# CS181 Section 0

Due: Never

The goal of these section notes is to cover some material that is mostly review for CS 181. There are a number of problems to test your understanding and readiness for the course. (*) indicates challenge sections or challenge problems. Do not worry if you cannot solve these problems as the corresponding material will not be necessary as prerequisites.

## 1 Linear Algebra

A great reference for this material is Sheldon Axler's *Linear Algebra Done Right*, which can be found on *Hollis*.

### 1.1 Scalars and Vectors

A **scalar** is a single element of the real numbers. $a \in \mathbb{R}$ is a scalar. We usually denote scalars using lowercase letters, such as $a$ or $x$.

A **vector** of $n$ dimensions is an ordered collection of $n$ coordinates, where each coordinate is a scalar. An $n$-dimensional vector $\mathbf{v}$ with real coordinates is an element of $\mathbb{R}^n$. Equivalently, the coordinates specify as single point in an $n$-dimensional space, just like you may have seen with cartesian coordiates where $(1, 3)$ might denote a point. By default, vectors will be columns and their transposes will be rows. We write vectors in bold lowercase, and the vector itself as a column of scalars:

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} x_1 & x_2 & \dots & x_n \end{bmatrix}^T.$$

This is the default format. Sometimes vectors will be in row form and their symbols may not be bolded. If you find this confusing at first please reach out to one of the course staff.
Vectors may be scaled. $a\mathbf{x}$ scales each element of $\mathbf{x}$ by scalar $a$ so that

$$a\mathbf{x} = \begin{bmatrix} ax_1 \\ ax_2 \\ \vdots \\ ax_n \end{bmatrix}.$$

Vectors of the same dimension may be added coordinate-wise:

$$\mathbf{x} + \mathbf{y} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} + \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_1 + y_1 \\ x_2 + y_2 \\ \vdots \\ x_n + y_n \end{bmatrix}.$$

Vectors have both a **direction** and a **magnitude**. The magnitude of a vector (or its length) is typically the vector's $\mathbf{L_2}$ norm, which can be computed as the square root of the sum of the squares of the coordinates:

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

There are a number of other vector norms such as the $\mathbf{L_1}, \mathbf{L_p}, \mathbf{L_\infty}$ norms:

$$||\mathbf{x}||_1 = \sum_{i=1}^{n} |x_i|,$$

$$||\mathbf{x}||_p = \left( \sum_{i=1}^{n} |x_i|^p \right)^{1/p},$$

$$||\mathbf{x}||_\infty = \max_{i=1,\ldots,n} |x_i|.$$

**Definition 1.1** (Norm)**.** We say that $\| \cdot \|$ is a norm if it satisfies the following properties:

- Triangle inequality: $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$.
- $\|a\mathbf{x}\| = |a| \cdot \|\mathbf{x}\|$ for a scalar $a$.
- $\|\mathbf{x}\| = 0$ if and only if $\mathbf{x} = \mathbf{0}$.

**Problem 1**
(*) *Challenge*: Show that the $\mathbf{L_p}$ norms are indeed norms for $p \in [1, \infty)$ and $p = \infty$. We will mostly work with $L_1$ and $L_2$ so it is recommended you understand these two norms.

The direction of a vector can be represented using a vector of magnitude one (according to some norm):

$$\hat{\mathbf{x}} = \frac{\mathbf{x}}{\|\mathbf{x}\|} = \begin{bmatrix} x_1/\|x\| \\ x_2/\|x\| \\ \vdots \\ x_n/\|x\| \end{bmatrix}.$$

We often use the "hat" symbol (i.e. $\hat{\mathbf{x}}$) to denote that a vector has magnitude one, or is a unit vector.
An important product between vectors of the same dimension is the **inner product** (also called dot product or scalar product). For two vectors $\mathbf{u}$ and $\mathbf{v}$, this is defined as

$$\mathbf{u} \cdot \mathbf{v} = \sum_{i=1}^{n} u_i v_i.$$

It is also written as $\langle \mathbf{u}, \mathbf{v} \rangle$. We can introduce **cosine similarity** through the formula

$$\cos(\theta) = \frac{\mathbf{u} \cdot \mathbf{v}}{||\mathbf{u}||_2 \, ||\mathbf{v}||_2},$$

where $\theta$ is the angle between $\mathbf{u}$ and $\mathbf{v}$. The cosine similarity ranges from -1 (exactly opposite) to 1 (exactly the same), with 0 indicating orthogonal vectors. If $\mathbf{v}$ is a unit vector then $\mathbf{u} \cdot \mathbf{v}$ gives us the magnitude of the projection of $\mathbf{u}$ onto the direction of $\mathbf{v}$. Thus it makes sense that a vector $\mathbf{u}$ dotted with itself equals the square of its L2 norm: $\langle \mathbf{u}, \mathbf{u} \rangle = ||\mathbf{u}||_2^2$.

The **outer product** between two vectors is the matrix $\mathbf{W} = [w_{ij}]_{i,j \leq n}$ whose entries are $w_{ij} = u_i v_j$. When the two vectors are dimension $n$ and $m$, respectively, their outer product is an $n \times m$ matrix.

## 1.2 Linear Independence

A set of vectors $\{\mathbf{v}_1, ..., \mathbf{v}_n\}$ is **linearly independent** if and only if the equation $c_1\mathbf{v}_1 + c_2\mathbf{v}_2 +, ..., +c_n\mathbf{v}_n = \mathbf{0}$ for scalars $c_1, ..., c_n$ can only be satisfied by setting $c_1, ..., c_n$ all to 0. Intuitively, it means that none of the vectors (or linear combinations of them) are parallel.

## 1.3 Spaces and Subspaces

A **vector space** $\mathcal{V}$ is a collection of vectors that follow several axioms regarding the properties of scaling and addition described above, and most importantly:

- $\mathbf{0} \in \mathcal{V}$
- closure under scaling: $\forall\ \mathbf{v} \in \mathcal{V}$ and scalars $a \in \mathbb{R}$, $a\mathbf{v} \in \mathcal{V}$
- closure under addition: $\forall\ \mathbf{u}, \mathbf{v} \in \mathcal{V}, \mathbf{u} + \mathbf{v} \in \mathcal{V}$

The most intuitive vector space and the one most relevant to the course is $\mathbb{R}^n$, the space of $n$-dimensional vectors. $\mathbb{R}^2$ is the 2-dimensional Cartesian plane for example.

Now we define a **basis** for a vector space. First, we define a **linear combination** of a list of vectors $(v_1, \ldots, v_m)$ as any quantity of the form:

$$a_1v_1 + \ldots + a_mv_m \text{ where } a_1, \ldots, a_m \in \mathbb{R} \tag{1}$$

The **span** of $(v_1, \ldots, v_m)$ is the set of all linear combinations of $(v_1, \ldots, v_m)$. Moreover, if the span of $(v_1, \ldots, v_m)$ is equal to the vector space $V$, then we say that $(v_1, \ldots, v_m)$ **spans** V.

Then a **basis** of a vector space $V$ is a list of vectors in $V$ that both are linearly independent and also span $V$. For the space $\mathbb{R}^n$, the most intuitive basis, which we call the **standard basis** is the list:

$$((1, 0, \ldots, 0), (0, 1, 0, \ldots, 0), \ldots, (0, \ldots, 0, 1)) \tag{2}$$

The set of vectors $\{\mathbf{v}_1, ..., \mathbf{v}_n\}$ form an **orthonormal basis** for $\mathcal{V}$ if they are all unit vectors (normal) and if $\langle \mathbf{v}_i, \mathbf{v}_j \rangle = 0, \forall\ i \neq j$ (orthogonal) where $\langle \cdot, \cdot \rangle$ is the inner product. The standard basis that we defined above is also an orthonormal basis.

