CS 1810 Spring 2025 Section 0 Notes: Linear Algebra, Calculus, Probability

The goal of these section notes is to cover some material that is mostly review for CS 1810. 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. 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:

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

This is the default format. Sometimes vectors will be in row form and their symbols may not be bolded. 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:

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

More generally, we define the $\mathbf{L_p}$ norm to be:

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

and in this class we focus on the $\mathbf{L_2}$ and $\mathbf{L_1}$ norms.
Multiplication by constants (e.g. $a\mathbf{x}$) and vector addition (e.g. $\mathbf{x}+\mathbf{y}$) work element-wise, exactly as you'd expect. 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 $\langle \mathbf{u}, \mathbf{v} \rangle$ or simply $\mathbf{u}^\top \mathbf{v}$.

1.2 Linear Independence

A set of vectors $\{\mathbf{v}_1, ..., \mathbf{v}_n\}$ is **linearly independent** if and only if the equation

$$c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots + c_n\mathbf{v}_n = \mathbf{0}$$

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** V is a collection of vectors that follow several axioms regarding the properties of scaling and addition described above, and most importantly:

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

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.

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 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 $\mathcal{S} \subseteq V$, then $\mathcal{S}$ is a **subspace** of 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 $\mathbf{u}, \mathbf{v} \in V$ 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 $\mathbf{v}$.

Finally, it is possible to project a vector $\mathbf{u}$ in a vector space V onto a subspace S of V. If the set of vectors $\{\mathbf{s}_1, ..., \mathbf{s}_m\}$ form an orthonormal basis for S, then the **subspace projection** $\mathbf{u}^{\parallel}$ of $\mathbf{u}$ onto $S = \text{span}(\mathbf{s}_1, ..., \mathbf{s}_m)$ 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 S, that $\mathbf{u} = \mathbf{u}^{\parallel}$ if and only if $\mathbf{u} \in S$, and that $\mathbf{u}^{\parallel}$ is the closest vector in S to $\mathbf{u}$: $||\mathbf{u} - \mathbf{v}|| > ||\mathbf{u} - \mathbf{u}^{\parallel}||, \forall\ \mathbf{v} \neq \mathbf{u}^{\parallel}, \mathbf{v} \in S$.

1.5 Exercise: Concept Checks for Vectors

(a) Does linear independence of a set of vectors imply orthogonality? How about the other way around?

(b) Are the vectors in a basis always orthogonal to each other?

Solution:

(a) Linear independence does not imply orthogonality. Intuitively, it just means that the vectors are not parallel. However, orthogonality does imply linear independence.

(b) No, they only have to be linearly independent.

1.6 Matrices

A **matrix** is a rectangular array of scalars. We call a matrix with n rows and m columns an $n \times m$ matrix $\mathbf{A} \in \mathbb{R}^{n \times m}$ (note the bold uppercase), and we denote the elements A_{ij} to be the scalars found at row i, column j. Conceptually, you can think of matrices as **linear operators** transforming m-dimensional vectors to n-dimensional vectors. 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 λ_1 and λ_2. Finally, note that we can think of n-dimensional vectors as matrices with shape $n \times 1$.

1.7 Exercise: Linear Independence Revisited

Recall that our definition of linear independence made use of the equation

$$c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \ldots + c_n\mathbf{v}_n = \mathbf{0}$$

How can we construct a matrix $\mathbf{A}$ so that we can rewrite this equation as $\mathbf{Ac} = \mathbf{0}$, where $\mathbf{c} = [c_1, \ldots, c_n]^\top$?

Solution: Just let $\mathbf{v}_1, \ldots, \mathbf{v}_n$ be the columns of $\mathbf{A}$.

