^{n}\left|x_{i}\right|
$$

and the $\ell_{\infty}$ norm,

$$
\|x\|_{\infty}=\max _{i}\left|x_{i}\right|
$$

In fact, all three norms presented so far are examples of the family of $\ell_{p}$ norms, which are parameterized by a real number $p \geq 1$, and defined as

$$
\|x\|_{p}=\left(\sum_{i=1}^{n}\left|x_{i}\right|^{p}\right)^{1 / p}
$$

Norms can also be defined for matrices, such as the Frobenius norm,

$$
\|A\|_{F}=\sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} A_{i j}^{2}}=\sqrt{\operatorname{tr}\left(A^{T} A\right)}
$$

## Orthogonality

Two vectors $x, y \in \mathbb{R}^{n}$ are orthogonal if $x^{T} y=0$. A vector $x \in \mathbb{R}^{n}$ is normalized if $\|x\|_{2}=1$. A square matrix $U \in \mathbb{R}^{n \times n}$ is orthogonal (note the different meanings when talking about vectors versus matrices) if all its columns are orthogonal to each other and are normalized (the columns are then referred to as being orthonormal).

$$
U^{T} U=I=U U^{T}
$$

## Linear Independence

A set of vectors $\left\{x_{1}, x_{2}, \ldots x_{n}\right\} \subset \mathbb{R}^{m}$ is said to be (linearly) independent if no vector can be represented as a linear combination of the remaining vectors. Conversely, if one vector belonging to the set can be represented as a linear combination of the remaining vectors, then the vectors are said to be (linearly) dependent. That is, if

$$
x_{n}=\sum_{i=1}^{n-1} \alpha_{i} x_{i}
$$

for some scalar values $\alpha_{1}, \ldots, \alpha_{n-1} \in \mathbb{R}$, then we say that the vectors $x_{1}, \ldots, x_{n}$ are linearly dependent; otherwise, the vectors are linearly independent.

## Rank of a matrix

Rank of a matrix $A$ is the maximal number of linearly independent columns or rows.
A matrix is full ranked, if all of its columns/rows are linearly independent.

- For $A \in \mathbb{R}^{m \times n}, \operatorname{rank}(A) \leq \min (m, n)$. If $\operatorname{rank}(A)=\min (m, n)$, then $A$ is said to be full rank.
- For $A \in \mathbb{R}^{m \times n}, \operatorname{rank}(A)=\operatorname{rank}\left(A^{T}\right)$.
- For $A \in \mathbb{R}^{m \times n}, B \in \mathbb{R}^{n \times p}, \operatorname{rank}(A B) \leq \min (\operatorname{rank}(A), \operatorname{rank}(B))$.
- For $A, B \in \mathbb{R}^{m \times n}, \operatorname{rank}(A+B) \leq \operatorname{rank}(A)+\operatorname{rank}(B)$.


## Inverse

$$
A^{-1} A=I=A A^{-1}
$$

In order for a square matrix $A$ to have an inverse $A^{-1}$, then $A$ must be full rank. We will soon see that there are many alternative sufficient and necessary conditions, in addition to full rank, for invertibility.

The following are properties of the inverse; all assume that $A, B \in \mathbb{R}^{n \times n}$ are non-singular:

- $\left(A^{-1}\right)^{-1}=A$
- $(A B)^{-1}=B^{-1} A^{-1}$
- $\left(A^{-1}\right)^{T}=\left(A^{T}\right)^{-1}$. For this reason this matrix is often denoted $A^{-T}$.

As an example of how the inverse is used, consider the linear system of equations, $A x=b$ where $A \in \mathbb{R}^{n \times n}$, and $x, b \in \mathbb{R}^{n}$. If $A$ is nonsingular (i.e., invertible), then $x=A^{-1} b$. (What if $A \in \mathbb{R}^{m \times n}$ is not a square matrix? Does this work?)