The **dimension** of a vector space $V$ is the number of vectors of any basis of $V$. Since every basis of $V$ has the same number of vectors, this is uniquely defined.

Let $\mathcal{S}$ be a vector space. If $\mathcal{S} \subseteq \mathcal{V}$, then $\mathcal{S}$ is a **subspace** of $\mathcal{V}$. Intuitively, a subspace is a lower-dimensional space in a higher-dimensional space—think about the plane defined by the $x$ and $y$ axis in a 3-dimensional $x, y$ and $z$ space.

## 1.4 Scalar, Vector, and Subspace Projection

For vectors $\mathbf{u}, \mathbf{v} \in \mathcal{V}$ and $\mathbf{v} \neq \mathbf{0}$, the **scalar projection** $a$ of $\mathbf{u}$ onto $\mathbf{v}$ is computed as:

$$a = \frac{\langle \mathbf{u}, \mathbf{v} \rangle}{||\mathbf{v}||}$$

Think about this as the size of $\mathbf{u}$ along the direction of $\mathbf{v}$. Using scalar projection $a$, the **vector projection** $\mathbf{u}^{\parallel}$ of $\mathbf{u}$ onto $\mathbf{v}$ can be computed as:

$$\mathbf{u}^{\parallel} = a \cdot \frac{\mathbf{v}}{||\mathbf{v}||} = \frac{\langle \mathbf{u}, \mathbf{v} \rangle}{\langle \mathbf{v}, \mathbf{v} \rangle}\mathbf{v}.$$

Think about this as scaling by $a$ the unit vector in the direction of $\mathbf{v}$. For a projection onto $\mathbf{v}$, we can then write $\mathbf{u} = \mathbf{u}^{\parallel} + \mathbf{u}^{\perp}$, completing $\mathbf{u}$ with this new component $\mathbf{u}^{\perp}$. In particular, $\langle \mathbf{u}^{\parallel}, \mathbf{u}^{\perp} \rangle = 0$, and $\mathbf{u}^{\perp}$ is orthogonal to $\mathbf{v}$. It follows that $\mathbf{u} = \mathbf{u}^{\parallel}$ if and only if $\mathbf{u}$ is a scaled multiple of $\mathbf{v}$.

**Problem 2**
Verify that $\langle \mathbf{u}^{\parallel}, \mathbf{u}^{\perp} \rangle = 0$ and that $\mathbf{u} = \mathbf{u}^{\parallel}$ if and only if $\mathbf{u}$ is a scaled multiple of $\mathbf{v}$.

### 1.4.1 Subspace Projections

Finally, it is possible to project a vector $\mathbf{u}$ in a vector space $\mathcal{V}$ onto a subspace $\mathcal{S}$ of $\mathcal{V}$. If the set of vectors $\{\mathbf{s}_1, ..., \mathbf{s}_m\}$ form an orthonormal basis for $\mathcal{S}$, then the **subspace projection** $\mathbf{u}^{\parallel}$ of $\mathbf{u}$ onto $\mathcal{S} = \text{span}(\mathbf{s}_1, ..., \mathbf{s}_m)$ can be expressed as the sum of the projections of $\mathbf{u}$ onto each element of the basis of $\mathcal{S}$:

$$\mathbf{u}^{\parallel} = \sum_{i=1}^{m} \frac{\langle \mathbf{u}, \mathbf{s}_i \rangle}{\langle \mathbf{s}_i, \mathbf{s}_i \rangle} \mathbf{s}_i$$

This has the properties that the vector $\mathbf{u}^{\perp} = \mathbf{u} - \mathbf{u}^{\parallel}$ is orthogonal to all vectors in $\mathcal{S}$, that $\mathbf{u} = \mathbf{u}^{\parallel}$ if and only if $\mathbf{u} \in \mathcal{S}$, and that $\mathbf{u}^{\parallel}$ is the closest vector in $\mathcal{S}$ to $\mathbf{u}$: $||\mathbf{u} - \mathbf{v}|| > ||\mathbf{u} - \mathbf{u}^{\parallel}||, \forall\ \mathbf{v} \neq \mathbf{u}^{\parallel}, \mathbf{v} \in \mathcal{S}$.

**Problem 3** (Distance between a hyperplane and a point)
(*) *Challenge:* Suppose we have a hyperplane defined by $\mathbf{w}^T\mathbf{x} + w_0 = 0$. In this problem, we will derive the formula for the distance between the hyperplane and a point $\mathbf{x}'$.

(a) Imagine two points $\mathbf{x_1}$ and $\mathbf{x_2}$ on this hyperplane. Show that $\mathbf{w}$ is orthogonal to the difference $\mathbf{x_1} - \mathbf{x_2}$. Why does this imply that $\mathbf{w}$ is orthogonal to the hyperplane?

(b) Now, suppose we wish to find the distance $d$ between a point $\mathbf{x}'$ and the our hyperplane. Let $\mathbf{x}_p$ be the projection of $\mathbf{x}'$ onto the hyperplane. Find an expression for $\mathbf{x}'$ in terms of $d$, $\mathbf{w}$, and $w_0$. (*Hint*: use the fact from (a) that $\mathbf{w}$ is perpendicular to the hyperplane.)

(c) Using your expression from (b), show that the distance $d$ is the following:

$$d = \frac{\mathbf{w}^T\mathbf{x}' + w_0}{\|\mathbf{w}\|_2} \tag{3}$$

## 1.5 Matrices

A **matrix** is a rectangular array of scalars. Primarily, an $n \times m$ matrix $\mathbf{A} \in \mathbb{R}^{n \times m}$ is used to describe a **linear transformation** from $m$ to $n$ dimensions, where the matrix is an **operator**. To see this, note that the result of multiplying an $n \times m$ matrix and an $m \times 1$ vector is an $n \times 1$ vector. $A_{ij}$ is the scalar found at the $i^{th}$ row and $j^{th}$ column. We write matrices in bold uppercase.
A typical linear transformation looks like $\mathbf{y} = \mathbf{A}\mathbf{x}$ where $\mathbf{x} \in \mathbb{R}^m, \mathbf{y} \in \mathbb{R}^n, \mathbf{A} \in \mathbb{R}^{n \times m}$. The transformation $\mathbf{A}$ is linear because $\mathbf{A}(\lambda_1\mathbf{u} + \lambda_2\mathbf{v}) = \lambda_1\mathbf{A}\mathbf{u} + \lambda_2\mathbf{A}\mathbf{v}$ for scalars $\lambda_1$ and $\lambda_2$.

## 1.6 Matrix Multiplication Properties

**AB** is a valid **matrix product** if $\mathbf{A}$ is $p \times q$ and $\mathbf{B}$ is $q \times r$, or the left matrix has same number of columns $q$ as the right matrix has rows. The standard matrix product is defined as follow:

$$(\mathbf{AB})_{ij} = a_{i1}b_{1j} + a_{i2}b_{2j} + \cdots + a_{iq}b_{qj} = \sum_{k=1}^{q} a_{ik}b_{kj}; \quad i = 1, ..., p \text{ and } j = 1, ..., r.$$

In other words, $(\mathbf{AB})_{ij}$ is the dot product of the $i$th row of $\mathbf{A}$ with the $j$th column of $\mathbf{B}$.