1.8 Describing Matrices

- $\mathbf{A}^\top$ 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.
- **Diagonal** matrices have non-zero values on the main diagonal (top left to bottom right) and zeros elsewhere. Diagonal matrices are easy to take powers of because you just get another diagonal matrix, except with the entries being the corresponding powers of the original entries. The most classic example of a diagonal matrix is the **identity** matrix $\mathbf{I}$ (written as $\mathbf{I}_n$ for an $n \times n$ identity matrix), which has all ones on the diagonal.
- 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 Matrix Multiplication Properties

$\mathbf{AB}$ 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 ith row of $\mathbf{A}$ with the jth column of $\mathbf{B}$. This yields a $p \times r$ dimensional matrix $\mathbf{AB}$. The following are several important properties of matrix multiplication:

- Generally not commutative: $\mathbf{AB} \neq \mathbf{BA}$
- Associative: $(\mathbf{AB})\mathbf{C} = \mathbf{A}(\mathbf{BC})$
- Left and 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 any scalar λ: $\lambda(\mathbf{AB}) = (\lambda\mathbf{A})\mathbf{B} = \mathbf{A}(\lambda\mathbf{B})$.
- Transpose of product: $(\mathbf{AB})^\top = \mathbf{B}^\top\mathbf{A}^\top$

1.10 Exercise: Basic Computations

Given the matrix $\mathbf{X}$, the matrix $\mathbf{Y}$, and the vector $\mathbf{z}$ below:

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{pmatrix}, \quad \mathbf{Y} = \begin{pmatrix} y_{11} & y_{12} \\ y_{21} & y_{22} \end{pmatrix}, \quad \mathbf{z} = \begin{pmatrix} z_1 \\ z_2 \end{pmatrix}.$$

(a) Expand $\mathbf{XYz}$.

(b) Expand $(\mathbf{Xz})^\top \mathbf{Xz}$.

Solution:

(a)

$$\mathbf{XYz} = \begin{pmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{pmatrix} \begin{pmatrix} y_{11} & y_{12} \\ y_{21} & y_{22} \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \end{pmatrix}.$$

First, compute $\mathbf{Yz}$:

$$\mathbf{Yz} = \begin{pmatrix} y_{11} & y_{12} \\ y_{21} & y_{22} \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = \begin{pmatrix} y_{11}z_1 + y_{12}z_2 \\ y_{21}z_1 + y_{22}z_2 \end{pmatrix}.$$

Then compute $\mathbf{X}(\mathbf{Yz})$:

$$\mathbf{XYz} = \begin{pmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{pmatrix} \begin{pmatrix} y_{11}z_1 + y_{12}z_2 \\ y_{21}z_1 + y_{22}z_2 \end{pmatrix} = \begin{pmatrix} x_{11}(y_{11}z_1 + y_{12}z_2) + x_{12}(y_{21}z_1 + y_{22}z_2) \\ x_{21}(y_{11}z_1 + y_{12}z_2) + x_{22}(y_{21}z_1 + y_{22}z_2) \end{pmatrix}.$$

(b)

$$(\mathbf{Xz})^\top \mathbf{Xz} = \begin{pmatrix} z_1 & z_2 \end{pmatrix} \mathbf{X}^\top \mathbf{X} \begin{pmatrix} z_1 \\ z_2 \end{pmatrix}.$$

Compute $\mathbf{Xz}$:

$$\mathbf{Xz} = \begin{pmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = \begin{pmatrix} x_{11}z_1 + x_{12}z_2 \\ x_{21}z_1 + x_{22}z_2 \end{pmatrix}.$$

Then compute $(\mathbf{Xz})^\top$:

$$(\mathbf{Xz})^\top = \begin{pmatrix} x_{11}z_1 + x_{12}z_2 & x_{21}z_1 + x_{22}z_2 \end{pmatrix}.$$

Finally, expand:

$$(\mathbf{Xz})^\top \mathbf{Xz} = \begin{pmatrix} x_{11}z_1 + x_{12}z_2 & x_{21}z_1 + x_{22}z_2 \end{pmatrix} \begin{pmatrix} x_{11}z_1 + x_{12}z_2 \\ x_{21}z_1 + x_{22}z_2 \end{pmatrix}.$$