## Determinant

How do we find determinant of a nxn matrix?
See Leibniz formula for determinants

$$
\left|\left[a_{11}\right]\right|=a_{11}
$$

$$
\begin{aligned}
\left.\|\left[\begin{array}{ll}
a_{11} & a_{12} \\
a_{21} & a_{22}
\end{array}\right] \right\rvert\, & =a_{11} a_{22}-a_{12} a_{21} \\
\left|\left|\left[\begin{array}{lll}
a_{11} & a_{12} & a_{13} \\
a_{21} & a_{22} & a_{23} \\
a_{31} & a_{32} & a_{33}
\end{array}\right]\right|\right. & =\begin{array}{r}
a_{11} a_{22} a_{33}+a_{12} a_{23} a_{31}+a_{13} a_{21} a_{32} \\
-a_{11} a_{23} a_{32}-a_{12} a_{21} a_{33}-a_{13} a_{22} a_{31}
\end{array}
\end{aligned}
$$

- Determinant of $2 \times 2$ matrix is the area of the parallelogram formed by the column vectors of the matrix.
- Determinant of $3 \times 3$ matrix is the volume of a parallelopiped formed by the 3 column vectors of the matrix
- Sign indicates whether the transformation preserves or reverse orientation.



## Eigenvalue and Eigenvectors

Given a square matrix $A \in \mathbb{R}^{n \times n}$, we say that $\lambda \in \mathbb{C}$ is an eigenvalue of $A$ and $x \in \mathbb{C}^{n}$ is the corresponding eigenvector ${ }^{3}$ if

$$
A x=\lambda x, \quad x \neq 0 .
$$

$$
(\lambda I-A) x=0, \quad x \neq 0
$$

$|(\lambda I-A)|=0 \longrightarrow\left(\lambda_{1}-\lambda\right)\left(\lambda_{2}-\lambda\right) \cdots\left(\lambda_{n}-\lambda\right), \quad$ Characteristic polynomial

$$
A X=X \Lambda \quad \longrightarrow X \in \mathbb{R}^{n \times n}=\left[\begin{array}{cccc}
\mid & \mid & & \mid \\
x_{1} & x_{2} & \cdots & x_{n} \\
\mid & \mid & & \mid
\end{array}\right], \Lambda=\operatorname{diag}\left(\lambda_{1}, \ldots, \lambda_{n}\right)
$$

If the eigenvectors of $A$ are linearly independent, then the matrix $X$ will be invertible, so $A=X \Lambda X^{-1}$. A matrix that can be written in this form is called diagonalizable.

## Eigenvalue and Eigenvectors

- The trace of a $A$ is equal to the sum of its eigenvalues,

$$
\operatorname{tr} A=\sum_{i=1}^{n} \lambda_{i}
$$

- The determinant of $A$ is equal to the product of its eigenvalues,

$$
|A|=\prod_{i=1}^{n} \lambda_{i} .
$$

## Eigenvalue and Eigenvectors

Given a square matrix $A \in \mathbb{R}^{n \times n}$, we say that $\lambda \in \mathbb{C}$ is an eigenvalue of $A$ and $x \in \mathbb{C}^{n}$ is the corresponding eigenvector if

$$
A x=\lambda x, \quad x \neq 0 . \quad\left(\begin{array}{cc}
I & M \\
0 & I
\end{array}\right)
$$



In this shear mapping the red arrow changes direction, but the blue arrow does not. The blue arrow is an eigenvector of this shear mapping because it does not change direction, and since its length is unchanged, its eigenvalue is 1 .

$$
\binom{x^{\prime}}{y^{\prime}}=\binom{x+m y}{y}=\left(\begin{array}{cc}
1 & m \\
0 & 1
\end{array}\right)\binom{x}{y}
$$

$[a, 0]$ is an eigenvector for any value of $a$.