Properties of matrix multiplication:

- Generally not commutative: $\mathbf{AB} \neq \mathbf{BA}$
- Left/Right Distributive over addition: $\mathbf{A}(\mathbf{B}+\mathbf{C}) = \mathbf{AB}+\mathbf{AC}$. $(\mathbf{A}+\mathbf{B})\mathbf{C} = \mathbf{AC}+\mathbf{BC}$.
- For some scalar $\lambda$: $\lambda(\mathbf{AB}) = (\lambda\mathbf{A})\mathbf{B} = (\mathbf{AB})\lambda = \mathbf{A}(\mathbf{B}\lambda)$.
- Transpose of product: $(\mathbf{AB})^\top = \mathbf{B}^\top\mathbf{A}^\top$

**Problem 4**

Given the matrix $\mathbf{X}$ and the vectors $\mathbf{y}$ and $\mathbf{z}$ below:

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{pmatrix} \quad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \quad \mathbf{z} = \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} \tag{4}$$

(a) Expand $\mathbf{Xy} + \mathbf{z}$

(b) Expand $\mathbf{y^T Xy}$

## 1.7 Rank, Determinant, Inverse

The **column rank** of a matrix $\mathbf{A}$ is the **dimension** of the vector space spanned by its column vectors, i.e., the number of linearly independent columns. The **row rank** is the dimension of the space spanned by its row vectors. A fundamental result in linear algebra is that the column rank and the row rank are always equal and this number is the **rank** of a matrix. If $\mathbf{A}$ is $n \times m$, then $rank(A) \leq min(n, m)$.

A matrix is **full rank** if its rank equals the largest possible for a matrix with the same dimensions, i.e. $\min(n, m)$. For a square matrix, full rank requires all its column (or row) vectors to be linearly independent.

The **determinant** $\det(\mathbf{A})$ is defined for a square matrix $\mathbf{A}$ and is a scalar quantity with various uses. Its computation differs for square matrices of different sizes. An $n$-by-$n$ square matrix may have an **inverse**. There is a matrix inverse if and only if $\mathbf{A}$ has a non-zero determinant. A square matrix that is not invertible is called **singular**. $\det(\mathbf{A})$ is also the product of the **eigenvalues** of $\mathbf{A}$ (see Section **??**). We will denote the determinant with single bars, e.g. $\det(X) = |\mathbf{X}|$. Do not confuse $|\mathbf{X}|$ with double bars $||\mathbf{X}||$, which typically denote a norm.

A few properties of the determinant (it's okay if you understand but can't recall from memory the rest of this section):

- The determinant of a diagonal matrix is the product of its diagonal values, and in particular the determinant of the **identity matrix I** is 1: $|\mathbf{I}| = 1$.
- For an $n \times n$-matrix $\mathbf{A}$ and a scalar value $c$ we have $|c\mathbf{A}| = c^n|\mathbf{A}|$.
- The determinant factors over products: $|\mathbf{AB}| = |\mathbf{A}| \cdot |\mathbf{B}|$.

The **inverse** $\mathbf{A}^{-1}$ of matrix operator $\mathbf{A}$ "undoes" $\mathbf{A}$ much like multiplying by $\frac{1}{x}$ undoes multiplying by $x$. We have $\mathbf{AA}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$. $\mathbf{A}^{-1}$ exists if and only if $|\mathbf{A}| \neq 0$. In general, matrix inversion is a complicated operation, but special cases that are easy to work with come up in the machine learning literature. Often analytical solutions to systems depend on the existence of the inverse of a matrix.

**Problem 5**

For an invertible matrix $\mathbf{A}$ show that $|\mathbf{A}^{-1}| = \frac{1}{|\mathbf{A}|}$.

(*) The **Moore-Penrose pseudoinverse** $\mathbf{A}^+$ of $\mathbf{A}$ is a generalization of the inverse to non-square matrices, where $\mathbf{A}\mathbf{A}^+\mathbf{A} = \mathbf{A}$. Matrix $\mathbf{A}\mathbf{A}^+$ may not be the general identity matrix but maps all column vectors of $\mathbf{A}$ to themselves.

## 1.8 Matrix Properties

- $\mathbf{A}^\top$ is the **transpose** of $\mathbf{A}$ and has $A^\top_{ji} = A_{ij}$. This is just like flipping the two dimensions of your matrix.
- $\mathbf{A}$ is **symmetric** if $A_{ij} = A_{ji}$. That is, $\mathbf{A} = \mathbf{A}^\top$. Only square matrices can be symmetric.
- (*) $\mathbf{A}$ is said to be **orthogonal** if its rows and its columns are orthogonal unit vectors. Consequence: $\mathbf{A}^\top\mathbf{A} = \mathbf{A}\mathbf{A}^\top = \mathbf{I}$ where $\mathbf{I}$ is the **identity matrix** (ones on the main diagonal and zeros elsewhere). For an orthogonal matrix $\mathbf{A}$ we have $\mathbf{A}^\top = \mathbf{A}^{-1}$.
- **Diagonal** matrices have non-zero values on the main diagonal and zeros elsewhere. Diagonal matrices are easy to take powers of because you just take the powers of the diagonal entries. Under certain conditions a matrix may be diagonalized, see **eigen-decomposition** and **SVD** below.
- A matrix is **upper-triangular** if the only non-zero values are on the diagonal or above (top right of matrix). A matrix is **lower-triangular** if the only non-zero values are on the diagonal or below (bottom left of matrix).

## 1.9 Eigen-Everything

Recall that a matrix $\mathbf{A}$ can be thought of as an operator. Each square matrix $\mathbf{A}$ has some set of vectors $\mathbf{x} \in \mathbb{R}^n$ in its domain that are simply mapped to a scaled version of the vector in the codomain. The matrix preserves the direction of these vectors: $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$ for some scalar value $\lambda$. In this case, $\lambda$ is an **eigenvalue** of $\mathbf{A}$ and $\mathbf{x}$ is a corresponding **eigenvector**. Eigenvectors can also be seen as the invariant directions of the matrix.

**Problem 6**

An *eigenspace* of a matrix $\mathbf{A}$ is an eigenvalue $\lambda$ and the set $U_\lambda = \{\mathbf{v} \mid A\mathbf{v} = \lambda\mathbf{v}\}$. Show that $U_\lambda$ is a vector subspace of the span of the columns of $A$.

**Eigen-decomposition**: Let $\mathbf{A}$ be an $n \times n$ full-rank matrix that has $n$ linearly independent eigenvectors $\{\mathbf{q}_i\}_{i=1}^n$. In this case, $\mathbf{A}$ can be factored into $\mathbf{A} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^{-1}$ where $\mathbf{Q}$ is $n \times n$ and has eigenvector $\mathbf{q}_i$ for its $i^{th}$ column. $\mathbf{\Lambda}$ is a diagonal matrix whose elements are the corresponding eigenvalues: $\Lambda_{ii} = \lambda_i$. This is the **eigen-decomposition** of the matrix and we say the matrix has been **diagonalized**. If a matrix $\mathbf{A}$ can be eigen-decomposed and none of its eigenvalues are 0, then $\mathbf{A}$ is **nonsingular** (i.e., it is **invertible**) and its inverse is given by $\mathbf{A}^{-1} = \mathbf{Q}\mathbf{\Lambda}^{-1}\mathbf{Q}^{-1}$ with $\mathbf{\Lambda}^{-1}_{ii} = \frac{1}{\lambda_i}$.