Simplify:

$$(\mathbf{Xz})^\top \mathbf{Xz} = (x_{11}z_1 + x_{12}z_2)^2 + (x_{21}z_1 + x_{22}z_2)^2.$$

1.11 Rank, Inverse, and Determinant

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. $\min(n, m)$.
The **inverse** of a square $n \times n$ matrix $\mathbf{A}$ is another $n \times n$ matrix $\mathbf{A}^{-1}$ such that $\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$. By the invertible matrix theorem, $\mathbf{A}$ is invertible if and only if it is full rank. A non-invertible matrix may also be called *singular*. The following are two important properties of matrix inverses:

- $(\mathbf{A}^\top)^{-1} = (\mathbf{A}^{-1})^\top$
- $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$

Finally, the **determinant** $\det(\mathbf{A})$ is defined for a square matrix $\mathbf{A}$ and is a scalar quantity that captures some of the properties of $\mathbf{A}$. Namely, we have that $\det(\mathbf{A}) = 0$ if and only if $\mathbf{A}$ is singular, again by the invertible matrix theorem. The computation of the determinant differs for square matrices of different sizes, but you do not need to know this for CS 1810. The following are a few useful properties of the determinant:

- The determinant of a diagonal matrix is the product of its diagonal values.
- 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: $|\mathbf{AB}| = |\mathbf{A}| \cdot |\mathbf{B}|$.

1.12 Exercise: Reasoning About Shapes

Assume matrix $\mathbf{X}$ has shape $(n \times d)$, and vector $\mathbf{w}$ has shape $(d \times 1)$.

(a) What shape is $\mathbf{y} = \mathbf{X}\mathbf{w}$?

(b) What shape is $(\mathbf{X}^\top\mathbf{X})^{-1}$?

(c) Using $\mathbf{y}$ from part (a), what shape is $(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$?

Solution:

(a) $(n \times 1)$

(b) $(d \times d)$

(c) $(d \times 1)$. This is the OLS linear regression estimator!

1.13 Exercise: Some Basic Matrix Algebra

Suppose $\mathbf{A}$ is an $n \times n$ matrix that is invertible, and that $\mathbf{x}, \mathbf{y}, \mathbf{z}$ are all $n \times 1$ column vectors. Solve the following matrix equations for $\mathbf{x}$:

(a) $\mathbf{A}\mathbf{x} - \mathbf{y} = \mathbf{z}$

(b) $-2\mathbf{A}^\top(\mathbf{y} - \mathbf{x}) = \mathbf{0}$

Solution:

(a) $\mathbf{x} = \mathbf{A}^{-1}(\mathbf{y} + \mathbf{z})$

(b) $\mathbf{x} = \mathbf{y}$

1.14 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. Sometimes the matrix preserves the direction of these vectors: $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$ for some scalar value λ. In this case, λ 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. To solve for eigenvalues, we solve the equations $|\mathbf{A} - \lambda\mathbf{I}| = 0$ for λ. Given an eigenvalue λ, we then solve for the corresponding eigenvector by solving $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$ for $\mathbf{x}$.

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 $\mathbf{\Lambda}_{ii}^{-1} = \frac{1}{\lambda_i}$.

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**, meaning that $\mathbf{U}$ has orthonormal (orthogonal unit vectors) rows and columns. Orthogonal matrices $\mathbf{U}$ follow the property that $\mathbf{U}^\top = \mathbf{U}^{-1}$. 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 σ_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 $\mathbf{A}$.

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 ∇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 1810 can be found on p. 8-11 of *The Matrix Cookbook* by Petersen and Pedersen. We've reproduced a few important

derivatives here:

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

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

2.2 Exercise: A Potpourri of Matrix Calculus

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

(b) Let $f(\mathbf{w}) = (1 - \mathbf{w}^\top \mathbf{x})^2$. 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 $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top \mathbf{A}\mathbf{x} + \mathbf{w}^\top \mathbf{x}$, find $\nabla f(\mathbf{x})$.

