Mathematics for Machine Learning

Garrett Thomas
Department of Electrical Engineering and Computer Sciences
University of California, Berkeley

August 6, 2017

1 About

Machine learning uses tools from a variety of mathematical fields. This document is an attempt to provide a summary of the mathematical background needed for an introductory class in machine learning, which at UC Berkeley is known as CS 189/289A.

Our assumption is that the reader is already familiar with the basic concepts of multivariable calculus and linear algebra (at the level of UCB Math 53/54). We emphasize that this document is **not** a replacement for the prerequisite classes. Most subjects presented here are covered rather minimally; we intend to give an overview and point the interested reader to more comprehensive treatments for further details.

Note that this document concerns math background for machine learning, not machine learning itself. We will not discuss specific machine learning models or algorithms except possibly in passing to highlight the relevance of a mathematical concept.

Earlier versions of this document did not include proofs. We have begun adding in proofs where they are reasonably short and aid in understanding. These proofs are not necessary background for CS 189 but can be used to deepen the reader's understanding.

You are free to distribute this document as you wish. The latest version can be found at http://gwthomas.github.io/docs/math4ml.pdf. Please report any mistakes to gwthomas@berkeley.edu.

Contents

2 Notation

Notation	Meaning
$\mathbb{R}$	set of real numbers
$\mathbb{R}^n$	set (vector space) of n-tuples of real numbers, endowed with the usual inner product
$\mathbb{R}^{m \times n}$	set (vector space) of m-by-n matrices
δ_{ij}	Kronecker delta, i.e. $\delta_{ij} = 1$ if $i = j$, 0 otherwise
$\nabla f(\mathbf{x})$	gradient of the function f evaluated at $\mathbf{x}$
$\nabla^2 f(\mathbf{x})$	Hessian of the function f evaluated at $\mathbf{x}$
$\mathbf{A}^\top$	transpose of the matrix $\mathbf{A}$
Ω	sample space
$\mathbb{P}(A)$	probability of event A
$p(X)$	distribution of random variable X
$p(x)$	probability density/mass function evaluated at x
A^c	complement of event A
$A \dot\cup B$	union of A and B, with the extra requirement that $A \cap B = \varnothing$
$\mathbb{E}[X]$	expected value of random variable X
$\mathrm{Var}(X)$	variance of random variable X
$\mathrm{Cov}(X, Y)$	covariance of random variables X and Y

Other notes:

- Vectors and matrices are in bold (e.g. $\mathbf{x}, \mathbf{A}$). This is true for vectors in $\mathbb{R}^n$ as well as for vectors in general vector spaces. We generally use Greek letters for scalars and capital Roman letters for matrices and random variables.
- To stay focused at an appropriate level of abstraction, we restrict ourselves to real values. In many places in this document, it is entirely possible to generalize to the complex case, but we will simply state the version that applies to the reals.
- We assume that vectors are column vectors, i.e. that a vector in $\mathbb{R}^n$ can be interpreted as an n-by-1 matrix. As such, taking the transpose of a vector is well-defined (and produces a row vector, which is a 1-by-n matrix).

3 Linear Algebra

In this section we present important classes of spaces in which our data will live and our operations will take place: vector spaces, metric spaces, normed spaces, and inner product spaces. Generally speaking, these are defined in such a way as to capture one or more important properties of Euclidean space but in a more general way.

3.1 Vector spaces

Vector spaces are the basic setting in which linear algebra happens. A vector space V is a set (the elements of which are called **vectors**) on which two operations are defined: vectors can be added together, and vectors can be multiplied by real numbers[1] called **scalars**. V must satisfy

(i) There exists an additive identity (written $\mathbf{0}$) in V such that $\mathbf{x} + \mathbf{0} = \mathbf{x}$ for all $\mathbf{x} \in V$

(ii) For each $\mathbf{x} \in V$, there exists an additive inverse (written $-\mathbf{x}$) such that $\mathbf{x} + (-\mathbf{x}) = \mathbf{0}$

(iii) There exists a multiplicative identity (written 1) in $\mathbb{R}$ such that $1\mathbf{x} = \mathbf{x}$ for all $\mathbf{x} \in V$

(iv) Commutativity: $\mathbf{x} + \mathbf{y} = \mathbf{y} + \mathbf{x}$ for all $\mathbf{x}, \mathbf{y} \in V$

(v) Associativity: $(\mathbf{x} + \mathbf{y}) + \mathbf{z} = \mathbf{x} + (\mathbf{y} + \mathbf{z})$ and $\alpha(\beta\mathbf{x}) = (\alpha\beta)\mathbf{x}$ for all $\mathbf{x}, \mathbf{y}, \mathbf{z} \in V$ and $\alpha, \beta \in \mathbb{R}$

(vi) Distributivity: $\alpha(\mathbf{x} + \mathbf{y}) = \alpha\mathbf{x} + \alpha\mathbf{y}$ and $(\alpha + \beta)\mathbf{x} = \alpha\mathbf{x} + \beta\mathbf{x}$ for all $\mathbf{x}, \mathbf{y} \in V$ and $\alpha, \beta \in \mathbb{R}$

3.1.1 Euclidean space

The quintessential vector space is **Euclidean space**, which we denote $\mathbb{R}^n$. The vectors in this space consist of n-tuples of real numbers:

$$\mathbf{x} = (x_1, x_2, \ldots, x_n)$$

For our purposes, it will be useful to think of them as $n \times 1$ matrices, or **column vectors**:

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

Addition and scalar multiplication are defined component-wise on vectors in $\mathbb{R}^n$:

$$\mathbf{x} + \mathbf{y} = \begin{bmatrix} x_1 + y_1 \\ \vdots \\ x_n + y_n \end{bmatrix}, \qquad \alpha\mathbf{x} = \begin{bmatrix} \alpha x_1 \\ \vdots \\ \alpha x_n \end{bmatrix}$$

Euclidean space is used to mathematically represent physical space, with notions such as distance, length, and angles. Although it becomes hard to visualize for $n > 3$, these concepts generalize mathematically in obvious ways. Tip: even when you're working in more general settings than $\mathbb{R}^n$, it is often useful to visualize vector addition and scalar multiplication in terms of 2D vectors in the plane or 3D vectors in space.

[1] More generally, vector spaces can be defined over any **field** $\mathbb{F}$. We take $\mathbb{F} = \mathbb{R}$ in this document to avoid an unnecessary diversion into abstract algebra.

3.2 Metric spaces

Metrics generalize the notion of distance from Euclidean space (although metric spaces need not be vector spaces).

A **metric** on a set S is a function $d : S \times S \to \mathbb{R}$ that satisfies

(i) $d(x, y) \geq 0$, with equality if and only if $x = y$

(ii) $d(x, y) = d(y, x)$

(iii) $d(x, z) \leq d(x, y) + d(y, z)$ (the so-called **triangle inequality**)

for all $x, y, z \in S$.

A key motivation for metrics is that they allow limits to be defined for mathematical objects other than real numbers. We say that a sequence $\{x_n\} \subseteq S$ converges to the limit x if for any $\epsilon > 0$, there exists $N \in \mathbb{N}$ such that $d(x_n, x) < \epsilon$ for all $n \geq N$. Note that the definition for limits of sequences of real numbers, which you have likely seen in a calculus class, is a special case of this definition when using the metric $d(x, y) = |x - y|$.

3.3 Normed spaces

Norms generalize the notion of length from Euclidean space.

A **norm** on a real vector space V is a function $\| \cdot \| : V \to \mathbb{R}$ that satisfies

(i) $\|\mathbf{x}\| \geq 0$, with equality if and only if $\mathbf{x} = \mathbf{0}$

(ii) $\|\alpha\mathbf{x}\| = |\alpha|\|\mathbf{x}\|$

(iii) $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$ (the **triangle inequality** again)

for all $\mathbf{x}, \mathbf{y} \in V$ and all $\alpha \in \mathbb{R}$. A vector space endowed with a norm is called a **normed vector space**, or simply a **normed space**.

Note that any norm on V induces a distance metric on V:

$$d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|$$

One can verify that the axioms for metrics are satisfied under this definition and follow directly from the axioms for norms. Therefore any normed space is also a metric space.[2]

[2] If a normed space is complete with respect to the distance metric induced by its norm, we say that it is a **Banach space**.

We will typically only be concerned with a few specific norms on $\mathbb{R}^n$:

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

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

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

$$\|\mathbf{x}\|_\infty = \max_{1 \leq i \leq n} |x_i|$$

Note that the 1- and 2-norms are special cases of the p-norm, and the ∞-norm is the limit of the p-norm as p tends to infinity. We require $p \geq 1$ for the general definition of the p-norm because the triangle inequality fails to hold if $p < 1$. (Try to find a counterexample!)

Here's a fun fact: for any given finite-dimensional vector space V, all norms on V are equivalent in the sense that for two norms $\|\cdot\|_A, \|\cdot\|_B$, there exist constants $\alpha, \beta > 0$ such that

$$\alpha \|\mathbf{x}\|_A \leq \|\mathbf{x}\|_B \leq \beta \|\mathbf{x}\|_A$$

for all $\mathbf{x} \in V$. Therefore convergence in one norm implies convergence in any other norm. This rule may not apply in infinite-dimensional vector spaces such as function spaces, though.

3.4 Inner product spaces

An **inner product** on a real vector space V is a function $\langle \cdot, \cdot \rangle : V \times V \to \mathbb{R}$ satisfying

(i) $\langle \mathbf{x}, \mathbf{x} \rangle \geq 0$, with equality if and only if $\mathbf{x} = \mathbf{0}$

(ii) $\langle \alpha\mathbf{x} + \beta\mathbf{y}, \mathbf{z} \rangle = \alpha\langle \mathbf{x}, \mathbf{z} \rangle + \beta\langle \mathbf{y}, \mathbf{z} \rangle$

(iii) $\langle \mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{y}, \mathbf{x} \rangle$

for all $\mathbf{x}, \mathbf{y}, \mathbf{z} \in V$ and all $\alpha, \beta \in \mathbb{R}$. A vector space endowed with an inner product is called an **inner product space**.

Note that any inner product on V induces a norm on V:

$$\|\mathbf{x}\| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$$

One can verify that the axioms for norms are satisfied under this definition and follow directly from the axioms for inner products. Therefore any inner product space is also a normed space (and hence also a metric space).[3]

Two vectors $\mathbf{x}$ and $\mathbf{y}$ are said to be **orthogonal** if $\langle \mathbf{x}, \mathbf{y} \rangle = 0$. Orthogonality generalizes the notion of perpendicularity from Euclidean space. If two orthogonal vectors $\mathbf{x}$ and $\mathbf{y}$ additionally have unit length (i.e. $\|\mathbf{x}\| = \|\mathbf{y}\| = 1$), then they are described as **orthonormal**.

[3] If an inner product space is complete with respect to the distance metric induced by its inner product, we say that it is a **Hilbert space**.