**Singular Value Decomposition** is a useful generalization of eigen-decomposition to rectangular matrices. Let $\mathbf{A}$ be an $m \times n$ matrix. Then $\mathbf{A}$ can be factored into $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{-1}$ where

- $\mathbf{U}$ is $m \times m$ and orthogonal. The columns of $\mathbf{U}$ are called the **left-singular vectors** of $\mathbf{A}$.
- $\mathbf{\Sigma}$ is an $m \times n$ diagonal matrix with non-negative real entries. The diagonal values $\sigma_i$ of $\mathbf{\Sigma}$ are known as the **singular values** of $\mathbf{A}$. These are also the square roots of the eigenvalues of $\mathbf{A}^\top\mathbf{A}$.
- $\mathbf{V}$ is an $n \times n$ orthogonal matrix. The columns of $\mathbf{V}$ are called the **right-singular** vectors of $\mathbf{A}$.

### 1.10 Positive Definiteness

The symmetric matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ is said to be **positive definite** if, for every non-zero vector $\mathbf{x} \in \mathbb{R}^n$, it satisfies the property

$$\mathbf{x}^\top \mathbf{A}\mathbf{x} > 0$$

and **positive semi-definite** if it satisfies

$$\mathbf{x}^\top \mathbf{A}\mathbf{x} \geq 0.$$

**Problem 7**
(*) Show that positive definite matrices have all eigenvalues $> 0$ and positive semi-definite matrices have all eigenvalues $\geq 0$.

## 2 Calculus

Khan Academy has good reference material for calculus and multivariable calculus. For matrix calculus see *The Matrix Cookbook* by Petersen and Pedersen, specifically sections 2.4, 2.6, and 2.7.

### 2.1 Differentiation

You should be familiar with single-variable differentiation, including properties like:

$$\begin{aligned}
\text{Chain rule: } & \frac{d}{dx} f(g(x)) = f'(g(x))g'(x) \\
\text{Product rule: } & \frac{d}{dx} f(x)g(x) = f'(x)g(x) + f(x)g'(x) \\
\text{Linearity: } & \frac{d}{dx}(af(x) + bg(x)) = af'(x) + bg'(x)
\end{aligned}$$

for scalars $a$ and $b$. In multivariable calculus, a function may have some number of inputs (say $n$) and some number of outputs (say $m$). In general, there is a partial derivative for every input-output pair. This is called the **Jacobian**. The $j^{th}$ column of the Jacobian is made up of the partial derivatives of $f_j$ (the $j^{th}$ output value of $\mathbf{f}$) with respect to all input elements, rows $i = 1$ to $n$.

$$\frac{d\mathbf{f}(\mathbf{x})}{d\mathbf{x}} = \begin{bmatrix} \frac{\partial f_1(\mathbf{x})}{\partial x_1} & \cdots & \frac{\partial f_m(\mathbf{x})}{\partial x_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_1(\mathbf{x})}{\partial x_n} & \cdots & \frac{\partial f_m(\mathbf{x})}{\partial x_n} \end{bmatrix}$$

If $f$ is scalar-valued (has only 1 output), its derivative is a column vector we call the **gradient vector**, written as $\nabla f$:

$$\nabla f = \frac{df(\mathbf{x})}{d\mathbf{x}} = \begin{bmatrix} \frac{\partial f(\mathbf{x})}{\partial x_1} \\ \frac{\partial f(\mathbf{x})}{\partial x_2} \\ \ldots \\ \frac{\partial f(\mathbf{x})}{\partial x_n} \end{bmatrix}$$

The gradient vector points in the direction of steepest ascent in $f(\mathbf{x})$. This is useful for optimization.

The **Hessian** matrix is like the Jacobian but with second-order derivatives. There are many interesting optimization topics related to the Hessian.

The most important vector or matrix derivatives that we will use in CS 181 can be found on p. 8-10 of *The Matrix Cookbook* by Petersen and Pedersen. We've reproduced a few important derivatives here:

$$\begin{aligned}
\frac{d\mathbf{x}^\top \mathbf{a}}{d\mathbf{x}} &= \frac{d\mathbf{a}^\top \mathbf{x}}{d\mathbf{x}} = \mathbf{a} \\
\frac{d\mathbf{a}^\top \mathbf{X}\mathbf{b}}{d\mathbf{X}} &= \mathbf{a}\mathbf{b}^\top \\
\frac{d\mathbf{a}^\top \mathbf{X}^\top \mathbf{b}}{d\mathbf{X}} &= \mathbf{b}\mathbf{a}^\top \\
\frac{d\mathbf{a}^\top \mathbf{X}\mathbf{a}}{d\mathbf{X}} &= \frac{d\mathbf{a}^\top \mathbf{X}^\top \mathbf{a}}{d\mathbf{X}} = \mathbf{a}\mathbf{a}^\top \\
\frac{d\mathbf{X}}{dX_{ij}} &= \mathbf{B}^{ij} \quad ***
\end{aligned}$$

*** $\mathbf{B}$ is a matrix with all zeros except for a 1 in the $i, j$ entry.

Have you ever wondered how to differentiate the norm of a matrix? The eigenvalues? For more, see *The Matrix Cookbook*.

## 2.2 Optimization

**Local Extrema**: Recall that the local extrema of a single-variable function can be found by setting its derivative to 0. The same is true here, using the condition $\frac{d\mathbf{f}(\mathbf{x})}{d\mathbf{x}} = \mathbf{0}$. However, this equation is often intractable. We can also search for local minima numerically using gradient-based methods.

**Gradient Descent** (we will learn this in class): We start with an initial guess at at a useful value for a parameter $\mathbf{w}$: $\mathbf{w}_0$. Then at each step $i$ we update our guess by going in the direction of greatest descent of a loss function (opposite the direction of the gradient vector):

$$\mathbf{w}_{i+1} = \mathbf{w}_i - \eta \frac{df(\mathbf{w})}{d\mathbf{w}}$$

where $\eta > 0$ is the *step size*. We stop updating $\mathbf{w}_i$ when the value of the gradient is close to 0.

**Lagrange Multipliers**: This technique is used to optimize a function $f(\mathbf{x})$ given some constraint $g(\mathbf{x}) = 0$. First construct what is called the **Lagrangian function** $L(\mathbf{x}, \lambda)$:

$$L(\mathbf{x}, \lambda) = f(\mathbf{x}) + \lambda g(\mathbf{x})$$

Then, set the derivative of $L$ with respect to both $\mathbf{x}$ and $\lambda$ equal to 0:

$$\nabla L_{\mathbf{x}} = \nabla f(\mathbf{x}) + \lambda \nabla g(\mathbf{x}) = 0, \qquad \frac{\partial L}{\partial \lambda} = g(\mathbf{x}) = 0$$

If $\mathbf{x}$ is $d$-dimensional, this will give you a system of $d + 1$ equations. In this way, you can solve analytically for $\mathbf{x}$ to find the optimal value of $f(\mathbf{x})$ subject to the constraint $g(\mathbf{x})$. As with unconstrained optimization, this too is intractable and gradient descent is used to make progress.

**Problem 8**

Solve the following vector/matrix calculus problems.

(a) Let $f(\mathbf{x}) = \mathbf{x}^T\mathbf{x}$. Find $\nabla f(\mathbf{x})$.

(b) Let $f(\mathbf{w}) = (1 - \mathbf{w}^T\mathbf{x})^2$. Find $\nabla f(\mathbf{w})$ where the gradient is taken with respect to $\mathbf{w}$.

(c) Let $f(\mathbf{x}) = g(h(\mathbf{x}))$, where $g : \mathbb{R} \to \mathbb{R}$ and $h : \mathbb{R}^d \to \mathbb{R}$ are both differentiable. Find $\nabla f(\mathbf{x})$.

(d) Let $\mathbf{A}$ be a symmetric $n$-by-$n$ matrix. If $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x} + \mathbf{w}^T\mathbf{x}$, find $\nabla f(\mathbf{x})$.