Solution:

(a) $\nabla_x f(\mathbf{x}) = (\mathbf{I} + \mathbf{I}^\top)\mathbf{x} = 2\mathbf{x}$.

(b) $\nabla_w f(\mathbf{w}) = 2\left(1 - \mathbf{w}^\top \mathbf{x}\right) \cdot \nabla_w \left(1 - \mathbf{w}^\top \mathbf{x}\right) = 2\left(1 - \mathbf{w}^\top \mathbf{x}\right) \cdot -\nabla_w \mathbf{w}^\top \mathbf{x} = -2\left(1 - \mathbf{w}^\top \mathbf{x}\right) \cdot \mathbf{x}$, using the very first derivative property.

(c) $\nabla_x f(\mathbf{x}) = g'(h(\mathbf{x})) \cdot \nabla_x h(\mathbf{x})$ using the Chain Rule.

(d) $\nabla_x f(\mathbf{x}) = \frac{1}{2}\nabla_x \mathbf{x}^\top \mathbf{A}\mathbf{x} + \nabla_x \mathbf{w}^\top \mathbf{x} = \frac{1}{2}\left(\mathbf{A} + \mathbf{A}^\top\right)\mathbf{x} + \mathbf{w}$. Since $\mathbf{A}$ is symmetric, we have that this is equal to $\mathbf{A}\mathbf{x} + \mathbf{w}$.

2.3 Optimization

Local Extrema: Recall that the local extrema of a single-variable function can be found by setting its derivative to 0. We can furthermore characterize each of these as a maximum, minimum, or neither through the second derivative test—if $f''(x) > 0$ then minimum, if $f''(x) < 0$ then minimum, and neither otherwise. This generalizes to the multivariable setting, using the condition

$\frac{d\mathbf{f}(\mathbf{x})}{d\mathbf{x}} = \mathbf{0}$ and the second derivative test being generalized to whether the Hessian is positive definite (minimum), negative definite (maximum), or neither. Note that the first order condition is often intractable, in which case we can use numerical methods to search for local minima (we will cover some of these later in the course).

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 λ 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 often intractable and requires numerical methods.

2.4 Exercise: Constrained Optimization with Cobb-Douglas

The Cobb-Douglas function is a common functional form used in economics. A simple form is $f(x, y) = x^\alpha y^\beta$. Solve for (x, y) that maximize $f(x, y)$ subject to the constraint that $x + y = 1$.

Solution: Our constraint can be written as

$$g(x, y) = 1 - (x + y) = 0$$

Hence, the Lagrangian is

$$L(x, y, \lambda) = x^\alpha y^\beta + \lambda(1 - x - y)$$

Differentiating with respect to x, y, λ and setting these partials equal to zero, we have

$$\alpha x^{\alpha-1} y^\beta = \lambda, \quad x^\alpha \beta y^{\beta-1} = \lambda, \quad x + y = 1$$

Setting the first two equal, we have that

$$\alpha x^{\alpha-1} y^\beta = x^\alpha \beta y^{\beta-1} \implies y = \frac{\beta}{\alpha} x$$

Finally, we substitute this into the third first order condition:

$$x + \frac{\beta}{\alpha} x = 1 \implies x = \frac{1}{1 + \beta/\alpha} = \boxed{\frac{\alpha}{\alpha + \beta}}, \quad y = \boxed{\frac{\beta}{\alpha + \beta}}$$

3 Probability

As part of the prerequisites, you should be comfortable with probability theory at the level of Harvard's Stat 110. In this section, we review the relevant concepts that will be used in this course. For your reference, a public version of Stat 110 can be found here.