## Singular Value Decomposition

$$
A=U D V^{T} \quad A=\sum_{i=1}^{r} \sigma_{i} \mathbf{u}_{\mathbf{i}} \mathbf{v}_{\mathbf{i}}^{T}
$$



U and V are orthonormal matrices, i.e. $U^{\top} U=I$ and $V^{\top} V=1$
$D$ is a diagonal matrix, where each diagonal element is known as singular values. $\mathrm{D}_{\mathrm{ii}}=\sigma_{\mathrm{i}}$
$r$ is the rank of the matrix
$r<=\min (n, d)$

## Singular Value Decomposition <br> $$
A=U D V^{T} \quad A=\sum_{i=1}^{r} \sigma_{i} u_{i} \mathrm{v}_{\mathrm{i}}^{T} .
$$



| $D$ |
| :---: | :---: |
| $r \times r$ |
| $T \times d$ |

Consider matrix $A$ as collection of ' $n$ ' d-dimensional vectors $\mathrm{a}_{\mathrm{i}}$. Let us consider vas unit vector in d-dimensional space.

Then $\left|a_{i} \cdot v\right|$ is the magnitude of project of each data point $a_{i}$ onto $v$. $|A . v|^{2}$ is the sum of the squared distances of all the data points to the line $v$.


$$
\mathbf{v}_{\mathbf{1}}=\arg \max _{|\mathbf{v}|=1}|A \mathbf{v}|
$$

Finding $\mathrm{v}_{1}$ indicates the direction in which the data is most spread. -> Most informative direction of the data

## Singular Value Decomposition

$$
A x=U S V^{T} x \quad \text { Change of basis }
$$



Applying A to any vector x can be visualized as...

## Eigen-decomposition vs SVD

## A = P.D.P ${ }^{-1}$

$$
A=U \cdot D \cdot V^{\top}
$$

- The vectors in the eigen-decomposition matrix $P$ are not necessarily orthogonal, so the change of basis isn't a simple rotation. On the other hand, the vectors in the matrices $U$ and $V$ in the SVD are orthonormal, so they do represent rotations (and possibly flips).
- In the SVD, the nondiagonal matrices $U$ and $V$ are not necessarily the inverse of one another. They are usually not related to each other at all. In the eigen decomposition the nondiagonal matrices $P$ and $P^{-1}$ are inverses of each other.
- The SVD always exists for any sort of rectangular or square matrix, whereas the eigen decomposition can only exists for square matrices, and even among square matrices sometimes it doesn't exist (eigen vectors need to be linearly independent).

They are same when $A$ is positive semi-definite matrix, i.e.
An $n \times n$ symmetric real matrix $M$ is said to be positive-semidefinite or non-negative-definite if $\mathbf{x}^{\top} M \mathbf{x} \geq 0$ for all $\mathbf{x}$ in $\mathbb{R}^{n}$. Formally,
$M$ positive semi-definite $\Longleftrightarrow \mathbf{x}^{\top} M \mathbf{x} \geq 0$ for all $\mathbf{x} \in \mathbb{R}^{n}$

## Overview

- Linear Algebra Review
- Multivariate Calculus Review
- Probability Review


## Gradient

Suppose that $f: \mathbb{R}^{m \times n} \rightarrow \mathbb{R}$ is a function that takes as input a matrix $A$ of size $m \times n$ and returns a real value. Then the gradient of $f$ (with respect to $A \in \mathbb{R}^{m \times n}$ ) is the matrix of partial derivatives, defined as:

$$
\nabla_{A} f(A) \in \mathbb{R}^{m \times n}=\left[\begin{array}{cccc}
\frac{\partial f(A)}{\partial A_{11}} & \frac{\partial f(A)}{\partial A_{12}} & \cdots & \frac{\partial f(A)}{\partial A_{1 n}} \\
\frac{\partial f(A)}{\partial A_{21}} & \frac{\partial f(A)}{\partial A_{22}} & \cdots & \frac{\partial f(A)}{\partial A_{2 n}} \\
\vdots & \vdots & \ddots & \vdots \\
\frac{\partial f(A)}{\partial A_{m 1}} & \frac{\partial f(A)}{\partial A_{m 2}} & \cdots & \frac{\partial f(A)}{\partial A_{m n}}
\end{array}\right]
$$

i.e., an $m \times n$ matrix with

$$
\left(\nabla_{A} f(A)\right)_{i j}=\frac{\partial f(A)}{\partial A_{i j}} .
$$