# 3 Probability Theory

Harvard's Statistics 110 is the most relevant source of material. A public version of the course can be found here.

## 3.1 Probability

Probability provides a measure of how likely it is that some *event* will occur or that some proposition will be true. In order to define probability, we define the *sample space* $\Omega$ as the set of events that represent all possible outcomes of a particular process, or *experiment*. For example, if you are flipping a coin, the sample space consists of two outcomes: $S = \{H, T\}$. We then represent the "probability of getting a head" as $P(H) = \frac{1}{2}$. Formally, we define $P(A)$ as a function that maps events to probability values between 0 and 1. (You don't need to know the formal mathematical underpinnings behind probability theory for CS 181, but if you're interested, see optional Section 3.2 below.)

Below are some of the fundamental concepts and formulas in probability that will be important for this course:

- **Conditional probability**: $P(A|B)$ represents the probability of $A$ given $B$. In other words, what is the probability that $A$ is true given that we already know that $B$ is true.

  For example, an unconditional probability might be the probability of arriving late to work today, $P$(late to work). A conditional probability would be the probability of arriving late to work, given that you know that traffic is bad, $P$(late to work|bad traffic). Intuitively, think of conditioning as updating your probabilities based on what you already know.

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

- **Independence**: Events $A$ and $B$ are independent if knowing whether $A$ occurred gives no information about whether $B$ occurred. If $A$ and $B$ are independent, then all the following hold:

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

- **Probability of intersection or union**: Note that $P(A \cap B)$ means the probability of $A$ *and* $B$, and

$P(A \cup B)$ means the probability of $A$ *or* $B$:

$$P(A \cap B) = P(A)P(B|A) \quad \text{if } A \text{ and } B \text{ not independent}$$
$$P(A \cap B) = P(A)P(B) \quad \text{if } A \text{ and } B \text{ independent}$$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \quad \text{if } A \text{ and } B \text{ not disjoint}$$
$$P(A \cup B) = P(A) + P(B) \quad \text{if } A \text{ and } B \text{ disjoint}$$

- **Bayes Rule**: Fundamental rule for calculating conditional probabilities:

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

- **Law of Total Probability**: Useful rule for calculating the probability of an event $A$ in situations where it's easier to calculate probabilities conditioned on some other event.

  In calculating $P(A)$, we make some *partition* of the sample space $B_1, B_2, B_3, ..., B_n$ (i.e. the $B_i$ are disjoint and their union is the entire sample space). Then we can write:

$$P(A) = \sum_{i}^{n} P(A|B_i)P(B_i) = P(A|B_1)P(B_1) + \ldots + P(A|B_n)P(B_n)$$

  In the case where the partition is simply $B$ and $B^c$, then:

$$P(A) = P(A|B)P(B) + P(A|B^c)P(B^c)$$

**Problem 9** (Example 2.3.9 from the *Stat 110* textbook)

A patient named Fred is tested for a disease called conditionitis, a medical condition that afflicts 1% of the population. The test result is positive, i.e., the test claims that Fred has the disease. Let $D$ be the event that Fred has the disease and $T$ be the event that he tests positive.

Suppose that the test is "95% accurate." What that means is $P(T|D) = 0.95$ and $P(T^c|D^c) = 0.95$. Find the conditional probability that Fred has conditionitis, given his positive test result.

### 3.1.1 (*) Formal definition of probability

The axioms of probability are a special case of general measure theory. This section will introduce some of that formalism, but this will not be necessary for CS181. In general, we will work with a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ where $\Omega$ is the sample space (the set of all possible outcomes), $\mathcal{F}$ is a set of events (subsets of $\Omega$), and $\mathbb{P}$ is a probability measure. $\Omega$ is an abstract set consisting of whatever we deem to be 'outcomes' (this could be a set of numbers, environmental states such as the weather, or really anything). $\mathcal{F}$ is a $\sigma$-algebra on $\Omega$, which means it satisfies certain axioms:

- $\emptyset \in \mathcal{F}$ and $\Omega \in \mathcal{F}$
- If $A \in \mathcal{F}$ then $A^c = \Omega \setminus A \in \mathcal{F}$
- If $A_1, A_2, ... \in \mathcal{F}$ then $\bigcup_{n=1}^{\infty} A_n \in \mathcal{F}$.

These axioms may seem arbitrary, but they are all quite intuitive to probability. Firstly, $\emptyset \in \mathcal{F}$ says that the event that *nothing* happens is in $\mathcal{F}$ and $\Omega \in \mathcal{F}$ says that the event that *something* happens is in $\mathcal{F}$. The second condition tells us that if event $A \subset \Omega$ is in $\mathcal{F}$ then the *opposite* of $A$, which is everything but $A$ i.e. $A^c$, is also in $\mathcal{F}$. If some event could happen then surely that event could not happen, which is the same as the opposite of said event occurring. The third condition tells us that for a countable collection of possible events $A_1, A_2, ...$ the event $\bigcup_{n=1}^{\infty} A_n$, which is the logical equivalent of $A_1$ OR $A_2$ OR $A_3$ OR ..., is also possible.

Now we are left with our probability measure $\mathbb{P} : \mathcal{F} \to [0, 1]$ which is a function from our collection of events to the interval $[0, 1]$. We have $\mathbb{P}(\emptyset) = 0$, which is to say that the probability of nothing happening is 0, and $\mathbb{P}(\Omega) = 1$, which says that the probability of something happening is 1. $\mathbb{P}(A) + \mathbb{P}(\Omega \setminus A) = 1$, which means that one of $A$ or $\Omega \setminus A$ will occur. $\mathbb{P}$ satisfies countable subadditivity:

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) \leq \sum_{n=1}^{\infty} \mathbb{P}(A_n)$$

and countable additivity for disjoint sets:

$$A_i \cap A_j = \emptyset \ \forall i \neq j \implies \mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mathbb{P}(A_n).$$

**Problem 10**
(*) Show that if $A \subset B$ then $\mathbb{P}(A) \leq \mathbb{P}(B)$. What does this mean?

Let's give an explicit example of a probability space. Consider a fair flip of a coin, which lands on Heads (H) or Tails (T) with equal probability. Then our set of outcomes is $\Omega = \{H, T\}$. What is our collection of events? It is the *power set* of $\Omega$ (the collection of all subsets of $\Omega$): $\mathcal{F} = 2^{\Omega} = \{\emptyset, \{H\}, \{T\}, \{H, T\}\}$. And what is our probability measure? It is defined as follows:

- $\mathbb{P}(\emptyset) = 0$
- $\mathbb{P}(\{H\}) = \mathbb{P}(\{T\}) = 1/2$
- $\mathbb{P}(\{H, T\}) = 1$

*Note*: For brevity we generally drop the set brackets and write $\mathbb{P}(H)$ instead of $\mathbb{P}(\{H\})$.

## 3.2 Random Variables

A **random variable** is a variable whose value is determined randomly as the result of some kind of random process or experiment. For example, suppose that for an experiment you flip a coin ten times. A random variable is any outcome that results from this experiment. Random variables do not need to be numerical. For example, they could represent the result of choosing a molecule according to some sample distribution from a set of interesting molecules. Here are some examples of numerical random variables:

- $X =$ the number of heads
- $X =$ the number of tails
- $X =$ the number of heads minus the number of tails
- $X = \begin{cases} 1, & \text{if the 5th flip is heads} \\ 0, & \text{if the 5th flip is tails} \end{cases}$

Remember that the value of a random variable $X$ is unknown before we observe the outcome of the experiment. Once we observe the experiment's outcome (for example, we actually flip the coin ten times), we then have an observation which we denote using $x$.