3.1 Probability Basics

Probability provides a measure of how likely it is that some **event** will occur. An event is a subset of the sample space (the set of all possible outcomes of some process). For example, if you are rolling a six-sided die, the sample space consists of six outcomes: $S = \{1, 2, 3, 4, 5, 6\}$. We can then represent the "probability of getting an even number" as $P(E) = \frac{1}{2}$, where $E = \{2, 4, 6\}$. 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 that A occurs given that B occurs. You can think of conditional probability as an updated version of the marginal probability $P(A)$, now incorporating the information gained from knowing that B occurred. We have the following definition:

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

  It follows that

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

- **Independence**: Events A and B are independent if knowing whether A occurred gives no information about whether B occurred. More precisely, we have:

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

  Looking at our conditional probability formulas, this also yields

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

- **Probability of union**: We don't use this identity much in this course, but here it is still important and left here for reference:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

- **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 (LOTP)**: Sometimes it is quite hard to calculate marginal probabilities $P(A)$. However, conditional probabilities of A could be easier to calculate. We can partition the entire sample space into **disjoint** (non-overlapping) events $B_1, B_2, B_3, ..., B_n$ and compute:

$$\begin{aligned} P(A) &= \sum_{i}^{n} P(A \cap B_i) \\ &= \sum_{i}^{n} P(A|B_i)P(B_i) \end{aligned}$$

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

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

3.2 Exercise: 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.

Solution: We want to find $P(D|T)$. Using Bayes' Rule, we have

$$P(D|T) = \frac{P(T|D)P(D)}{P(D)} = \frac{0.95 \times 0.01}{P(T)}$$

It remains to find $P(T)$. We can do so using LOTP as such:

$$\begin{aligned} P(T) &= P(T|D)P(D) + P(T|D^c)P(D^c) \\ &= 0.95 \times 0.01 + 0.05 \times 0.99 \\ &= 0.059 \end{aligned}$$

Substituting this into the fraction above, we get that

$$P(D|T) \approx 0.161$$

Surprisingly, this means that Fred likely still doesn't have the disease! Intuitively, this is because only a very small fraction of the entire population actually has the disease.

3.3 Exercise: Derivations and More Identities

(a) How do you derive Bayes' rule (the version with just A, B)?

(b) How do you derive Bayes' rule with extra conditioning?

(c) Derive an LOTP-style identity for $P(A|C)$.

Solution:

(a) Recall that

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

We simply substitute the left equation into the right one.

(b) The key is that $B \cap C$ is one event, so the definition of conditional probability gives us:

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

Now we note that $A \cap C$ is also one event, so we can write the numerator as

$$P(A,B,C) = P(B|A,C)P(A,C) = P(B|A,C)P(A|C)P(C)$$

Finally, the denominator is equal to $P(B|C)P(C)$ so putting this all together:

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

as desired.

(c) Suppose we partition the entire sample space into disjoint events $B_1, \ldots, B_n$. Then we have

$$\begin{aligned} P(A|C) &= \frac{P(A,C)}{P(C)} \\ &= \frac{1}{P(C)} \sum_{i=1}^{n} P(A, B_i, C) \\ &= \frac{1}{P(C)} \sum_{i=1}^{n} P(A|B_i, C)P(B_i, C) \\ &= \sum_{i=1}^{n} P(A|B_i, C)P(B_i|C) \end{aligned}$$

3.4 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 encodes any outcome that results from this experiment. We generally define random variables to be real-valued, although the sample space that they reflect can be almost anything. We usually denote random variables with capital letters, like X. The following are some examples of random variables:

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

A random variable is characterized by its **distribution**, which intuitively captures how likely all the possible values of X are. A random variable can be *discrete* or *continuous*, depending on the set of possible values it can take on, which we call its **support**. We define the distribution of a discrete random variable through its **probability mass function** (PMF). In particular, we write $p(x)$ to capture the probability of X taking on the value x, i.e. $P(X = x)$.
The analog of the PMF for continuous random variables is the **probability density function** (PDF), which we also use $p(x)$ to represent. However, it is important to note that for continuous r.v.s, $p(x)$ is different from $P(X = x)$. In fact, we have that for any x in the support of X, $P(X = x) = 0$ even if $p(x)$ is positive. Also, note that $p(x)$ can take on any nonnegative real value, meaning that it can be greater than one. Intuitively, we should think of the function $p(x)$ as assigning *densities* that behave like *relative probabilities* rather than absolute probabilities. We can easily go between the probability density and probability using integration:

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

We usually work with variables that have named distributions, like the Binomial or Normal. The standard notation that you'll see for these distributions looks like $X \sim \text{Binom}(n, p)$, meaning that X is distributed according to the Binomial distribution with parameters n and p.