The standard inner product on $\mathbb{R}^n$ is given by

$$\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{i=1}^{n} x_i y_i = \mathbf{x}^\top \mathbf{y}$$

The matrix notation on the righthand side (see the Transposition section if it's unfamiliar) arises because this inner product is a special case of matrix multiplication where we regard the resulting 1×1 matrix as a scalar. The inner product on $\mathbb{R}^n$ is also often written $\mathbf{x} \cdot \mathbf{y}$ (hence the alternate name **dot product**). The reader can verify that the two-norm $\|\cdot\|_2$ on $\mathbb{R}^n$ is induced by this inner product.

3.4.1 Pythagorean Theorem

The well-known Pythagorean theorem generalizes naturally to arbitrary inner product spaces.

Theorem 1. *If* $\langle \mathbf{x}, \mathbf{y} \rangle = 0$*, then*

$$\|\mathbf{x} + \mathbf{y}\|^2 = \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2$$

Proof. Suppose $\langle \mathbf{x}, \mathbf{y} \rangle = 0$. Then

$$\|\mathbf{x} + \mathbf{y}\|^2 = \langle \mathbf{x} + \mathbf{y}, \mathbf{x} + \mathbf{y} \rangle = \langle \mathbf{x}, \mathbf{x} \rangle + \langle \mathbf{y}, \mathbf{x} \rangle + \langle \mathbf{x}, \mathbf{y} \rangle + \langle \mathbf{y}, \mathbf{y} \rangle = \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2$$

as claimed. □

3.4.2 Cauchy-Schwarz inequality

This inequality is sometimes useful in proving bounds:

$$|\langle \mathbf{x}, \mathbf{y} \rangle| \leq \|\mathbf{x}\| \cdot \|\mathbf{y}\|$$

for all $\mathbf{x}, \mathbf{y} \in V$. Equality holds exactly when $\mathbf{x}$ and $\mathbf{y}$ are scalar multiples of each other (or equivalently, when they are linearly dependent).

3.5 Transposition

If $\mathbf{A} \in \mathbb{R}^{m \times n}$, its **transpose** $\mathbf{A}^\top \in \mathbb{R}^{n \times m}$ is given by $(\mathbf{A}^\top)_{ij} = A_{ji}$ for each (i, j). In other words, the columns of $\mathbf{A}$ become the rows of $\mathbf{A}^\top$, and the rows of $\mathbf{A}$ become the columns of $\mathbf{A}^\top$.

The transpose has several nice algebraic properties that can be easily verified from the definition:

(i) $(\mathbf{A}^\top)^\top = \mathbf{A}$

(ii) $(\mathbf{A} + \mathbf{B})^\top = \mathbf{A}^\top + \mathbf{B}^\top$

(iii) $(\alpha \mathbf{A})^\top = \alpha \mathbf{A}^\top$

(iv) $(\mathbf{A}\mathbf{B})^\top = \mathbf{B}^\top \mathbf{A}^\top$

3.6 Eigenthings

For a square matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$, there may be vectors which, when $\mathbf{A}$ is applied to them, are simply scaled by some constant. We say that a nonzero vector $\mathbf{x} \in \mathbb{R}^n$ is an **eigenvector** of $\mathbf{A}$ corresponding to **eigenvalue** λ if

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$$

The zero vector is excluded from this definition because $\mathbf{A}\mathbf{0} = \mathbf{0} = \lambda\mathbf{0}$ for every λ.

We now give some useful results about how eigenvalues change after various manipulations.

Proposition 1. *Let $\mathbf{x}$ be an eigenvector of $\mathbf{A}$ with corresponding eigenvalue λ. Then*

(i) For any $\gamma \in \mathbb{R}$, $\mathbf{x}$ is an eigenvector of $\mathbf{A} + \gamma\mathbf{I}$ with eigenvalue $\lambda + \gamma$.

(ii) If $\mathbf{A}$ is invertible, then $\mathbf{x}$ is an eigenvector of $\mathbf{A}^{-1}$ with eigenvalue λ^{-1}.

(iii) $\mathbf{A}^k\mathbf{x} = \lambda^k\mathbf{x}$ for any $k \in \mathbb{Z}$ (where $\mathbf{A}^0 = \mathbf{I}$ by definition).

Proof. (i) follows readily:

$$(\mathbf{A} + \gamma\mathbf{I})\mathbf{x} = \mathbf{A}\mathbf{x} + \gamma\mathbf{I}\mathbf{x} = \lambda\mathbf{x} + \gamma\mathbf{x} = (\lambda + \gamma)\mathbf{x}$$

(ii) Suppose $\mathbf{A}$ is invertible. Then

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{A}\mathbf{x} = \mathbf{A}^{-1}(\lambda\mathbf{x}) = \lambda\mathbf{A}^{-1}\mathbf{x}$$

Dividing by λ, which is valid because the invertibility of $\mathbf{A}$ implies $\lambda \neq 0$, gives $\lambda^{-1}\mathbf{x} = \mathbf{A}^{-1}\mathbf{x}$.

(iii) The case $k \geq 0$ follows immediately by induction on k. Then the general case $k \in \mathbb{Z}$ follows by combining the $k \geq 0$ case with (ii). □

3.7 Trace

The **trace** of a square matrix is the sum of its diagonal entries:

$$\operatorname{tr}(\mathbf{A}) = \sum_{i=1}^{n} A_{ii}$$

The trace has several nice algebraic properties:

(i) $\operatorname{tr}(\mathbf{A} + \mathbf{B}) = \operatorname{tr}(\mathbf{A}) + \operatorname{tr}(\mathbf{B})$

(ii) $\operatorname{tr}(\alpha\mathbf{A}) = \alpha \operatorname{tr}(\mathbf{A})$

(iii) $\operatorname{tr}\left(\mathbf{A}^\top\right) = \operatorname{tr}(\mathbf{A})$

(iv) $\operatorname{tr}(\mathbf{ABCD}) = \operatorname{tr}(\mathbf{BCDA}) = \operatorname{tr}(\mathbf{CDAB}) = \operatorname{tr}(\mathbf{BADC})$

The first three properties follow readily from the definition. The last is known as **invariance under cyclic permutations**. Note that the matrices cannot be reordered arbitrarily, for example $\operatorname{tr}(\mathbf{ABCD}) \neq \operatorname{tr}(\mathbf{BACD})$ in general.

Interestingly, the trace of a matrix is equal to the sum of its eigenvalues (repeated according to multiplicity):

$$\operatorname{tr}(\mathbf{A}) = \sum_i \lambda_i(\mathbf{A})$$

3.8 Determinant

The **determinant** of a square matrix can be defined in several different confusing ways, none of which are particularly important for our purposes; go look at an introductory linear algebra text (or Wikipedia) if you need a definition. But it's good to know the properties:

(i) $\det(\mathbf{I}) = 1$

(ii) $\det\big(\mathbf{A}^\top\big) = \det(\mathbf{A})$

(iii) $\det(\mathbf{AB}) = \det(\mathbf{A})\det(\mathbf{B})$

(iv) $\det\big(\mathbf{A}^{-1}\big) = \det(\mathbf{A})^{-1}$

(v) $\det(\alpha\mathbf{A}) = \alpha^n \det(\mathbf{A})$

Interestingly, the determinant of a matrix is equal to the product of its eigenvalues (repeated according to multiplicity):

$$\det(\mathbf{A}) = \prod_i \lambda_i(\mathbf{A})$$

3.9 Special kinds of matrices

There are several ways matrices can be classified. Each categorization implies some potentially desirable properties, so it's always good to know what kind of matrix you're dealing with.

3.9.1 Orthogonal matrices

A matrix $\mathbf{Q} \in \mathbb{R}^{n\times n}$ is said to be **orthogonal** if its columns are pairwise orthonormal. This definition implies that

$$\mathbf{Q}^\top\mathbf{Q} = \mathbf{QQ}^\top = \mathbf{I}$$

or equivalently, $\mathbf{Q}^\top = \mathbf{Q}^{-1}$. A nice thing about orthogonal matrices is that they preserve inner products:

$$(\mathbf{Qx})^\top(\mathbf{Qy}) = \mathbf{x}^\top\mathbf{Q}^\top\mathbf{Qy} = \mathbf{x}^\top\mathbf{Iy} = \mathbf{x}^\top\mathbf{y}$$

A direct result of this fact is that they also preserve 2-norms:

$$\|\mathbf{Qx}\|_2 = \sqrt{(\mathbf{Qx})^\top(\mathbf{Qx})} = \sqrt{\mathbf{x}^\top\mathbf{x}} = \|\mathbf{x}\|_2$$

Therefore multiplication by an orthogonal matrix can be considered as a transformation that preserves length, but may rotate or reflect the vector about the origin.

3.9.2 Symmetric matrices

A matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ is said to be **symmetric** if it is equal to its own transpose ($\mathbf{A} = \mathbf{A}^\top$). This definition seems harmless enough but turns out to have some strong implications. We summarize the most important of these as

Theorem 2. *(Spectral Theorem) Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ be symmetric. Then there exists an orthonormal basis for $\mathbb{R}^n$ consisting of eigenvectors of $\mathbf{A}$.*

This theorem allows us to factor symmetric matrices as follows:

$$\mathbf{A} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^\top$$

Here $\mathbf{Q}$ is an orthogonal matrix with the aforementioned orthogonal basis as its columns, and $\mathbf{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_n)$, where $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$ are the corresponding eigenvalues[4] of $\mathbf{A}$. This is referred to as the **eigendecomposition** or **spectral decomposition** of $\mathbf{A}$.

3.9.3 Positive (semi-)definite matrices

A symmetric matrix $\mathbf{A}$ is **positive definite** if for all nonzero $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{x}^\top\mathbf{A}\mathbf{x} > 0$. Sometimes people write $\mathbf{A} \succ 0$ to indicate that $\mathbf{A}$ is positive definite. Positive definite matrices have all positive eigenvalues and diagonal entries.

A symmetric matrix $\mathbf{A}$ is **positive semi-definite** if for all $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{x}^\top\mathbf{A}\mathbf{x} \geq 0$. Sometimes people write $\mathbf{A} \succeq 0$ to indicate that $\mathbf{A}$ is positive semi-definite. Positive semi-definite matrices have all nonnegative eigenvalues and diagonal entries.

Positive definite and positive semi-definite matrices will come up very frequently! Note that since these matrices are also symmetric, the properties of symmetric matrices apply here as well.

As an example of how these matrices arise, the matrix $\mathbf{A}^\top\mathbf{A}$ is positive semi-definite for any $\mathbf{A} \in \mathbb{R}^{m\times n}$, since

$$\mathbf{x}^\top(\mathbf{A}^\top\mathbf{A})\mathbf{x} = (\mathbf{A}\mathbf{x})^\top(\mathbf{A}\mathbf{x}) = \|\mathbf{A}\mathbf{x}\|_2^2 \geq 0$$

for any $\mathbf{x} \in \mathbb{R}^n$.