A random variable can be *discrete* or *continuous*. A discrete random variable $X$ takes one of a finite set of values in the sample space $\Omega$, each with a corresponding probability $p(X = x)$ (or written $p_X(x)$ or simply $p(x)$), for event $x \in \Omega$. $p(x)$ is the **probability mass function** of $X$. We say that $x \sim X$ ($x$ is sampled from $X$) when the value of $x$ is picked in accordance with the distribution on $X$.

A continuous random variable can take on a continuous range of values. We use $p(x)$ or $p_X(x)$ for the **probability density function** of a continuous random variable ($f(x)$ is also common notation). For a continuous r.v., it's important to note that the probability of any particular value is zero (but the probability of a range of values can be nonzero). It's important to think of the function as assigning densities that behave like *relative probabilities* rather than absolute masses. Among other things, the probability density function (PDF) $p(x)$ can be greater than 1. We can easily go between the probability density and probability using integration:

$$\mathbb{P}(A) = \int_{x \in A} p(x)\ dx.$$

### 3.2.1 Expectation

The **expected value** (or *expectation, mean*) of a numerical random variable can be thought of as the "weighted average" of the possible outcomes of the random variable. For discrete random variables:

$$\mathbb{E}_{x \sim p(x)}[X] = \sum_{x \in \Omega} x \cdot p(x) \qquad \mathbb{E}[g(X)] = \sum_{x \in \Omega} g(x)p(x)$$

where $g : \Omega \to \mathbb{R}$. Note that we often drop the subscript underneath the $\mathbb{E}$. For a numerical, continuous random variable:

$$\mathbb{E}[X] = \int_{x \in \Omega} x \cdot p(x)dx \qquad \mathbb{E}[g(X)] = \int_{x \in \Omega} g(x)p(x)dx$$

The most important property of expected values is the **linearity of expectation**. For **any** two random variables $X$ and $Y$ (regardless of independence)

$$\mathbb{E}[aX + bY + c] = a\mathbb{E}[X] + b\mathbb{E}[Y] + c.$$

If $X$ and $Y$ are independent, then $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$.

### 3.2.2 Variance

The variance of a numerical random variable is its expected squared deviation from its mean:

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

Variance is a measure of the spread of a random variable. Random variables with high variance are more spread out. Consider two normal distributions with different variances. The distribution with low variance is tall and skinny, and the distribution with high variance is shorter and wider.

**Problem 11**

An example of a discrete distribution $X$ is the result from rolling a standard, fair 6-sided die.

(a) What is the set of outcomes $\Omega$?

(b) Calculate $\mathbb{E}(X)$ and $\text{Var}(X)$

**Problem 12**
Verify that $\text{Var}(aX + b) = a^2\text{Var}(X)$.

### 3.2.3 Independence of random variables

Two random variables $X, Y$ are said to be *independent* if $p(x, y) = p(x)p(y)$. This tells us that knowing $X$ tells us nothing about $Y$ and vice-versa. Independence is often denoted using the $\perp$ symbol, where $X \perp Y$ implies $X$ is independent of $Y$.

**Problem 13**
Show that if $X$ and $Y$ are independent then $p(x|y) = p(x)$. Interpret this.

**Definition 3.1** (Independent and Identically Distributed)**.** We say that random variables $X_1, X_2, ...$ are *independent and identically distributed* (often abbreviated as i.i.d. or iid) if $X_i \sim_p X$ (each $X_i$ is sampled from the same distribution $p$) and $X_i$ is independent of $X_j$ for $i \neq j$.

### 3.2.4 Conditional Independence

Two random variables $X, Y$ are said to be *conditionally independent* given another random variable $Z$ if $p(x, y|z) = p(x|z)p(y|z)$. This tells us that if given $Z$, then knowing $X$ tells us nothing about $Y$ and vice-versa (given $Z$, knowing $Y$ tells us nothing about $X$).

**Problem 14**
Does independence imply conditional independence? Does conditional independence imply independence?

### 3.2.5 Joint Distributions

The **joint probability distribution** of $X = x$ and $Y = y$ is written as $p(x, y)$ or $p_{X,Y}(x, y)$. For independent random variables $X$ and $Y$, the joint distribution factors into the product $p(x, y) = p(x)p(y)$. However, in the more general case we must **condition**: $p(x, y) = p(x)p(y|x) = p(y)p(x|y)$ (see next section). When you have a joint distribution of two or more random variables, its a common situation to want the **marginal distribution** of a single variable. For a pair of random variables $X$ and $Y$, use the **sum rule**:

$$\text{Discrete: } p(x) = \sum_{y \in \mathcal{Y}} p(x, y)$$

$$\text{Continuous: } p(x) = \int_{y \in \mathcal{Y}} p(x, y)dy$$

Think about the marginal distribution as that you would obtain by running an experiment, sampling both r.v.s, but only recording the observations on one of them. This generalizes. For example, with four r.v.s then the marginal distribution on two of them is attained by "summing out" over the other two.

### 3.2.6 Conditional Distributions

Receiving information about the value of a random variable $Y$ can change the distribution of another variable $X$. For example, if you know the 49ers won the Super Bowl (!) and their opponents scored 20 points, then you know that the 49ers scored at least 21 points. We write the conditional random variable as $X|Y$, and the

conditional distribution as $p(x|y)$. Manipulating the definition for the joint probability of random variables that may be dependent, we get:

$$p(x|y) = \frac{p(x,y)}{p(y)}$$

As mentioned above, when dealing with the joint probability of several dependent variables, we can factor into chains of conditional probabilities with the **product rule**:

$$\begin{aligned} p(x,y,z) &= p(x)p(y|x)p(z|x,y) \\ &= p(y)p(x|y)p(z|x,y) \\ &= p(z)p(x|z)p(y|x,z) \\ &= \text{etc...} \end{aligned}$$

This is a very useful tool. Optional: See `http://colah.github.io/posts/2015-09-Visual-Information/` for some interesting visualizations of conditional probability and information theory.

### 3.2.7 Bayes' Theorem for Conditional Distributions

This is a central theorem that we will use repeatedly in this course, and is an extension of the product rule:

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}.$$

**Problem 15**
Prove Bayes' theorem.

(*) Since we are conditioning on $y$, then $y$ is held constant, and that means $p(y)$ is just a normalization constant. As a result, we often write the above property as

$$p(x|y) \propto p(y|x)p(x)$$

where the symbol $\propto$ can be interpreted as "is proportional to".

### 3.2.8 Covariance

The **covariance** between two jointly distributed, numerical random variables $X$ and $Y$ with finite variances is defined as the expected product of their deviations from their individual expected values. Intuitively, this asks: are $X$ and $Y$ likely to tend above $\mathbb{E}[X]$ and $\mathbb{E}[Y]$ jointly (high covariance)? Or does $X$ tend below $\mathbb{E}[X]$ while $Y$ tends above $\mathbb{E}[Y]$ and vise versa (low covariance)? To oversimplify, do $X$ and $Y$ tend to increase and decrease together?

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

When considering data in $n$ dimensions, compute the $n \times n$ **covariance matrix** (often denoted $\mathbf{\Sigma}$), where $\mathbf{\Sigma}_{ij} = \text{Cov}(X_i, X_j)$ is the empirical covariance between the $i^{th}$ and $j^{th}$ features ("empirical" in the sense that it is calculated from sample data).