## Gradient

$\frac{\partial y}{\partial \mathbf{x}} \quad$| y is scalar |
| :--- |
| x is a $\mathrm{n} \times 1 \operatorname{dim}$ vector |$\longrightarrow \quad$ What is the dimension?


| Types of matrix derivative |  |  |  |
| :---: | :---: | :---: | :---: |
| Types | Scalar | Vector | Matrix |
| Scalar | $\frac{\partial y}{\partial x}$ | $\frac{\partial \mathbf{y}}{\partial x}$ | $\frac{\partial \mathbf{Y}}{\partial x}$ |
| Vector | $\frac{\partial y}{\partial \mathbf{x}}$ | $\frac{\partial \mathbf{y}}{\partial \mathbf{x}}$ |  |
| Matrix | $\frac{\partial y}{\partial \mathbf{X}}$ |  |  |

$\frac{\partial \operatorname{tr}(\mathbf{X})}{\partial \mathbf{X}} \longrightarrow$ What is the result?
$\frac{\partial \operatorname{tr}\left(\mathbf{X}^{\top} \mathbf{A} \mathbf{X}\right)}{\partial \mathbf{X}}=\longrightarrow$ What is the result?

## Hessian

Suppose that $f: \mathbb{R}^{n} \rightarrow \mathbb{R}$ is a function that takes a vector in $\mathbb{R}^{n}$ and returns a real number. Then the Hessian matrix with respect to $x$, written $\nabla_{x}^{2} f(x)$ or simply as $H$ is the $n \times n$ matrix of partial derivatives,

$$
\left.\nabla_{x}^{2} f(x) \in \mathbb{R}^{n \times n}=\left[\begin{array}{cccc}
\frac{\partial^{2} f(x)}{\partial x_{1}^{2}} & \frac{\partial^{2} f(x)}{\partial x_{1} \partial x_{2}} & \cdots & \frac{\partial^{2} f(x)}{\partial x_{1} \partial x_{n}} \\
\frac{\partial^{2} f(x)}{\partial x_{2} \partial x_{1}} & \frac{\partial^{2} f(x)}{\partial x_{2}^{2}} & \cdots & \frac{\partial^{2} f(x)}{\partial x_{2} \partial x_{n}} \\
\vdots & \vdots & \ddots & \vdots \\
\frac{\partial^{2} f(x)}{\partial x_{n} \partial x_{1}} & \frac{\partial^{2} f(x)}{\partial x_{n} \partial x_{2}} & \cdots & \frac{\partial^{2} f(x)}{\partial x_{n}^{2}}
\end{array}\right] \cdots \quad \nabla_{x f(x)=}^{\frac{\partial f(x)}{\partial x_{1}}} \begin{array}{c}
\frac{\partial f(x)}{\partial x_{2}} \\
\vdots \\
\frac{\partial f(x)}{\partial x_{n}}
\end{array}\right]
$$

Note that the Hessian is always symmetric, since

$$
\frac{\partial^{2} f(x)}{\partial x_{i} \partial x_{j}}=\frac{\partial^{2} f(x)}{\partial x_{j} \partial x_{i}} .
$$