3.10 Singular value decomposition

Singular value decomposition (SVD) is a widely applicable tool in linear algebra. Its strength stems partially from the fact that *every matrix* $\mathbf{A} \in \mathbb{R}^{m\times n}$ has an SVD (even non-square matrices)! The decomposition goes as follows:

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$$

where $\mathbf{U} \in \mathbb{R}^{m\times m}$ and $\mathbf{V} \in \mathbb{R}^{n\times n}$ are orthogonal matrices and $\mathbf{\Sigma} \in \mathbb{R}^{m\times n}$ is a diagonal matrix with the **singular values** of $\mathbf{A}$ (denoted σ_i) on its diagonal. The singular values of $\mathbf{A}$ are defined as the square roots of the eigenvalues of $\mathbf{A}^\top\mathbf{A}$ (or equivalently, of $\mathbf{A}\mathbf{A}^\top$).

By convention, the singular values are given in non-increasing order, i.e.

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{\min(m,n)} \geq 0$$

Only the first r singular values are nonzero, where r is the rank of $\mathbf{A}$.

The columns of $\mathbf{U}$ are called the **left-singular vectors** of $\mathbf{A}$, and they are eigenvectors of $\mathbf{A}\mathbf{A}^\top$. (Try showing this!) The columns of $\mathbf{V}$ are called the **right-singular vectors** of $\mathbf{A}$, and they are eigenvectors of $\mathbf{A}^\top\mathbf{A}$.

[4] The fact that the eigenvalues are real also follows from the symmetry of $\mathbf{A}$.

3.11 Some useful matrix identities

3.11.1 Matrix-vector product as linear combination of matrix columns

Proposition 2. *Let $\mathbf{x} \in \mathbb{R}^n$ be a vector and $\mathbf{A} \in \mathbb{R}^{m \times n}$ a matrix with columns $\mathbf{a}_1, \ldots, \mathbf{a}_n$. Then*

$$\mathbf{A}\mathbf{x} = \sum_{i=1}^{n} x_i \mathbf{a}_i$$

This identity is extremely useful in understanding linear operators in terms of their matrices' columns. The proof is very simple (consider each element of $\mathbf{A}\mathbf{x}$ individually and expand by definitions) but it is a good exercise to convince yourself.

3.11.2 Sum of outer products as matrix-matrix product

An **outer product** is an expression of the form $\mathbf{a}\mathbf{b}^\top$, where $\mathbf{a} \in \mathbb{R}^m$ and $\mathbf{b} \in \mathbb{R}^n$. By inspection it is not hard to see that such an expression yields an $m \times n$ matrix such that

$$[\mathbf{a}\mathbf{b}^\top]_{ij} = a_i b_j$$

It is not immediately obvious, but the sum of outer products is actually equivalent to an appropriate matrix-matrix product! We formalize this statement as

Proposition 3. *Let $\mathbf{a}_1, \ldots, \mathbf{a}_k \in \mathbb{R}^m$ and $\mathbf{b}_1, \ldots, \mathbf{b}_k \in \mathbb{R}^n$. Then*

$$\sum_{\ell=1}^{k} \mathbf{a}_\ell \mathbf{b}_\ell^\top = \mathbf{A}\mathbf{B}^\top$$

where

$$\mathbf{A} = \begin{bmatrix} \mathbf{a}_1 & \cdots & \mathbf{a}_k \end{bmatrix}, \qquad \mathbf{B} = \begin{bmatrix} \mathbf{b}_1 & \cdots & \mathbf{b}_k \end{bmatrix}$$

Proof. For each (i, j), we have

$$\left[\sum_{\ell=1}^{k} \mathbf{a}_\ell \mathbf{b}_\ell^\top \right]_{ij} = \sum_{\ell=1}^{k} [\mathbf{a}_\ell \mathbf{b}_\ell^\top]_{ij} = \sum_{\ell=1}^{k} [\mathbf{a}_\ell]_i [\mathbf{b}_\ell]_j = \sum_{\ell=1}^{k} A_{i\ell} B_{j\ell}$$

This last expression should be recognized as an inner product between the ith row of $\mathbf{A}$ and the jth row of $\mathbf{B}$, or equivalently the jth column of $\mathbf{B}^\top$. Hence by the definition of matrix multiplication, it is equal to $[\mathbf{A}\mathbf{B}^\top]_{ij}$. □

3.12 Quadratic forms

Let $\mathbf{A} \in \mathbb{R}^{n \times n}$ be a symmetric matrix. The expression $\mathbf{x}^\top \mathbf{A}\mathbf{x}$ is called a **quadratic form** and comes up all the time. It is in some cases helpful to rewrite quadratic forms in terms of the individual elements that make up $\mathbf{A}$ and $\mathbf{x}$:

$$\mathbf{x}^\top \mathbf{A}\mathbf{x} = \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij} x_i x_j$$

This identity is not hard to show, but the derivation is somewhat tedious, so we omit it. The result can be used, for example, to derive $\nabla_{\mathbf{x}}(\mathbf{x}^\top \mathbf{A}\mathbf{x})$, as well as to prove that all the diagonal entries of a positive-definite matrix are positive.

3.12.1 Rayleigh quotients

There turns out to be an interesting connection between the quadratic form of a symmetric matrix and its eigenvalues. This connection is provided by the **Rayleigh quotient**

$$R_{\mathbf{A}}(\mathbf{x}) = \frac{\mathbf{x}^\top \mathbf{A}\mathbf{x}}{\mathbf{x}^\top \mathbf{x}}$$

The Rayleigh quotient has a couple of important properties which the reader can (and should!) easily verify from the definition:

(i) **Scale invariance**: for any vector $\mathbf{x} \neq \mathbf{0}$ and any scalar $\alpha \neq 0$, $R_{\mathbf{A}}(\mathbf{x}) = R_{\mathbf{A}}(\alpha\mathbf{x})$.

(ii) If $\mathbf{x}$ is an eigenvector of $\mathbf{A}$ with eigenvalue λ, then $R_{\mathbf{A}}(\mathbf{x}) = \lambda$.

We can further show that the Rayleigh quotient is bounded by the largest and smallest eigenvalues of $\mathbf{A}$. But first we will show a useful special case of the final result.

Proposition 4. *For any* $\mathbf{x}$ *such that* $\|\mathbf{x}\|_2 = 1$*,*

$$\lambda_{\min}(\mathbf{A}) \leq \mathbf{x}^\top \mathbf{A}\mathbf{x} \leq \lambda_{\max}(\mathbf{A})$$

with equality if and only if $\mathbf{x}$ *is a corresponding eigenvector.*

Proof. We show only the max case because the argument for the min case is entirely analogous.

Since $\mathbf{A}$ is symmetric, we can decompose it as $\mathbf{A} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^\top$. Then use the change of variable $\mathbf{y} = \mathbf{Q}^\top\mathbf{x}$, noting that the relationship between $\mathbf{x}$ and $\mathbf{y}$ is one-to-one and that $\|\mathbf{y}\|_2 = 1$ since $\mathbf{Q}$ is orthogonal. Hence

$$\max_{\|\mathbf{x}\|_2=1} \mathbf{x}^\top \mathbf{A}\mathbf{x} = \max_{\|\mathbf{y}\|_2=1} \mathbf{y}^\top \mathbf{\Lambda}\mathbf{y} = \max_{y_1^2+\cdots+y_n^2=1} \sum_{i=1}^n \lambda_i y_i^2$$

Written this way, it is clear that $\mathbf{y}$ maximizes this expression exactly if and only if it satisfies $\sum_{i\in I} y_i^2 = 1$ where $I = \{i : \lambda_i = \max_{j=1,\ldots,n} \lambda_j = \lambda_{\max}(\mathbf{A})\}$ and $y_j = 0$ for $j \notin I$. That is, I contains the index or indices of the largest eigenvalue. In this case, the maximal value of the expression is

$$\sum_{i=1}^n \lambda_i y_i^2 = \sum_{i\in I} \lambda_i y_i^2 = \lambda_{\max}(\mathbf{A}) \sum_{i\in I} y_i^2 = \lambda_{\max}(\mathbf{A})$$

Then writing $\mathbf{q}_1, \ldots, \mathbf{q}_n$ for the columns of $\mathbf{Q}$, we have

$$\mathbf{x} = \mathbf{Q}\mathbf{Q}^\top\mathbf{x} = \mathbf{Q}\mathbf{y} = \sum_{i=1}^n y_i \mathbf{q}_i = \sum_{i\in I} y_i \mathbf{q}_i$$

where we have used the matrix-vector product identity.

Recall that $\mathbf{q}_1, \ldots, \mathbf{q}_n$ are eigenvectors of $\mathbf{A}$ and form an orthonormal basis for $\mathbb{R}^n$. Therefore by construction, the set $\{\mathbf{q}_i : i \in I\}$ forms an orthonormal basis for the eigenspace of $\lambda_{\max}(\mathbf{A})$. Hence $\mathbf{x}$, which is a linear combination of these, lies in that eigenspace and thus is an eigenvector of $\mathbf{A}$ corresponding to $\lambda_{\max}(\mathbf{A})$.

We have shown that $\max_{\|\mathbf{x}\|_2=1} \mathbf{x}^\top \mathbf{A}\mathbf{x} = \lambda_{\max}(\mathbf{A})$, from which we have the general inequality $\mathbf{x}^\top \mathbf{A}\mathbf{x} \leq \lambda_{\max}(\mathbf{A})$ for all unit-length $\mathbf{x}$. □

By the scale invariance of the Rayleigh quotient, we immediately have as a corollary (since $\mathbf{x}^\top \mathbf{A}\mathbf{x} = R_{\mathbf{A}}(\mathbf{x})$ for unit $\mathbf{x}$)

Theorem 3. *(Min-max theorem) For all* $\mathbf{x} \neq \mathbf{0}$*,*

$$\lambda_{\min}(\mathbf{A}) \leq R_{\mathbf{A}}(\mathbf{x}) \leq \lambda_{\max}(\mathbf{A})$$

with equality if and only if $\mathbf{x}$ *is a corresponding eigenvector.*

3.12.2 The geometry of positive definite quadratic forms

A useful way to understand quadratic forms is by the geometry of their level sets. Recall that a **level set** or **isocontour** of a function is the set of all inputs such that the function applied to those inputs yields a given output. Mathematically, the c-isocontour of f is $\{\mathbf{x} \in \operatorname{dom} f : f(\mathbf{x}) = c\}$.