Properties of covariance: (supposing $X, Y, Z$ have mean 0 and finite variances)

- Symmetric: $\text{Cov}(X, Y) = \text{Cov}(Y, X)$
- Positive Semi-definite: $\text{Cov}(X, X) \geq 0$
- $\text{Cov}(X, X) = 0$ implies $X$ always takes the same value, its mean
- Bilinear: $\text{Cov}(aX + bY, Z) = a\text{Cov}(X, Z) + b\text{Cov}(Y, Z)$
- Triangle Inequality: $|\text{Cov}(X, Y)| \leq \sqrt{\text{Var}(X)\text{Var}(Y)}$

**Problem 16**
(*) Prove these five properties. The last one is tough!

**Problem 17**
Show that for random variables $X, Y$ that $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y)$.

**Problem 18**
(*) Show that for random variables $X_1, ..., X_n$ that

$$\text{Var}(X_1 + \cdots + X_n) = \sum_{i=1}^{n} \text{Var}(X_i) + 2 \sum_{1 \leq i < j \leq n} \text{Cov}(X_i, X_j).$$

*Hint:* Use induction and the problem above.

## 3.3 Conditional Expectation and Conditional Variance

$\mathbb{E}[X|Y = y]$ is the expected (or average) value of the random variable $X$ given a particular observed value of $Y$ (such as the expected temperature, given no rain). This is the **conditional expectation** of $X$ given $Y = y$.

Similarly, we can define **conditional variance** as

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

**Adam's law** (law of total/iterated expectations) gives

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Y]]$$

**Problem 19**
(*) Prove Adam's law. This is quite tough so feel free to look it up on Wikipedia if needed.

**Eve's Law** (or law of total variance) is the analogous case for variance:[1]

$$\text{Var}[X] = \mathbb{E}[\text{Var}[X|Y]] + \text{Var}[\mathbb{E}[X|Y]]$$

**Problem 20**
Prove Eve's law using Adam's law.

[1] These two components are also the source of the term "Eve's law", from the initials EV VE for "expectation of variance" and "variance of expectation".

## 3.4 Gaussians (Normal Distribution)

### 3.4.1 Univariate PDF

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

The univariate Gaussian is often referred to as a bell curve. The 'bell' corresponding to the PDF above is centered at the mean $\mu$ and has width proportional to the variance $\sigma^2$.

**Problem 21**

Using the probability density function of $X \sim \mathcal{N}(0,1)$ show that $X$ has mean 0 and variance 1.

*Hint*: The PDF is $p(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right)$. For the mean, you can reason about the properties of the PDF itself to get the answer without integration techniques. For the variance, use integration by parts and the fact that the PDF itself integrates to 1.

**Problem 22**

Solve the following problems:

(a) Let $Z \sim \mathcal{N}(0,1)$. Find a random variable in terms of $Z$ that has the distribution $\mathcal{N}(-2,4)$.

(b) (*) Show that in general, if $X \sim \mathcal{N}(\mu, \sigma^2)$ then $aX + b \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$.

Properties of Gaussians:

- If $X, Y$ are independent normals then $X + Y \sim \mathcal{N}(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$
- Any PDF proportional to $\exp(ax^2 + bx + c)$ must be a Gaussian PDF.

### 3.4.2 Multivariate PDF

Given dimension $m$, mean vector $\mu \in \mathbb{R}^m$, and covariance matrix $\mathbf{\Sigma} \in \mathbb{R}^{m \times m}$, we say that $\mathbf{x}$ is distributed multivariate normal if its PDF is given by:

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

Note: there are many ways to write the multivariate normal PDF. You may notice the absence of $m$ in the coefficient. This works because the $2\pi$ distributes nicely over $\mathbf{\Sigma}$ in the determinant.

## 3.5 Markov Chains

A sequence of random variables $X_1, X_2, ...$ is said to be a *Markov chain* if it satisfies the *Markov property*:

$$X_{n+1}|X_1, ..., X_n \sim X_{n+1}|X_n,$$

i.e. that knowing the value of $X_n$ tells you the same amount of information about $X_{n+1}$ as knowing all of $X_1, ..., X_n$. If the $X_i$'s are a discrete distribution the Markov property can be written as:

$$\mathbb{P}(X_{n+1} = j_{n+1}|X_n = j_n, ..., X_1 = j_1) = \mathbb{P}(X_{n+1} = j_{n+1}|X_n = j_n).$$

**Problem 23**

A simple random walk is defined by setting $X_0 = 0$ and letting $X_{i+1} = X_i + R_i$ where the $R_i$'s are independent variables taking value $+1$ or $-1$ with equal probability: $\mathbb{P}(R_i = 1) = \mathbb{P}(R_i = -1) = 1/2$. Show that simple random walk is a Markov chain.

An $m$-state Markov chain consists of a Markov chain $X_1, ..., X_n$ for which $X_i \in \{1, ..., m\}$, which is the set of states. Don't confuse the set of $m$ states with the length of the chain. Such a Markov chain is typically defined with a matrix of probabilities. Let $p_{ij} = \mathbb{P}(X_{n+1} = j | X_n = i)$ be the *transition probability* from state $i$ to state $j$. Note that $p_{ij} \neq p_{ji}$ in general. The *transition matrix* defines the properties of the Markov chain:

$$P = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1m} \\ p_{21} & p_{22} & \cdots & p_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ p_{m1} & p_{m2} & \cdots & p_{mm} \end{pmatrix}.$$

A distribution on this Markov chain is a row vector $\pi = (q_1, ..., q_m)$ with $q_i \geq 0$ and $\sum_{i=1}^{m} q_i = 1$ and where $q_i$ represents the probability of being at state $i$. $\pi P$ represents the distribution after 1 sample from the Markov chain and $\pi P^k$ represents the distribution after $k$ samples from the Markov chain. If $\pi^* P = \pi^*$ then we say $\pi^*$ is a **stationary distribution** of Markov chain $P$.

**Problem 24**

Consider the following Markovian environment. The weather is either sunny (state 1) or rainy (state 2). If it's sunny today it will be sunny tomorrow with probability 0.7 and if it's rainy today it will be rainy tomorrow with probability 0.5.

- What is the transition matrix $P$ for this environment?
- What is the corresponding stationary distribution?
- Let's say we start with the initial distribution $\begin{pmatrix} 0.5 & 0.5 \end{pmatrix}$. What does the distribution look like after 1 sample of the Markov chain? After 2? After 10? Please feel free to use a computer algebra system like WolframAlpha. Do you see any similarities with the stationary distribution? Check out the concept of the *"limiting distribution"* of a Markov chain if you're interested.

## 3.6 Inference

We will not need too much inference as a prerequisite, but the more familiar you are with basic concepts such as the likelihood function, the easier it will be to pick up the methods in CS 181. If you would like to dig deeper take a look at Harvard's Statistics 111. Some section notes from a previous iteration of the course can be found here.

### 3.6.1 The Likelihood Function

Let $Y$ be a probability distribution that is dependent on some set of **parameters** $\theta$. The parameters may be the result of using machine learning to try to model a distribution on data that has been observed. For example, if you're familiar with logistic regression (and not to worry if not), the parameters of the logistic regression. Write $f(y|\theta)$ to denote the probability mass function if $Y$ is discrete and the probability density function if $Y$ is continuous. Let $D$ denote the *data*, a set of $n$, independent and identically distributed (i.i.d.) samples $y_1, ..., y_n \sim_f Y$. The **likelihood function** on data $D$ is defined as

$$L(\theta; D) = f(y_1, .., y_n|\theta) = \prod_{i=1}^{n} f(y_i|\theta).$$

Often times, to transform the product into a sum we take the log of the likelihood $L(\theta)$, called the log-likelihood $l(\theta)$:

$$\ell(\theta; D) = \log L(\theta; D) = \log \left( \prod_{i=1}^{n} f(y_i|\theta) \right) = \sum_{i=1}^{n} \log f(y_i|\theta).$$

It is convenient to work with the log to the base $e$.