## Overview

- Linear Algebra Review
- Multivariate Calculus Review
- Probability Review

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

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

Let $S_{1}, \ldots, S_{k}$ be events, $P\left(S_{i}\right)>0$. Then the chain rule states that:

$$
\begin{aligned}
& P\left(S_{1} \cap S_{2} \cap \cdots \cap S_{k}\right) \\
& =P\left(S_{1}\right) P\left(S_{2} \mid S_{1}\right) P\left(S_{3} \mid S_{2} \cap S_{1}\right) \cdots P\left(S_{k} \mid S_{1} \cap S_{2} \cap \cdots \cap S_{k-1}\right)
\end{aligned}
$$

## Chain Rule

Two events are called independent if $P(A \cap B)=P(A) P(B)$, or equivalently, $P(A \mid B)=P(A)$. Intuitively, $A$ and $B$ are independent means that observing $B$ does not have any effect on the probability of $A$.

Bayes Rule

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

## Random Variables

Consider an experiment in which we flip 10 coins, and we want to know the number of coins that come up heads. Here, the elements of the sample space $\Omega$ are 10-length sequences of heads and tails. For example, we might have $\omega_{0}=\langle H, H, T, H, T, H, H, T, T, T\rangle \in \Omega$. However, in practice, we usually do not care about the probability of obtaining any particular sequence of heads and tails. Instead we usually care about real-valued functions of outcomes, such as the number of heads that appear among our 10 tosses, or the length of the longest run of tails. These functions, under some technical conditions, are known as random variables.

## Random Variables

Example: In our experiment above, suppose that $X(\omega)$ is the number of heads which occur in the sequence of tosses $\omega$. Given that only 10 coins are tossed, $X(\omega)$ can take
only a finite number of values, so it is known as a discrete random variable. Here, the

Discrete Random Variable probability of the set associated with a random variable $X$ taking on some specific value $k$ is $P(X=k):=P(\{\omega: X(\omega)=k\})$.

Example: Suppose that $X(\omega)$ is a random variable indicating the amount of time it takes for a radioactive particle to decay. In this case, $X(\omega)$ takes on a infinite number of possible values, so it is called a continuous random variable. We denote the probability that $X$ takes on a value between two real constants $a$ and $b$ (where $a<b$ ) as $P(a \leq X \leq b):=P(\{\omega: a \leq X(\omega) \leq b\})$.

## PDF \& CDFs

$$
F_{X}(x)=P(X \leq x) .
$$

By using this function, one can calculate the probability that $X$ takes on a value

$$
f_{X}(x)=\frac{d F_{X}(x)}{d x} .
$$ between any two real constants $a$ and $b$ (where $a<b$ ).

## Properties:

- $0 \leq F_{X}(x) \leq 1$.
- $\lim _{x \rightarrow-\infty} F_{X}(x)=0$.
- $\lim _{x \rightarrow+\infty} F_{X}(x)=1$.
- $x \leq y \Longrightarrow F_{X}(x) \leq F_{X}(y)$.


## Discrete random variables

- $X \sim \operatorname{Bernoulli}(p)$ (where $0 \leq p \leq 1$ ): the outcome of a coin flip $(H=1, T=0)$ for a coin that comes up heads with probability $p$.

$$
p(x)= \begin{cases}p, & \text { if } x=1 \\ 1-p, & \text { if } x=0\end{cases}
$$

- $X \sim \operatorname{Binomial}(n, p)$ (where $0 \leq p \leq 1$ ): the number of heads in $n$ independent flips of a coin with heads probability $p$.

$$
p(x)=\binom{n}{x} \cdot p^{x}(1-p)^{n-x}
$$

- $X \sim \operatorname{Geometric}(p)$ (where $p>0$ ): the number of flips of a coin until the first heads, for a coin that comes up heads with probability $p$.

$$
p(x)=p(1-p)^{x-1}
$$

- $X \sim \operatorname{Poisson}(\lambda)($ where $\lambda>0)$ : a probability distribution over the nonnegative integers used for modeling the frequency of rare events.