An important note on notation concerns lowercase symbols such as x. Something potentially confusing in ML is that we often treat x itself as a random variable, while in statistics we generally treat x as a crystallized (known) value coming from the distribution of a random variable X. Usually, the context will make it clear enough which of these to follow. A common scenario for treating x as random is when we tell you that x is a data point sampled from some distribution. We often do this by writing $x \sim p(x)$, where $p(x)$ refers to the PMF/PDF of x. Note that this is not a perfect choice of notation since in this context $p(x)$ should really mean the PMF/PDF of x evaluated at x. Thus, you could imagine replacing the first notation with something like $x \sim p(.)$. That being said, we will stick with using $p(x)$ to refer to both the PMF/PDF of x and the evaluation of that function at the value x, since we can generally infer the meaning from the context.

3.5 Exercise: PDF Concept Check

Why is $P(X = x) = 0$ for a continuous random variable X?

Solution: X can take on infinitely many values due to being continuous, so the probability of taking on a specific value is 0.

3.6 Expectation and Variance

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:

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

where $g : \Omega \to \mathbb{R}$. Note that we often drop the subscript underneath the $\mathbb{E}$. For a 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 $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$.

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. An important property of variance is that it is always nonnegative.

3.6.1 Joint and Conditional Distributions

The **joint distribution** of two random variables X and Y captures the relationship between these variables. We write $p(x, y)$ to capture the likelihood (remember that for continuous variables, this is not the same as probability) of observing values x and y from the joint distribution of X and Y.

Receiving information about the value of a random variable Y can change the distribution of another variable X. We call this the **conditional distribution** of X, given that $Y = y$. 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 $p(x|y)$ to capture the likelihood of observing the

value x from the distribution of X given that we observed $Y = y$. The definition of this quantity is analogous to the definition of conditional probability:

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

It follows that we can factor joint PMFs/PDFs into chains of conditional PMFs/PDFs as such:

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

3.7 Exercise: More Identities

Fill in the "etc..."' above with any additional expressions equal to $p(x,y,z)$.

Solution:

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

3.8 Bayes' Rule and Marginalizing

As you'd expect, **Bayes' Rule** extends to conditional distributions:

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

We often only care about $p(x|y)$ as a function of x, in which case we just write

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

and compute the RHS.

When you have a joint distribution of two or more random variables, you may just want the **marginal distribution** of a single variable. For instance, we may want to compute $p(y)$ for the denominator of the above expressions for $p(x|y)$. We can use the **sum rule** (this is just LOTP) to compute this:

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

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

This process of summing over the support of one of the variables is called *marginalizing*. Think about the marginal distribution as what 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 marginalizing over the other two.

3.9 Independence of Random Variables

The notion of **independence** extends to random variables as well. Two random variables X, Y are 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 $\perp\!\!\!\perp$ symbol, where $X \perp\!\!\!\perp Y$ implies X is independent of Y.

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

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 we are given Z, then knowing X tells us nothing about Y and vice-versa (given Z, knowing Y tells us nothing about X).