**Problem 25**

(*) Derive the likelihood and log-likelihood functions for i.i.d. samples $y_1, ..., y_n \sim \mathcal{N}(\mu, \sigma^2)$.

### 3.6.2 (*) Maximum Likelihood Estimation

What $L(\theta; D)$ represents is a product of densities on elements of data $D$ as a function of the parameters $\theta$, where we think about parameters $\theta$ as being selected through a machine learning procedure. Intuitively, $L(\theta_1; D) > L(\theta_2; D)$ implies that $\theta_1$ is a better set of parameters than $\theta_2$ for modeling data $D$. Maximizing the likelihood will then produce the best set of parameters according to the data. This is referred to as **maximum likelihood estimation**.

The *maximum likelihood estimate (MLE)* is the set of parameters $\theta$ that maximizes $L(\theta; D)$ given data $D$ (or equivalently the set of parameters that maximizes $\ell(\theta; D)$ since $L(\theta_1; D) > L(\theta_2; D) \iff \ell(\theta_1; D) > \ell(\theta_2; D)$). There is a rich theory underlying the MLE with rigorous convergence results and guarantees. We will not investigate this theory here but there are many good references available online and this is part of what will be covered in CS 181.

**Problem 26**

(*) Compute the MLE estimates for i.i.d. samples $y_1, ..., y_n \sim \mathcal{N}(\mu, \sigma^2)$.

## 3.7 Bayesian Inference

In this course, we will use Bayesian inference. While thorough knowledge of Bayesian inference is not required coming into the course, it will be helpful to have at least a general idea of the principles of Bayesian inference.

Suppose that we have a weighted coin which lands Heads with probability $p$, where we don't know the value of $p$. Our goal is to infer the value of $p$ after observing $n$ coin flips. In frequentist inference, $p$ is treated as a constant whose value we do not know, and therefore does not have a distribution. Therefore, statements such as "the probability that $p$ is greater than 0.5" are not logical since $p$ does not have a distribution. It is either greater than 0.5 or it isn't.

In Bayesian inference, $p$ is treated as a random variable and therefore has a distribution. In Bayesian inference, the distribution of a random variable can represent our belief or uncertainty about that random variable. We know that since $p$ is a probability, it must have support $[0, 1]$. Thus, we might suppose initially that $p \sim Beta(a, b)$ This initial distribution we give to $p$ is a **prior distribution**.

Suppose then we observe the value of $X$, the outcome of $n$ coin flips, i.e. $X \sim Bin(n, p)$. The distribution of the data given the parameter, $f(X|p)$ is called the **likelihood**. Suppose we observe $X = x$. Then we are interested in the **posterior distribution** of $p|X = x$, the distribution of $p$ given that we observe $X = x$. The posterior reflects our updated belief about $p$ given the data that we observe, and is given by Bayes Rule:

$$f(p|X) = \frac{f(p)f(X|p)}{f(X)} \tag{5}$$

As you can see, the posterior includes both the prior $f(p)$ and the likelihood $f(X|p)$ distributions. Intuitively, the posterior combines our prior belief as well as the data that we observe.

### 3.7.1 Conjugate priors

To find the posterior, we can use Bayes Rule as above. However, in certain cases, the posterior distribution is part of the same distribution family as the prior. In that case, the prior is called a **conjugate prior**. For example, in this case, the beta distribution is the conjugate prior for the binomial distribution. If $p \sim Beta(a, b)$ and $X|p \sim Bin(n, p)$, then the posterior is also Beta. Specifically, $p|X = x \sim Beta(a+x, b+n-x)$.

### 3.7.2 Normal-Normal Conjugacy

Let $x|\mu \sim \mathcal{N}(\mu, \sigma^2)$ and $\mu \sim \mathcal{N}(\mu_0, \sigma_0^2)$. Then the posterior distribution $\mu|x$ also follows a Gaussian distribution. Because of this we say that the Gaussian distribution is *self-conjugate*.

(*) Proof: First, note that for $a > 0$ and $b, c \in \mathbb{R}$ we have

$$\int_{-\infty}^{\infty} \exp(-ax^2 + bx + c)\ dx = \sqrt{\frac{\pi}{a}} \exp\left(\frac{b^2}{4a} + c\right).$$

We will now show that

$$\mu|x \sim \mathcal{N}\left(\frac{\mu_0\sigma^2 + x\sigma_0^2}{\sigma^2 + \sigma_0^2}, \left(\frac{1}{\sigma^2} + \frac{1}{\sigma_0^2}\right)^{-1}\right).$$

We know by Bayes' theorem that

$$p(\mu|x) = \frac{p(x|\mu)p(\mu)}{p(x)}.$$

We have

$$p(x|\mu) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right), \quad p(\mu) = \frac{1}{\sigma_0\sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{(\mu-\mu_0)^2}{\sigma_0^2}\right).$$

By LOTP we have

$$p(x) = \int_{-\infty}^{\infty} p(x|\mu')p(\mu')\ d\mu' = \frac{1}{2\pi\sigma\sigma_0}\int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}\left(\frac{(x-\mu')^2}{\sigma_0^2} + \frac{(\mu'-\mu_0)^2}{\sigma_0^2}\right)\right)\ d\mu' =$$

$$\frac{1}{2\pi\sigma\sigma_0}\int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}\left(\frac{1}{\sigma^2} + \frac{1}{\sigma_0^2}\right)(\mu')^2 + \left(\frac{x}{\sigma^2} + \frac{\mu_0}{\sigma_0^2}\right)\mu' - \frac{1}{2}\left(\frac{x^2}{\sigma^2} + \frac{\mu_0^2}{\sigma_0^2}\right)\right)\ d\mu' =$$

$$\frac{1}{2\pi\sigma\sigma_0}\sqrt{\frac{\pi}{\frac{1}{2}\left(\frac{1}{\sigma^2} + \frac{1}{\sigma_0^2}\right)}} \exp\left(\frac{\left(\frac{x}{\sigma^2} + \frac{\mu_0}{\sigma_0^2}\right)^2}{2\left(\frac{1}{\sigma^2} + \frac{1}{\sigma_0^2}\right)} - \frac{1}{2}\left(\frac{x^2}{\sigma^2} + \frac{\mu_0^2}{\sigma_0^2}\right)\right) =$$

$$\frac{1}{\sigma\sigma_0\sqrt{2\pi}\sqrt{\left(\frac{1}{\sigma^2} + \frac{1}{\sigma_0^2}\right)}} \exp\left(\frac{\left(\frac{x}{\sigma^2} + \frac{\mu_0}{\sigma_0^2}\right)^2}{2\left(\frac{1}{\sigma^2} + \frac{1}{\sigma_0^2}\right)} - \frac{1}{2}\left(\frac{x^2}{\sigma^2} + \frac{\mu_0^2}{\sigma_0^2}\right)\right).$$

Now using this we arrive at the desired result:

$$p(\mu|x) = \frac{1}{\sqrt{2\pi\left(\frac{1}{\sigma^2} + \frac{1}{\sigma_0^2}\right)}} \exp\left(-\frac{1}{2}\left(\frac{1}{\sigma^2} + \frac{1}{\sigma_0^2}\right)^2\left(\mu - \frac{\mu_0\sigma^2 + x\sigma_0^2}{\sigma^2 + \sigma_0^2}\right)^2\right).$$