Let us consider the special case $f(\mathbf{x}) = \mathbf{x}^\top \mathbf{A}\mathbf{x}$ where $\mathbf{A}$ is a positive definite matrix. Since $\mathbf{A}$ is positive definite, it has a unique matrix square root $\mathbf{A}^{\frac{1}{2}} = \mathbf{Q}\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{Q}^\top$, where $\mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^\top = \mathbf{A}$ is the eigendecomposition of $\mathbf{A}$ and $\mathbf{\Lambda}^{\frac{1}{2}} = \operatorname{diag}(\sqrt{\lambda_1}, \dots \sqrt{\lambda_d})$. It is easy to see that this matrix $\mathbf{A}^{\frac{1}{2}}$ is positive definite and satisfies $\mathbf{A}^{\frac{1}{2}}\mathbf{A}^{\frac{1}{2}} = \mathbf{A}$. Fixing a value $c \geq 0$, the c-isocontour of f is the set of $\mathbf{x} \in \mathbb{R}^d$ such that

$$c = \mathbf{x}^\top \mathbf{A}\mathbf{x} = \mathbf{x}^\top \mathbf{A}^{\frac{1}{2}}\mathbf{A}^{\frac{1}{2}}\mathbf{x} = \|\mathbf{A}^{\frac{1}{2}}\mathbf{x}\|_2^2$$

where we have used the symmetry of $\mathbf{A}^{\frac{1}{2}}$. Making the change of variable $\mathbf{z} = \mathbf{A}^{\frac{1}{2}}\mathbf{x}$, we have the condition $\|\mathbf{z}\|_2 = \sqrt{c}$. That is, the values $\mathbf{z}$ lie on a sphere of radius $\sqrt{c}$. These can be parameterized as $\mathbf{z} = \sqrt{c}\hat{\mathbf{z}}$ where $\hat{\mathbf{z}}$ has $\|\hat{\mathbf{z}}\|_2 = 1$. Then since $\mathbf{A}^{-\frac{1}{2}} = \mathbf{Q}\mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{Q}^\top$, we have

$$\mathbf{x} = \mathbf{A}^{-\frac{1}{2}}\mathbf{z} = \mathbf{Q}\mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{Q}^\top\sqrt{c}\hat{\mathbf{z}} = \mathbf{Q}(\sqrt{c}\mathbf{\Lambda}^{-\frac{1}{2}})\tilde{\mathbf{z}}$$

where $\tilde{\mathbf{z}} = \mathbf{Q}^\top\hat{\mathbf{z}}$ also satisfies $\|\tilde{\mathbf{z}}\|_2 = 1$ since $\mathbf{Q}$ is orthogonal. Using this parameterization, we see that the solution set $\{\mathbf{x} \in \mathbb{R}^d : f(\mathbf{x}) = c\}$ is the image of the unit sphere $\{\tilde{\mathbf{z}} \in \mathbb{R}^d : \|\tilde{\mathbf{z}}\|_2 = 1\}$ under the invertible linear map $\mathbf{x} = \mathbf{Q}(\sqrt{c}\mathbf{\Lambda}^{-\frac{1}{2}})\tilde{\mathbf{z}}$.

What we have gained with all these manipulations is a clear algebraic understanding of the c-isocontour of f in terms of a sequence of linear transformations applied to a well-understood set. We begin with the unit sphere, then scale every axis i by $\sqrt{c}\lambda_i^{-\frac{1}{2}}$, resulting in an axis-aligned ellipsoid. Observe that the axis lengths of the ellipsoid are proportional to the inverse square roots of the eigenvalues of $\mathbf{A}$. Hence larger eigenvalues correspond to shorter axis lengths, and vice-versa.

Then this axis-aligned ellipsoid undergoes a rigid transformation (i.e. one that preserves length and angles, such as a rotation/reflection) given by $\mathbf{Q}$. The result of this transformation is that the axes of the ellipse are no longer along the coordinate axes in general, but rather along the directions given by the corresponding eigenvectors. To see this, consider the unit vector $\mathbf{e}_i \in \mathbb{R}^d$ that has $[\mathbf{e}_i]_j = \delta_{ij}$. In the pre-transformed space, this vector points along the axis with length proportional to $\lambda_i^{-\frac{1}{2}}$. But after applying the rigid transformation $\mathbf{Q}$, the resulting vector points in the direction of the corresponding eigenvector $\mathbf{q}_i$, since

$$\mathbf{Q}\mathbf{e}_i = \sum_{j=1}^{d} [\mathbf{e}_i]_j \mathbf{q}_j = \mathbf{q}_i$$

where we have used the matrix-vector product identity from earlier.

In summary: the isocontours of $f(\mathbf{x}) = \mathbf{x}^\top \mathbf{A}\mathbf{x}$ are ellipsoids such that the axes point in the directions of the eigenvectors of $\mathbf{A}$, and the radii of these axes are proportional to the inverse square roots of the corresponding eigenvalues.

4 Calculus and Optimization

Much of machine learning is about minimizing a **cost function** (also called an **objective function** in the optimization community), which is a scalar function of several variables that typically measures how poorly our model fits the data we have.

4.1 Extrema

Optimization is about finding **extrema**, which depending on the application could be minima or maxima. When defining extrema, it is necessary to consider the set of inputs over which we're optimizing. This set $\mathcal{X} \subseteq \mathbb{R}^d$ is called the **feasible set**. If $\mathcal{X}$ is the entire domain of the function being optimized (as it often will be for our purposes), we say that the problem is **unconstrained**. Otherwise the problem is **constrained** and may be much harder to solve, depending on the nature of the feasible set.

Suppose $f : \mathbb{R}^d \to \mathbb{R}$. A point $\mathbf{x}$ is said to be a **local minimum** (resp. **local maximum**) of f in $\mathcal{X}$ if $f(\mathbf{x}) \leq f(\mathbf{y})$ (resp. $f(\mathbf{x}) \geq f(\mathbf{y})$) for all $\mathbf{y}$ in some neighborhood $\mathcal{N} \subseteq \mathcal{X}$ that contains $\mathbf{x}$. Furthermore, if $f(\mathbf{x}) \leq f(\mathbf{y})$ for all $\mathbf{y} \in \mathcal{X}$, then $\mathbf{x}$ is a **global minimum** of f in $\mathcal{X}$ (similarly for global maximum). If the phrase "in $\mathcal{X}$" is unclear from context, assume we are optimizing over the whole domain of the function.

The qualifier **strict** (as in e.g. a strict local minimum) means that the inequality sign in the definition is actually a $>$ or $<$, with equality not allowed. This indicates that the extremum is unique.

Observe that maximizing a function f is equivalent to minimizing $-f$, so optimization problems are typically phrased in terms of minimization without loss of generality. This convention (which we follow here) eliminates the need to discuss minimization and maximization separately.

4.2 Gradients

The single most important concept from calculus in the context of machine learning is the **gradient**. Gradients generalize derivatives to scalar functions of several variables. The gradient of $f : \mathbb{R}^d \to \mathbb{R}$, denoted ∇f, is given by

$$\nabla f = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix} \qquad \text{i.e.} \qquad [\nabla f]_i = \frac{\partial f}{\partial x_i}$$

Gradients have the following very important property: $\nabla f(\mathbf{x})$ points in the direction of **steepest ascent** from $\mathbf{x}$. Similarly, $-\nabla f(\mathbf{x})$ points in the direction of **steepest descent** from $\mathbf{x}$. We will use this fact frequently when iteratively minimizing a function via **gradient descent**.

4.3 The Jacobian

The **Jacobian** of $f : \mathbb{R}^n \to \mathbb{R}^m$ is a matrix of first-order partial derivatives:

$$\mathbf{J}_f = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix} \qquad \text{i.e.} \qquad [\mathbf{J}_f]_{ij} = \frac{\partial f_i}{\partial x_j}$$

Note the special case $m = 1$, where $\nabla f = \mathbf{J}_f^\top$.

4.4 The Hessian

The **Hessian** matrix of $f : \mathbb{R}^d \to \mathbb{R}$ is a matrix of second-order partial derivatives:

$$\nabla^2 f = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix} \quad \text{i.e.} \quad [\nabla^2 f]_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j}$$

Recall that if the partial derivatives are continuous, the order of differentiation can be interchanged (Clairaut's theorem), so the Hessian matrix will be symmetric. This will typically be the case for differentiable functions that we work with.

The Hessian is used in some optimization algorithms such as Newton's method. It is expensive to calculate but can drastically reduce the number of iterations needed to converge to a local minimum by providing information about the curvature of f.

4.5 Matrix calculus

Since a lot of optimization reduces to finding points where the gradient vanishes, it is useful to have differentiation rules for matrix and vector expressions. We give some common rules here. Probably the two most important for our purposes are

$$\nabla_{\mathbf{x}}(\mathbf{a}^\top \mathbf{x}) = \mathbf{a}$$
$$\nabla_{\mathbf{x}}(\mathbf{x}^\top \mathbf{A}\mathbf{x}) = (\mathbf{A} + \mathbf{A}^\top)\mathbf{x}$$

Note that this second rule is defined only if $\mathbf{A}$ is square. Furthermore, if $\mathbf{A}$ is symmetric, we can simplify the result to $2\mathbf{A}\mathbf{x}$.

4.5.1 The chain rule

Most functions that we wish to optimize are not completely arbitrary functions, but rather are composed of simpler functions which we know how to handle. The chain rule gives us a way to calculate derivatives for a composite function in terms of the derivatives of the simpler functions that make it up.

The chain rule from single-variable calculus should be familiar:

$$(f \circ g)'(x) = f'(g(x))g'(x)$$

where $\circ$ denotes function composition. There is a natural generalization of this rule to multivariate functions.

Proposition 5. *Suppose $f : \mathbb{R}^m \to \mathbb{R}^k$ and $g : \mathbb{R}^n \to \mathbb{R}^m$. Then $f \circ g : \mathbb{R}^n \to \mathbb{R}^k$ and*

$$\mathbf{J}_{f \circ g}(\mathbf{x}) = \mathbf{J}_f(g(\mathbf{x}))\mathbf{J}_g(\mathbf{x})$$

In the special case $k = 1$ we have the following corollary since $\nabla f = \mathbf{J}_f^\top$.

Corollary 1. *Suppose $f : \mathbb{R}^m \to \mathbb{R}$ and $g : \mathbb{R}^n \to \mathbb{R}^m$. Then $f \circ g : \mathbb{R}^n \to \mathbb{R}$ and*

$$\nabla(f \circ g)(\mathbf{x}) = \mathbf{J}_g(\mathbf{x})^\top \nabla f(g(\mathbf{x}))$$

4.6 Taylor's theorem

Taylor's theorem has natural generalizations to functions of more than one variable. We give the version presented in [1].