$$
p(x)=e^{-\lambda} \frac{\lambda^{x}}{x!}
$$

## Continuous random variables

- $X \sim \operatorname{Uniform}(a, b)$ (where $a<b$ ): equal probability density to every value between $a$ and $b$ on the real line.

$$
f(x)= \begin{cases}\frac{1}{b-a}, & \text { if } a \leq x \leq b \\ 0, & \text { otherwise }\end{cases}
$$

- $X \sim \operatorname{Exponential}(\lambda)$ (where $\lambda>0$ ): decaying probability density over the nonnegative reals.

$$
f(x)= \begin{cases}\lambda e^{-\lambda x}, & \text { if } x \geq 0 \\ 0, & \text { otherwise }\end{cases}
$$

- $X \sim \operatorname{Normal}\left(\mu, \sigma^{2}\right)$ : also known as the Gaussian distribution

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

## Expectation

Suppose that $X$ is a discrete random variable with PMF $p_{X}(x)$ and $g: \mathbb{R} \rightarrow \mathbb{R}$ is an arbitrary function. In this case, $g(X)$ can be considered a random variable, and we define the expectation or expected value of $g(X)$ as

$$
\mathbb{E}[g(X)]=\sum_{x \in \operatorname{Val}(X)} g(x) p_{X}(x)
$$

If $X$ is a continuous random variable with PDF $f_{X}(x)$, then the expected value of $\mathrm{g}(\mathrm{X})$ is defined as

$$
\mathbb{E}[g(X)]=\int_{-\infty}^{\infty} g(x) f_{X}(x) d x
$$

## Variance

The variance of a random variable $X$ is a measure of how concentrated the distribution of a random variable $X$ is around its mean. Formally, the variance of a random variable $X$ is defined as $\operatorname{Var}[X]=\mathbb{E}\left[(X-\mathbb{E}[X])^{2}\right]$.

$$
\begin{aligned}
& \mathbb{E}\left[(X-\mathbb{E}[X])^{2}\right] \\
& =\mathbb{E}\left[X^{2}-2 \mathbb{E}[X] X+\mathbb{E}[X]^{2}\right] \\
& =\mathbb{E}\left[X^{2}\right]-2 \mathbb{E}[X] \mathbb{E}[X]+\mathbb{E}[X]^{2} \\
& =\mathbb{E}\left[X^{2}\right]-\mathbb{E}[X]^{2}
\end{aligned}
$$

## Covariance

We can use the concept of expectation to study the relationship of two random variables with each other. In particular, the covariance of two random variables $X$ and $Y$ is defined as

$$
\operatorname{Cov}[X, Y]=\mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])]
$$

Using an argument similar to that for variance, we can rewrite this as

$$
\begin{aligned}
\operatorname{Cov}[X, Y] & =\mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])] \\
& =\mathbb{E}[X Y-X \mathbb{E}[Y]-Y \mathbb{E}[X]+\mathbb{E}[X] \mathbb{E}[Y]] \\
& =\mathbb{E}[X Y]-\mathbb{E}[X] \mathbb{E}[Y]-\mathbb{E}[Y] \mathbb{E}[X]+\mathbb{E}[X] \mathbb{E}[Y] \\
& =\mathbb{E}[X Y]-\mathbb{E}[X] \mathbb{E}[Y] .
\end{aligned}
$$

- $\operatorname{Var}[X+Y]=\operatorname{Var}[X]+\operatorname{Var}[Y]+2 \operatorname{Cov}[X, Y]$.
- If $X$ and $Y$ are independent, then $\operatorname{Cov}[X, Y]=0$.
- If $X$ and $Y$ are independent, then $\mathbb{E}[f(X) g(Y)]=\mathbb{E}[f(X)] \mathbb{E}[g(Y)]$.


## Slide Credits

- Stanford CS 229, Linear Algebra Review, Zico Kolter.
- Stanford CS 229, Probability Review, Maleki \& Do.