3.10 Covariance

The **covariance** between two jointly distributed random variables X and Y with finite variances is defined as the expected product of their deviations from their individual expected values. Intuitively, this captures the direction and strength of a linear relationship between these two variables. To oversimplify, do X and Y tend to increase and decrease together? We define this quantity as such:

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

The following are several essential properties of covariance:

- $\text{Cov}(X, X) = \text{Var}(X)$, by definition.
- Symmetric: $\text{Cov}(X, Y) = \text{Cov}(Y, X)$
- Constants don't vary: $\text{Cov}(X, c) = 0$ for a constant c.
- Bilinear: $\text{Cov}(aX + bY, Z) = a\text{Cov}(X, Z) + b\text{Cov}(Y, Z)$
- $X \perp\!\!\!\perp Y \Rightarrow \text{Cov}(X, Y) = 0$, but the converse is not necessarily true.

3.11 Exercise: Proving Some Useful Identities

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

(b) Show that for random variables X, Y that $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y)$. Note that this implies $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$ when $X \perp\!\!\!\perp Y$.

Solution:

(a) We can expand the terms to get

$$\begin{aligned}\text{Var}(aX + b) &= \text{Cov}(aX + b, aX + b) \\ &= a^2\text{Cov}(X, X) + 0 + 0 + 0 \\ &= a^2\text{Var}(X)\end{aligned}$$

(b)

$$\text{Var}(X + Y) = \text{Cov}(X + Y, X + Y) = \text{Cov}(X, X) + \text{Cov}(Y, Y) + 2\text{Cov}(X, Y)$$

Since $\text{Cov}(Z, Z) = \text{Var}(Z)$ for any r.v. Z, we are done.

3.12 Exercise: Sample Mean Properties

Suppose that $X_1, \ldots, X_n$ are i.i.d. scalar random variables with mean μ and variance σ^2. Let $\bar{X}$ be the sample mean $\frac{1}{n}\sum_{i=1}^n X_i$. Find $\mathbb{E}(\bar{X})$ and $\text{Var}(\bar{X})$.

Solution:

$$\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n X_i\right) = \frac{1}{n}\sum_{i=1}^n \mathbb{E}(X_i) = \frac{n\mu}{n} = \mu$$

$$\text{Var}\left(\frac{1}{n}\sum_{i=1}^n X_i\right) = \frac{1}{n^2}\sum_{i=1}^n \text{Var}(X_i) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

3.13 Conditional Expectation and Conditional Variance

We can also define the **conditional expectation** of X given $Y = y$ as the quantity $\mathbb{E}[X|Y = y]$, which is the expected value of the random variable X given a particular observed value y from the distribution of Y. An example is the expected temperature, given no rain.

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

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

Similar to how it may sometimes be easier to compute conditional probabilities as opposed to marginal probabilities, the same can be true for expectations and variance. Thus, we have two useful tools:

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

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

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

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

3.14 Exercise: Proving Eve's Law

Prove Eve's law using Adam's law.

Solution: The first line is by the definition of variance, the second is by Adam's law, and the fourth line is by definition of variance.

$$\begin{aligned}
\text{Var}[X] &= \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 \\
&= \mathbb{E}\left[\mathbb{E}\left[X^2 \mid Y\right]\right] - \mathbb{E}[\mathbb{E}[X \mid Y]]^2 \\
&= \mathbb{E}\left[\mathbb{E}\left[X^2 \mid Y\right]\right] - \mathbb{E}\left[\mathbb{E}[X \mid Y]^2\right] + \mathbb{E}\left[\mathbb{E}[X \mid Y]^2\right] - \mathbb{E}[\mathbb{E}[X \mid Y]]^2 \\
&= \mathbb{E}[\text{Var}[X \mid Y]] + \text{Var}[\mathbb{E}[X \mid Y]]
\end{aligned}$$

3.15 Some Named Distributions

You won't have to know many named distributions off the top of your head for CS 1810, but it will be helpful to review the most common ones.

3.15.1 Univariate Normal

The univariate Normal (AKA Gaussian) is a continuous distribution over all real numbers with a mean parameter μ and a variance parameter σ^2. It has the following PDF:

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

Please keep in mind the notation $\mathcal{N}(x; \mu, \sigma^2)$. Here it is referring to the PDF of a Normal random variable (which we notate differently as $X \sim \mathcal{N}(\mu, \sigma^2)$) evaluated at the scalar value x. The univariate Normal is often referred to as a bell curve.