Theorem 4. *(Taylor's theorem) Suppose $f : \mathbb{R}^d \to \mathbb{R}$ is continuously differentiable, and let $\mathbf{h} \in \mathbb{R}^d$. Then there exists $t \in (0, 1)$ such that*

$$f(\mathbf{x} + \mathbf{h}) = f(\mathbf{x}) + \nabla f(\mathbf{x} + t\mathbf{h})^\top \mathbf{h}$$

Furthermore, if f is twice continuously differentiable, then

$$\nabla f(\mathbf{x} + \mathbf{h}) = \nabla f(\mathbf{x}) + \int_0^1 \nabla^2 f(\mathbf{x} + t\mathbf{h})\mathbf{h}\, \mathrm{d}t$$

and there exists $t \in (0, 1)$ such that

$$f(\mathbf{x} + \mathbf{h}) = f(\mathbf{x}) + \nabla f(\mathbf{x})^\top \mathbf{h} + \frac{1}{2}\mathbf{h}^\top \nabla^2 f(\mathbf{x} + t\mathbf{h})\mathbf{h}$$

This theorem is used in proofs about conditions for local minima of unconstrained optimization problems. Some of the most important results are given in the next section.

4.7 Conditions for local minima

Proposition 6. *If $\mathbf{x}^*$ is a local minimum of f and f is continuously differentiable in a neighborhood of $\mathbf{x}^*$, then $\nabla f(\mathbf{x}^*) = \mathbf{0}$.*

Proof. Let $\mathbf{x}^*$ be a local minimum of f, and suppose towards a contradiction that $\nabla f(\mathbf{x}^*) \neq \mathbf{0}$. Let $\mathbf{h} = -\nabla f(\mathbf{x}^*)$, noting that by the continuity of ∇f we have

$$\lim_{t \to 0} -\nabla f(\mathbf{x}^* + t\mathbf{h}) = -\nabla f(\mathbf{x}^*) = \mathbf{h}$$

Hence

$$\lim_{t \to 0} \mathbf{h}^\top \nabla f(\mathbf{x}^* + t\mathbf{h}) = \mathbf{h}^\top \nabla f(\mathbf{x}^*) = -\|\mathbf{h}\|_2^2 < 0$$

Thus there exists $T > 0$ such that $\mathbf{h}^\top \nabla f(\mathbf{x}^* + t\mathbf{h}) < 0$ for all $t \in [0, T]$. Now we apply Taylor's theorem: for any $t \in (0, T]$, there exists $t' \in (0, t)$ such that

$$f(\mathbf{x}^* + t\mathbf{h}) = f(\mathbf{x}^*) + t\mathbf{h}^\top \nabla f(\mathbf{x}^* + t'\mathbf{h}) < f(\mathbf{x}^*)$$

whence it follows that $\mathbf{x}^*$ is not a local minimum, a contradiction. Hence $\nabla f(\mathbf{x}^*) = \mathbf{0}$. □

The proof shows us why the vanishing gradient is necessary for an extremum: if $\nabla f(\mathbf{x})$ is nonzero, there always exists a sufficiently small step $\alpha > 0$ such that $f(\mathbf{x} - \alpha \nabla f(\mathbf{x}))) < f(\mathbf{x})$. For this reason, $-\nabla f(\mathbf{x})$ is called a **descent direction**.

Points where the gradient vanishes are called **stationary points**. Note that not all stationary points are extrema. Consider $f : \mathbb{R}^2 \to \mathbb{R}$ given by $f(x, y) = x^2 - y^2$. We have $\nabla f(\mathbf{0}) = \mathbf{0}$, but the point $\mathbf{0}$ is the minimum along the line $y = 0$ and the maximum along the line $x = 0$. Thus it is neither a local minimum nor a local maximum of f. Points such as these, where the gradient vanishes but there is no local extremum, are called **saddle points**.

We have seen that first-order information (i.e. the gradient) is insufficient to characterize local minima. But we can say more with second-order information (i.e. the Hessian). First we prove a necessary second-order condition for local minima.

Proposition 7. *If $\mathbf{x}^*$ is a local minimum of f and f is twice continuously differentiable in a neighborhood of $\mathbf{x}^*$, then $\nabla^2 f(\mathbf{x}^*)$ is positive semi-definite.*

Proof. Let $\mathbf{x}^*$ be a local minimum of f, and suppose towards a contradiction that $\nabla^2 f(\mathbf{x}^*)$ is not positive semi-definite. Let $\mathbf{h}$ be such that $\mathbf{h}^\top \nabla^2 f(\mathbf{x}^*)\mathbf{h} < 0$, noting that by the continuity of $\nabla^2 f$ we have

$$\lim_{t\to 0} \nabla^2 f(\mathbf{x}^* + t\mathbf{h}) = \nabla^2 f(\mathbf{x}^*)$$

Hence

$$\lim_{t\to 0} \mathbf{h}^\top \nabla^2 f(\mathbf{x}^* + t\mathbf{h})\mathbf{h} = \mathbf{h}^\top \nabla^2 f(\mathbf{x}^*)\mathbf{h} < 0$$

Thus there exists $T > 0$ such that $\mathbf{h}^\top \nabla^2 f(\mathbf{x}^* + t\mathbf{h})\mathbf{h} < 0$ for all $t \in [0, T]$. Now we apply Taylor's theorem: for any $t \in (0, T]$, there exists $t' \in (0, t)$ such that

$$f(\mathbf{x}^* + t\mathbf{h}) = f(\mathbf{x}^*) + \underbrace{t\mathbf{h}^\top \nabla f(\mathbf{x}^*)}_{0} + \frac{1}{2}t^2 \mathbf{h}^\top \nabla^2 f(\mathbf{x}^* + t'\mathbf{h})\mathbf{h} < f(\mathbf{x}^*)$$

where the middle term vanishes because $\nabla f(\mathbf{x}^*) = \mathbf{0}$ by the previous result. It follows that $\mathbf{x}^*$ is not a local minimum, a contradiction. Hence $\nabla^2 f(\mathbf{x}^*)$ is positive semi-definite. □

Now we give sufficient conditions for local minima.

Proposition 8. *Suppose f is twice continuously differentiable with $\nabla^2 f$ positive semi-definite in a neighborhood of $\mathbf{x}^*$, and that $\nabla f(\mathbf{x}^*) = \mathbf{0}$. Then $\mathbf{x}^*$ is a local minimum of f. Furthermore if $\nabla^2 f(\mathbf{x}^*)$ is positive definite, then $\mathbf{x}^*$ is a strict local minimum.*

Proof. Let $\mathcal{B}$ be an open ball of radius $r > 0$ centered at $\mathbf{x}^*$ which is contained in the neighborhood. Applying Taylor's theorem, we have that for any $\mathbf{h}$ with $\|\mathbf{h}\|_2 < r$, there exists $t \in (0, 1)$ such that

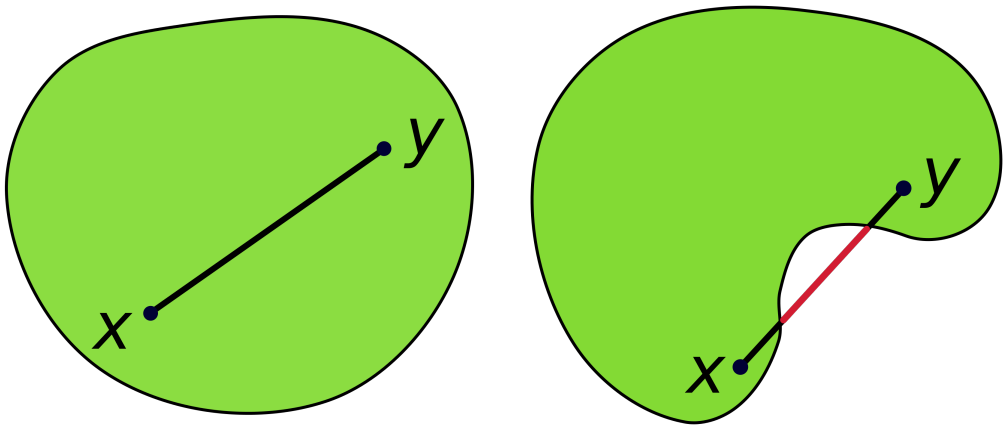
$$f(\mathbf{x}^* + \mathbf{h}) = f(\mathbf{x}^*) + \underbrace{\mathbf{h}^\top \nabla f(\mathbf{x}^*)}_{0} + \frac{1}{2}\mathbf{h}^\top \nabla^2 f(\mathbf{x}^* + t\mathbf{h})\mathbf{h} \geq f(\mathbf{x}^*)$$

The last inequality holds because $\nabla^2 f(\mathbf{x}^* + t\mathbf{h})$ is positive semi-definite (since $\|t\mathbf{h}\|_2 = t\|\mathbf{h}\|_2 < \|\mathbf{h}\|_2 < r$), so $\mathbf{h}^\top \nabla^2 f(\mathbf{x}^* + t\mathbf{h})\mathbf{h} \geq 0$. Since $f(\mathbf{x}^*) \leq f(\mathbf{x}^* + \mathbf{h})$ for all directions $\mathbf{h}$ with $\|\mathbf{h}\|_2 < r$, we conclude that $\mathbf{x}^*$ is a local minimum.

Now further suppose that $\nabla^2 f(\mathbf{x}^*)$ is strictly positive definite. Since the Hessian is continuous we can choose another ball $\mathcal{B}'$ with radius $r' > 0$ centered at $\mathbf{x}^*$ such that $\nabla^2 f(\mathbf{x})$ is positive definite for all $\mathbf{x} \in \mathcal{B}'$. Then following the same argument as above (except with a strict inequality now since the Hessian is positive definite) we have $f(\mathbf{x}^* + \mathbf{h}) > f(\mathbf{x}^*)$ for all $\mathbf{h}$ with $0 < \|\mathbf{h}\|_2 < r'$. Hence $\mathbf{x}^*$ is a strict local minimum. □

Note that, perhaps counterintuitively, the conditions $\nabla f(\mathbf{x}^*) = \mathbf{0}$ and $\nabla^2 f(\mathbf{x}^*)$ positive semi-definite are not enough to guarantee a local minimum at $\mathbf{x}^*$! Consider the function $f(x) = x^3$. We have $f'(0) = 0$ and $f''(0) = 0$ (so the Hessian, which in this case is the 1×1 matrix $\begin{bmatrix}0\end{bmatrix}$, is positive semi-definite). But f has a saddle point at $x = 0$. The function $f(x) = -x^4$ is an even worse offender – it has the same gradient and Hessian at $x = 0$, but $x = 0$ is a strict local maximum for this function!

For these reasons we require that the Hessian remains positive semi-definite as long as we are close to $\mathbf{x}^*$. Unfortunately, this condition is not practical to check computationally, but in some cases we can verify it analytically (usually by showing that $\nabla^2 f(\mathbf{x})$ is p.s.d. for all $\mathbf{x} \in \mathbb{R}^d$). Also, if $\nabla^2 f(\mathbf{x}^*)$ is strictly positive definite, the continuity assumption on f implies this condition, so we don't have to worry.



(a) A convex set (b) A non-convex set

Figure 1: What convex sets look like

4.8 Convexity

Convexity is a term that pertains to both sets and functions. For functions, there are different degrees of convexity, and how convex a function is tells us a lot about its minima: do they exist, are they unique, how quickly can we find them using optimization algorithms, etc. In this section, we present basic results regarding convexity, strict convexity, and strong convexity.

4.8.1 Convex sets

A set $\mathcal{X} \subseteq \mathbb{R}^d$ is **convex** if

$$t\mathbf{x} + (1-t)\mathbf{y} \in \mathcal{X}$$

for all $\mathbf{x}, \mathbf{y} \in \mathcal{X}$ and all $t \in [0, 1]$.

Geometrically, this means that all the points on the line segment between any two points in $\mathcal{X}$ are also in $\mathcal{X}$. See Figure 1 for a visual.

Why do we care whether or not a set is convex? We will see later that the nature of minima can depend greatly on whether or not the feasible set is convex. Undesirable pathological results can occur when we allow the feasible set to be arbitrary, so for proofs we will need to assume that it is convex. Fortunately, we often want to minimize over all of $\mathbb{R}^d$, which is easily seen to be a convex set.

4.8.2 Basics of convex functions

In the remainder of this section, assume $f : \mathbb{R}^d \to \mathbb{R}$ unless otherwise noted. We'll start with the definitions and then give some results.

A function f is **convex** if

$$f(t\mathbf{x} + (1-t)\mathbf{y}) \leq tf(\mathbf{x}) + (1-t)f(\mathbf{y})$$

for all $\mathbf{x}, \mathbf{y} \in \operatorname{dom} f$ and all $t \in [0, 1]$.

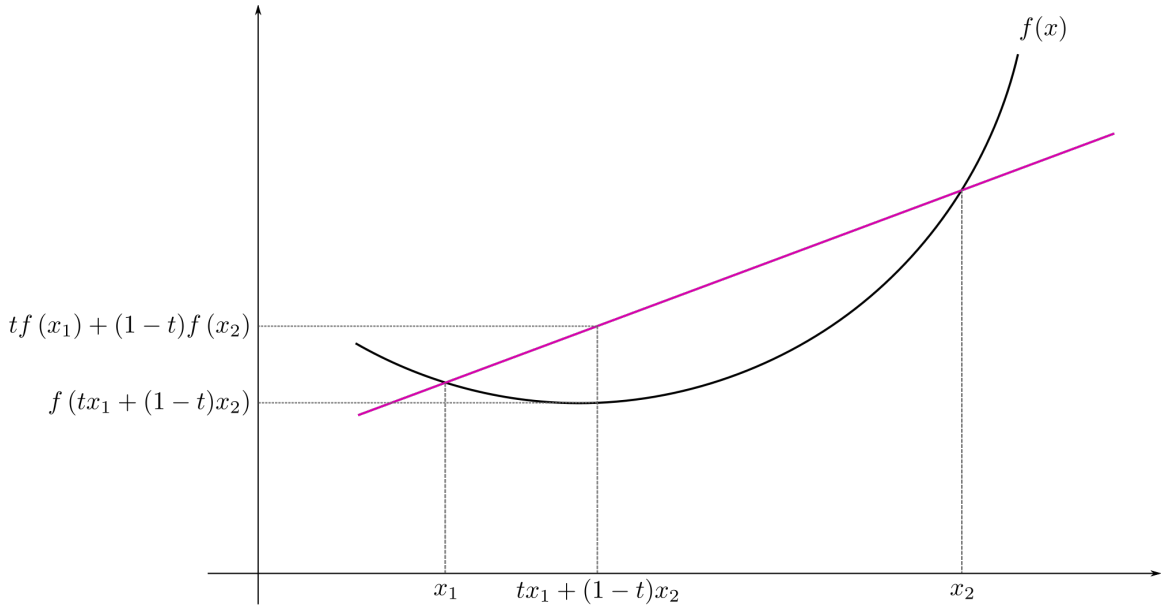


Figure 2: What convex functions look like

If the inequality holds strictly (i.e. $<$ rather than $\leq$) for all $t \in (0,1)$ and $\mathbf{x} \neq \mathbf{y}$, then we say that f is **strictly convex**.

A function f is **strongly convex with parameter** m (or m**-strongly convex**) if the function

$$\mathbf{x} \mapsto f(\mathbf{x}) - \frac{m}{2}\|\mathbf{x}\|_2^2$$

is convex.

These conditions are given in increasing order of strength; strong convexity implies strict convexity which implies convexity.

Geometrically, convexity means that the line segment between two points on the graph of f lies on or above the graph itself. See Figure 2 for a visual.

Strict convexity means that the graph of f lies strictly above the line segment, except at the segment endpoints. (So actually the function in the figure appears to be strictly convex.)

4.8.3 Consequences of convexity

Why do we care if a function is (strictly/strongly) convex?

Basically, our various notions of convexity have implications about the nature of minima. It should not be surprising that the stronger conditions tell us more about the minima.

Proposition 9. *Let $\mathcal{X}$ be a convex set. If f is convex, then any local minimum of f in $\mathcal{X}$ is also a global minimum.*

Proof. Suppose f is convex, and let $\mathbf{x}^*$ be a local minimum of f in $\mathcal{X}$. Then for some neighborhood $\mathcal{N} \subseteq \mathcal{X}$ about $\mathbf{x}^*$, we have $f(\mathbf{x}) \geq f(\mathbf{x}^*)$ for all $\mathbf{x} \in \mathcal{N}$. Suppose towards a contradiction that there exists $\tilde{\mathbf{x}} \in \mathcal{X}$ such that $f(\tilde{\mathbf{x}}) < f(\mathbf{x}^*)$.

Consider the line segment $\mathbf{x}(t) = t\mathbf{x}^* + (1-t)\tilde{\mathbf{x}},\ t \in [0,1]$, noting that $\mathbf{x}(t) \in \mathcal{X}$ by the convexity of $\mathcal{X}$. Then by the convexity of f,

$$f(\mathbf{x}(t)) \leq tf(\mathbf{x}^*) + (1-t)f(\tilde{\mathbf{x}}) < tf(\mathbf{x}^*) + (1-t)f(\mathbf{x}^*) = f(\mathbf{x}^*)$$

for all $t \in (0,1)$.

We can pick t to be sufficiently close to 1 that $\mathbf{x}(t) \in \mathcal{N}$; then $f(\mathbf{x}(t)) \geq f(\mathbf{x}^*)$ by the definition of $\mathcal{N}$, but $f(\mathbf{x}(t)) < f(\mathbf{x}^*)$ by the above inequality, a contradiction.

It follows that $f(\mathbf{x}^*) \leq f(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}$, so $\mathbf{x}^*$ is a global minimum of f in $\mathcal{X}$. □

Proposition 10. *Let $\mathcal{X}$ be a convex set. If f is strictly convex, then there exists at most one local minimum of f in $\mathcal{X}$. Consequently, if it exists it is the unique global minimum of f in $\mathcal{X}$.*

Proof. The second sentence follows from the first, so all we must show is that if a local minimum exists in $\mathcal{X}$ then it is unique.

Suppose $\mathbf{x}^*$ is a local minimum of f in $\mathcal{X}$, and suppose towards a contradiction that there exists a local minimum $\tilde{\mathbf{x}} \in \mathcal{X}$ such that $\tilde{\mathbf{x}} \neq \mathbf{x}^*$.

Since f is strictly convex, it is convex, so $\mathbf{x}^*$ and $\tilde{\mathbf{x}}$ are both global minima of f in $\mathcal{X}$ by the previous result. Hence $f(\mathbf{x}^*) = f(\tilde{\mathbf{x}})$. Consider the line segment $\mathbf{x}(t) = t\mathbf{x}^* + (1-t)\tilde{\mathbf{x}},\ t \in [0,1]$, which again must lie entirely in $\mathcal{X}$. By the strict convexity of f,

$$f(\mathbf{x}(t)) < tf(\mathbf{x}^*) + (1-t)f(\tilde{\mathbf{x}}) = tf(\mathbf{x}^*) + (1-t)f(\mathbf{x}^*) = f(\mathbf{x}^*)$$

for all $t \in (0,1)$. But this contradicts the fact that $\mathbf{x}^*$ is a global minimum. Therefore if $\tilde{\mathbf{x}}$ is a local minimum of f in $\mathcal{X}$, then $\tilde{\mathbf{x}} = \mathbf{x}^*$, so $\mathbf{x}^*$ is the unique minimum in $\mathcal{X}$. □

It is worthwhile to examine how the feasible set affects the optimization problem. We will see why the assumption that $\mathcal{X}$ is convex is needed in the results above.

Consider the function $f(x) = x^2$, which is a strictly convex function. The unique global minimum of this function in $\mathbb{R}$ is $x = 0$. But let's see what happens when we change the feasible set $\mathcal{X}$.

(i) $\mathcal{X} = \{1\}$: This set is actually convex, so we still have a unique global minimum. But it is not the same as the unconstrained minimum!

(ii) $\mathcal{X} = \mathbb{R} \setminus \{0\}$: This set is non-convex, and we can see that f has no minima in $\mathcal{X}$. For any point $x \in \mathcal{X}$, one can find another point $y \in \mathcal{X}$ such that $f(y) < f(x)$.

(iii) $\mathcal{X} = (-\infty, -1] \cup [0, \infty)$: This set is non-convex, and we can see that there is a local minimum ($x = -1$) which is distinct from the global minimum ($x = 0$).

(iv) $\mathcal{X} = (-\infty, -1] \cup [1, \infty)$: This set is non-convex, and we can see that there are two global minima ($x = \pm 1$).

4.8.4 Showing that a function is convex

Hopefully the previous section has convinced the reader that convexity is an important property. Next we turn to the issue of showing that a function is (strictly/strongly) convex. It is of course possible (in principle) to directly show that the condition in the definition holds, but this is usually not the easiest way.

Proposition 11. *Norms are convex.*

Proof. Let $\|\cdot\|$ be a norm on $\mathbb{R}^d$. Then for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ and $t \in [0,1]$,

$$\|t\mathbf{x} + (1-t)\mathbf{y}\| \leq \|t\mathbf{x}\| + \|(1-t)\mathbf{y}\| = |t|\|\mathbf{x}\| + |1-t|\|\mathbf{y}\| = t\|\mathbf{x}\| + (1-t)\|\mathbf{y}\|$$

where we have used respectively the triangle inequality, the homogeneity of norms, and the fact that t and $1-t$ are nonnegative. Hence $\|\cdot\|$ is convex. □

Proposition 12. *Suppose f is differentiable. Then f is convex if and only if*

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle$$

for all $\mathbf{x}, \mathbf{y} \in \operatorname{dom} f$.

Proof. To-do. □

Proposition 13. *Suppose f is twice differentiable. Then*

(i) f is convex if and only if $\nabla^2 f(\mathbf{x}) \succeq 0$ for all $\mathbf{x} \in \operatorname{dom} f$.

(ii) If $\nabla^2 f(\mathbf{x}) \succ 0$ for all $\mathbf{x} \in \operatorname{dom} f$, then f is strictly convex.