The following are several important properties of Normal:

- If $X \sim \mathcal{N}(\mu, \sigma^2)$, then $aX + b \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$.
- 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 Normal PDF.

[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".

3.15.2 Multivariate Normal

A **random vector** is a vector of random variables, as you would expect. The Multivariate Normal (MVN) is essentially the random vector version of the Normal with a key property: every linear combination of its components is distributed as a univariate Normal. This means that for a vector to be MVN, it is necessary for all its components to be Normal, but not sufficient.

Again, there are two parameters—assuming the random vector has dimension m, it has a mean vector $\mu \in \mathbb{R}^m$ and a covariance matrix $\mathbf{\Sigma} \in \mathbb{R}^{m \times m}$. For each index $i = 1, \ldots, m$, we have that $E(X_i) = \mu_i$. For each pair of indices (i, j), we have that $\text{Cov}(X_i, X_j) = \Sigma_{ij}$. We have the following PDF:

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

One useful property of the MVN is that if the covariance between two of its components is 0, then this implies those components are independent.

3.15.3 Binomial

The Binomial is a discrete distribution with parameters n and p over the integers 0 to n. The story of the Binomial is that it counts the number of successes in n independent trials, each having a probability p of success. It has the following PMF:

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

The mean of the Binomial is $E(X) = np$, while the variance is $\text{Var}(X) = np(1 - p)$.

When $n = 1$, we call this the Bernoulli distribution. Note that this is the distribution of the indicator variables I_A that we mentioned earlier, where p is the probability of the associated event A occuring.

3.15.4 Multinomial

The Multinomial is the multivariate generalization of the Binomial. It has parameters $n, k, p_1, \ldots, p_k$. For the story, it still holds that n is the number of independent trials, but now there are k possible categories, each with it s own associated success probability p_i. The PMF still looks similar:

$$P(X_1 = x_1, \ldots, X_k = x_k) = \frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k}$$

The mean for component X_i is simply np_i. You don't need to know this, but the covariance matrix is defined as such for arbitrary component index i and $j \neq i$:

$$\text{Var}(X_i) = np_i(1 - p_i), \quad \text{Cov}(X_i, X_j) = -np_i p_j$$

3.16 Exercise: Practice with Normal and Binomial

(a) Suppose $X_1, \ldots, X_n$ are i.i.d and distributed Normal with mean μ and variance σ^2. What is the distribution of the sample mean?

(b) Take the log of the MVN PDF.

(c) What is the PMF of X, where $X \sim \text{Bernoulli}(p)$? How about the joint PMF of n iid variables $X_1, \ldots, X_n$ where $X_i \sim \text{Bernoulli}(p)$?

Solution:

(a) Use the mean and variance formulas from earlier. Then note that the sum of independent Normals is still Normal. Thus,

$$\mathcal{N}(\mu, \sigma^2/n)$$

(b) The purpose of this problem is just to get familiar with the MVN density through writing it down. The log yields

$$-\frac{1}{2}(\mathbf{x}-\mu)^\top \mathbf{\Sigma}^{-1}(\mathbf{x}-\mu) - \frac{1}{2}\log(\det(2\pi\mathbf{\Sigma}))$$

(c) We have

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

and we have

$$P(X_1 = x_1, \ldots, X_n = x_n) = p^{\sum_{i=1}^n x_n}(1-p)^{(n-\sum_{i=1}^n x_n)}$$

Note that if we let $S = \sum_{i=1}^n X_i$, then it is clear that S is distributed Binomial(n, p) as the story of the Binomial suggests. This is because for a given sum s, we can add up joint PMFs of the above form for every possible tuple $(x_1, \ldots, x_n)$ such that $\sum_{i=1}^n x_i = s$. There are precisely $\binom{n}{s}$ of these, since each coordinate can be 0 or 1.