(iii) f is m-strongly convex if and only if $\nabla^2 f(\mathbf{x}) \succeq mI$ for all $\mathbf{x} \in \operatorname{dom} f$.

Proof. Omitted. □

Proposition 14. *If f is convex and $\alpha \geq 0$, then αf is convex.*

Proof. Suppose f is convex and $\alpha \geq 0$. Then for all $\mathbf{x}, \mathbf{y} \in \operatorname{dom}(\alpha f) = \operatorname{dom} f$,

$$\begin{aligned}
(\alpha f)(t\mathbf{x} + (1-t)\mathbf{y}) &= \alpha f(t\mathbf{x} + (1-t)\mathbf{y}) \\
&\leq \alpha \left(tf(\mathbf{x}) + (1-t)f(\mathbf{y}) \right) \\
&= t(\alpha f(\mathbf{x})) + (1-t)(\alpha f(\mathbf{y})) \\
&= t(\alpha f)(\mathbf{x}) + (1-t)(\alpha f)(\mathbf{y})
\end{aligned}$$

so αf is convex. □

Proposition 15. *If f and g are convex, then $f+g$ is convex. Furthermore, if g is strictly convex, then $f+g$ is strictly convex, and if g is m-strongly convex, then $f+g$ is m-strongly convex.*

Proof. Suppose f and g are convex. Then for all $\mathbf{x}, \mathbf{y} \in \operatorname{dom}(f+g) = \operatorname{dom} f \cap \operatorname{dom} g$,

$$\begin{aligned}
(f+g)(t\mathbf{x} + (1-t)\mathbf{y}) &= f(t\mathbf{x} + (1-t)\mathbf{y}) + g(t\mathbf{x} + (1-t)\mathbf{y}) \\
&\leq tf(\mathbf{x}) + (1-t)f(\mathbf{y}) + g(t\mathbf{x} + (1-t)\mathbf{y}) && \text{convexity of } f \\
&\leq tf(\mathbf{x}) + (1-t)f(\mathbf{y}) + tg(\mathbf{x}) + (1-t)g(\mathbf{y}) && \text{convexity of } g \\
&= t(f(\mathbf{x}) + g(\mathbf{x})) + (1-t)(f(\mathbf{y}) + g(\mathbf{y})) \\
&= t(f+g)(\mathbf{x}) + (1-t)(f+g)(\mathbf{y})
\end{aligned}$$

so $f+g$ is convex.

If g is strictly convex, the second inequality above holds strictly for $\mathbf{x} \neq \mathbf{y}$ and $t \in (0,1)$, so $f+g$ is strictly convex.

If g is m-strongly convex, then the function $h(\mathbf{x}) \equiv g(\mathbf{x}) - \frac{m}{2}\|\mathbf{x}\|_2^2$ is convex, so $f+h$ is convex. But

$$(f+h)(\mathbf{x}) \equiv f(\mathbf{x}) + h(\mathbf{x}) \equiv f(\mathbf{x}) + g(\mathbf{x}) - \frac{m}{2}\|\mathbf{x}\|_2^2 \equiv (f+g)(\mathbf{x}) - \frac{m}{2}\|\mathbf{x}\|_2^2$$

so $f+g$ is m-strongly convex. □

Proposition 16. *If $f_1, \ldots, f_n$ are convex and $\alpha_1, \ldots, \alpha_n \geq 0$, then*

$$\sum_{i=1}^{n} \alpha_i f_i$$

is convex.

Proof. Follows from the previous two propositions by induction. □

Proposition 17. *If f is convex, then $g(\mathbf{x}) \equiv f(A\mathbf{x} + \mathbf{b})$ is convex for any appropriately-sized A and $\mathbf{b}$.*

Proof. Suppose f is convex and g is defined like so. Then for all $\mathbf{x}, \mathbf{y} \in \operatorname{dom} g$,

$$\begin{aligned}
g(t\mathbf{x} + (1-t)\mathbf{y}) &= f(A(t\mathbf{x} + (1-t)\mathbf{y}) + \mathbf{b}) \\
&= f(tA\mathbf{x} + (1-t)A\mathbf{y} + \mathbf{b}) \\
&= f(tA\mathbf{x} + (1-t)A\mathbf{y} + t\mathbf{b} + (1-t)\mathbf{b}) \\
&= f(t(A\mathbf{x} + \mathbf{b}) + (1-t)(A\mathbf{y} + \mathbf{b})) \\
&\leq tf(A\mathbf{x} + \mathbf{b}) + (1-t)f(A\mathbf{y} + \mathbf{b}) && \text{convexity of } f \\
&= tg(\mathbf{x}) + (1-t)g(\mathbf{y})
\end{aligned}$$

Thus g is convex. □

Proposition 18. *If f and g are convex, then $h(\mathbf{x}) \equiv \max\{f(\mathbf{x}), g(\mathbf{x})\}$ is convex.*

Proof. Suppose f and g are convex and h is defined like so. Then for all $\mathbf{x}, \mathbf{y} \in \operatorname{dom} h$,

$$\begin{aligned}
h(t\mathbf{x} + (1-t)\mathbf{y}) &= \max\{f(t\mathbf{x} + (1-t)\mathbf{y}), g(t\mathbf{x} + (1-t)\mathbf{y})\} \\
&\leq \max\{tf(\mathbf{x}) + (1-t)f(\mathbf{y}), tg(\mathbf{x}) + (1-t)g(\mathbf{y})\} \\
&\leq \max\{tf(\mathbf{x}), tg(\mathbf{x})\} + \max\{(1-t)f(\mathbf{y}), (1-t)g(\mathbf{y})\} \\
&= t\max\{f(\mathbf{x}), g(\mathbf{x})\} + (1-t)\max\{f(\mathbf{y}), g(\mathbf{y})\} \\
&= th(\mathbf{x}) + (1-t)h(\mathbf{y})
\end{aligned}$$

Note that in the first inequality we have used convexity of f and g plus the fact that $a \leq c, b \leq d$ implies $\max\{a, b\} \leq \max\{c, d\}$. In the second inequality we have used the fact that $\max\{a+b, c+d\} \leq \max\{a, c\} + \max\{b, d\}$.

Thus h is convex. □

4.8.5 Examples

A good way to gain intuition about the distinction between convex, strictly convex, and strongly convex functions is to consider examples where the stronger property fails to hold.

Functions that are convex but not strictly convex:

(i) $f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + \alpha$ for any $\mathbf{w} \in \mathbb{R}^d, \alpha \in \mathbb{R}$. Such a function is called an **affine function**, and it is both convex and concave. (In fact, a function is affine if and only if it is both convex and concave.) Note that linear functions and constant functions are special cases of affine functions.

(ii) $f(\mathbf{x}) = \|\mathbf{x}\|_1$

Functions that are strictly but not strongly convex:

(i) $f(x) = x^4$. This example is interesting because it is strictly convex but you cannot show this fact via a second-order argument (since $f''(0) = 0$).

(ii) $f(x) = \exp(x)$. This example is interesting because it's bounded below but has no local minimum.

(iii) $f(x) = -\log x$. This example is interesting because it's strictly convex but not bounded below.

Functions that are strongly convex:

(i) $f(\mathbf{x}) = \|\mathbf{x}\|_2^2$

5 Probability

Probability theory provides powerful tools for modeling and dealing with uncertainty. It is used extensively in machine learning, particularly to construct and analyze classifiers.

5.1 Basics

Suppose we have some sort of randomized experiment (e.g. a coin toss, die roll) that has a fixed set of possible outcomes. This set is called the **sample space** and denoted Ω.

We would like to define probabilities for some **events**, which are subsets of Ω. The set of events is denoted $\mathcal{F}$.[5]

Then we can define a **probability measure** $\mathbb{P} : \mathcal{F} \to [0, 1]$ which must satisfy

(i) $\mathbb{P}(\Omega) = 1$

(ii) **Countable additivity**: for any countable collection of disjoint sets $\{A_i\} \subseteq \mathcal{F}$,

$$\mathbb{P}\left(\bigcup_i A_i\right) = \sum_i \mathbb{P}(A_i)$$

The triple $(\Omega, \mathcal{F}, \mathbb{P})$ is called a **probability space**.[6]

If $\mathbb{P}(A) = 1$, we say that A occurs **almost surely** (often abbreviated a.s.).[7], and conversely A occurs **almost never** if $\mathbb{P}(A) = 0$.

From these axioms, a number of useful rules can be derived.

Proposition 19. *Let A be an event. Then*

(i) $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$.

(ii) If B is an event and $B \subseteq A$, then $\mathbb{P}(B) \leq \mathbb{P}(A)$.

(iii) $0 = \mathbb{P}(\varnothing) \leq \mathbb{P}(A) \leq \mathbb{P}(\Omega) = 1$

Proof. (i) Using the countable additivity of $\mathbb{P}$, we have

$$\mathbb{P}(A) + \mathbb{P}(A^{\mathrm{c}}) = \mathbb{P}(A \,\dot\cup\, A^{\mathrm{c}}) = \mathbb{P}(\Omega) = 1$$

To show (ii), suppose $B \in \mathcal{F}$ and $B \subseteq A$. Then

$$\mathbb{P}(A) = \mathbb{P}(B \,\dot\cup\, (A \setminus B)) = \mathbb{P}(B) + \mathbb{P}(A \setminus B) \geq \mathbb{P}(B)$$

as claimed.

For (iii): the middle inequality follows from (ii) since $\varnothing \subseteq A \subseteq \Omega$. We also have

$$\mathbb{P}(\varnothing) = \mathbb{P}(\varnothing \,\dot\cup\, \varnothing) = \mathbb{P}(\varnothing) + \mathbb{P}(\varnothing)$$

by countable additivity, which shows $\mathbb{P}(\varnothing) = 0$. □

[5] $\mathcal{F}$ is required to be a σ-algebra for technical reasons; see [2].

[6] Note that a probability space is simply a measure space in which the measure of the whole space equals 1.

[7] This is a probabilist's version of the measure-theoretic term *almost everywhere*.

Proposition 20. *If A and B are events, then $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.*

Proof. The key is to break the events up into their various overlapping and non-overlapping parts.

$$\begin{aligned}\mathbb{P}(A \cup B) &= \mathbb{P}((A \cap B) \dot\cup (A \setminus B) \dot\cup (B \setminus A)) \\ &= \mathbb{P}(A \cap B) + \mathbb{P}(A \setminus B) + \mathbb{P}(B \setminus A) \\ &= \mathbb{P}(A \cap B) + \mathbb{P}(A) - \mathbb{P}(A \cap B) + \mathbb{P}(B) - \mathbb{P}(A \cap B) \\ &= \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)\end{aligned}$$

□

Proposition 21. *If $\{A_i\} \subseteq \mathcal{F}$ is a countable set of events, disjoint or not, then*

$$\mathbb{P}\left(\bigcup_i A_i\right) \leq \sum_i \mathbb{P}(A_i)$$

This inequality is sometimes referred to as **Boole's inequality** or the **union bound**.

Proof. Define $B_1 = A_1$ and $B_i = A_i \setminus (\bigcup_{j<i} A_j)$ for $i > 1$, noting that $\bigcup_{j \leq i} B_j = \bigcup_{j \leq i} A_j$ for all i and the B_i are disjoint. Then

$$\mathbb{P}\left(\bigcup_i A_i\right) = \mathbb{P}\left(\bigcup_i B_i\right) = \sum_i \mathbb{P}(B_i) \leq \sum_i \mathbb{P}(A_i)$$

where the last inequality follows by monotonicity since $B_i \subseteq A_i$ for all i. □

5.1.1 Conditional probability

The **conditional probability** of event A given that event B has occurred is written $\mathbb{P}(A|B)$ and defined as

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

assuming $\mathbb{P}(B) > 0$.[8]

5.1.2 Chain rule

Another very useful tool, the **chain rule**, follows immediately from this definition:

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

5.1.3 Bayes' rule

Taking the equality from above one step further, we arrive at the simple but crucial **Bayes' rule**:

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

[8] In some cases it is possible to define conditional probability on events of probability zero, but this is significantly more technical so we omit it.

It is sometimes beneficial to omit the normalizing constant and write

$$\mathbb{P}(A|B) \propto \mathbb{P}(A)\mathbb{P}(B|A)$$

Under this formulation, $\mathbb{P}(A)$ is often referred to as the **prior**, $\mathbb{P}(A|B)$ as the **posterior**, and $\mathbb{P}(B|A)$ as the **likelihood**.

In the context of machine learning, we can use Bayes' rule to update our "beliefs" (e.g. values of our model parameters) given some data that we've observed.

5.2 Random variables

A **random variable** is some uncertain quantity with an associated probability distribution over the values it can assume.

Formally, a random variable on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is a function[9] $X : \Omega \to \mathbb{R}$.[10]

We denote the range of X by $X(\Omega) = \{X(\omega) : \omega \in \Omega\}$. To give a concrete example (taken from [3]), suppose X is the number of heads in two tosses of a fair coin. The sample space is

$$\Omega = \{hh, tt, ht, th\}$$

and X is determined completely by the outcome ω, i.e. $X = X(\omega)$. For example, the event $X = 1$ is the set of outcomes $\{ht, th\}$.

It is common to talk about the values of a random variable without directly referencing its sample space. The two are related by the following definition: the event that the value of X lies in some set $S \subseteq \mathbb{R}$ is

$$X \in S = \{\omega \in \Omega : X(\omega) \in S\}$$

Note that special cases of this definition include X being equal to, less than, or greater than some specified value. For example

$$\mathbb{P}(X = x) = \mathbb{P}(\{\omega \in \Omega : X(\omega) = x\})$$

A word on notation: we write $p(X)$ to denote the entire probability distribution of X and $p(x)$ for the evaluation of the function p at a particular value $x \in X(\Omega)$. Hopefully this (reasonably standard) abuse of notation is not too distracting. If p is parameterized by some parameters θ, we write $p(X; \theta)$ or $p(x; \theta)$, unless we are in a Bayesian setting where the parameters are considered a random variable, in which case we condition on the parameters.

5.2.1 The cumulative distribution function

The **cumulative distribution function** (c.d.f.) gives the probability that a random variable is at most a certain value:

$$F(x) = \mathbb{P}(X \leq x)$$

The c.d.f. can be used to give the probability that a variable lies within a certain range:

$$\mathbb{P}(a < X \leq b) = F(b) - F(a)$$

[9] The function must be measurable.

[10] More generally, the codomain can be any measurable space, but $\mathbb{R}$ is the most common case by far and sufficient for our purposes.

5.2.2 Discrete random variables

A **discrete random variable** is a random variable that has a countable range and assumes each value in this range with positive probability. Discrete random variables are completely specified by their **probability mass function** (p.m.f.) $p : X(\Omega) \to [0, 1]$ which satisfies

$$\sum_{x \in X(\Omega)} p(x) = 1$$

For a discrete X, the probability of a particular value is given exactly by its p.m.f.:

$$\mathbb{P}(X = x) = p(x)$$

5.2.3 Continuous random variables

A **continuous random variable** is a random variable that has an uncountable range and assumes each value in this range with probability zero. Most of the continuous random variables that one would encounter in practice are **absolutely continuous random variables**[11], which means that there exists a function $p : \mathbb{R} \to [0, \infty)$ that satisfies

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

The function p is called a **probability density function** (abbreviated p.d.f.) and must satisfy

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

The values of this function are not themselves probabilities, since they could exceed 1. However, they do have a couple of reasonable interpretations. One is as relative probabilities; even though the probability of each particular value being picked is technically zero, some points are still in a sense more likely than others.

One can also think of the density as determining the probability that the variable will lie in a small range about a given value. Recall that for small ϵ,

$$\mathbb{P}(x - \epsilon/2 \leq X \leq x + \epsilon/2) = \int_{x-\epsilon/2}^{x+\epsilon/2} p(z) \, \mathrm{d}z \approx \epsilon p(x)$$

using a midpoint approximation to the integral.

Here are some useful identities that follow from the definitions above:

$$\mathbb{P}(a \leq X \leq b) = \int_{a}^{b} p(x) \, \mathrm{d}x$$
$$p(x) = F'(x)$$

5.2.4 Other kinds of random variables

There are random variables that are neither discrete nor continuous. For example, consider a random variable determined as follows: flip a fair coin, then the value is zero if it comes up heads, otherwise draw a number uniformly at random from $[1, 2]$. Such a random variable can take on uncountably many values, but only finitely many of these with positive probability. We will not discuss such random variables because they are rather pathological and require measure theory to analyze.

[11] Random variables that are continuous but not absolutely continuous are called **singular random variables**. We will not discuss them, assuming rather that all continuous random variables admit a density function.

5.3 Joint distributions

Often we have several random variables and we would like to get a distribution over some combination of them. A **joint distribution** is exactly this. For some random variables $X_1, \ldots, X_n$, the joint distribution is written $p(X_1, \ldots, X_n)$ and gives probabilities over entire assignments to all the X_i simultaneously.

5.3.1 Independence of random variables

We say that two variables X and Y are **independent** if their joint distribution factors into their respective distributions, i.e.

$$p(X, Y) = p(X)p(Y)$$

We can also define independence for more than two random variables, although it is more complicated. Let $\{X_i\}_{i \in I}$ be a collection of random variables indexed by I, which may be infinite. Then $\{X_i\}$ are independent if for every finite subset of indices $i_1, \ldots, i_k \in I$ we have

$$p(X_{i_1}, \ldots, X_{i_k}) = \prod_{j=1}^{k} p(X_{i_j})$$

For example, in the case of three random variables, X, Y, Z, we require that $p(X, Y, Z) = p(X)p(Y)p(Z)$ as well as $p(X, Y) = p(X)p(Y)$, $p(X, Z) = p(X)p(Z)$, and $p(Y, Z) = p(Y)p(Z)$.

It is often convenient (though perhaps questionable) to assume that a bunch of random variables are **independent and identically distributed** (i.i.d.) so that their joint distribution can be factored entirely:

$$p(X_1, \ldots, X_n) = \prod_{i=1}^{n} p(X_i)$$

where $X_1, \ldots, X_n$ all share the same p.m.f./p.d.f.

5.3.2 Marginal distributions

If we have a joint distribution over some set of random variables, it is possible to obtain a distribution for a subset of them by "summing out" (or "integrating out" in the continuous case) the variables we don't care about:

$$p(X) = \sum_{y} p(X, y)$$

5.4 Great Expectations

If we have some random variable X, we might be interested in knowing what is the "average" value of X. This concept is captured by the **expected value** (or **mean**) $\mathbb{E}[X]$, which is defined as

$$\mathbb{E}[X] = \sum_{x \in X(\Omega)} xp(x)$$

for discrete X and as

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} xp(x)\,\mathrm{d}x$$

for continuous X.

In words, we are taking a weighted sum of the values that X can take on, where the weights are the probabilities of those respective values. The expected value has a physical interpretation as the "center of mass" of the distribution.

5.4.1 Properties of expected value

A very useful property of expectation is that of linearity:

$$\mathbb{E}\left[\sum_{i=1}^{n} \alpha_i X_i + \beta\right] = \sum_{i=1}^{n} \alpha_i \mathbb{E}[X_i] + \beta$$

Note that this holds even if the X_i are not independent!

But if they are independent, the product rule also holds:

$$\mathbb{E}\left[\prod_{i=1}^{n} X_i\right] = \prod_{i=1}^{n} \mathbb{E}[X_i]$$

5.5 Variance

Expectation provides a measure of the "center" of a distribution, but frequently we are also interested in what the "spread" is about that center. We define the variance $\mathrm{Var}(X)$ of a random variable X by

$$\mathrm{Var}(X) = \mathbb{E}\left[\big(X - \mathbb{E}[X]\big)^2\right]$$

In words, this is the average squared deviation of the values of X from the mean of X. Using a little algebra and the linearity of expectation, it is straightforward to show that

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

5.5.1 Properties of variance

Variance is not linear (because of the squaring in the definition), but one can show the following:

$$\mathrm{Var}(\alpha X + \beta) = \alpha^2 \,\mathrm{Var}(X)$$

Basically, multiplicative constants become squared when they are pulled out, and additive constants disappear (since the variance contributed by a constant is zero).

Furthermore, if $X_1, \ldots, X_n$ are uncorrelated[12], then

$$\mathrm{Var}(X_1 + \cdots + X_n) = \mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_n)$$

5.5.2 Standard deviation

Variance is a useful notion, but it suffers from that fact the units of variance are not the same as the units of the random variable (again because of the squaring). To overcome this problem we can use **standard deviation**, which is defined as $\sqrt{\mathrm{Var}(X)}$. The standard deviation of X has the same units as X.

[12] We haven't defined this yet; see the Correlation section below

5.6 Covariance

Covariance is a measure of the linear relationship between two random variables. We denote the covariance between X and Y as $\text{Cov}(X, Y)$, and it is defined to be

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

Note that the outer expectation must be taken over the joint distribution of X and Y.

Again, the linearity of expectation allows us to rewrite this as

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

Comparing these formulas to the ones for variance, it is not hard to see that $\text{Var}(X) = \text{Cov}(X, X)$.

A useful property of covariance is that of **bilinearity**:

$$\text{Cov}(\alpha X + \beta Y, Z) = \alpha \,\text{Cov}(X, Z) + \beta \,\text{Cov}(Y, Z)$$
$$\text{Cov}(X, \alpha Y + \beta Z) = \alpha \,\text{Cov}(X, Y) + \beta \,\text{Cov}(X, Z)$$

5.6.1 Correlation

Normalizing the covariance gives the **correlation**:

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\,\text{Var}(Y)}}$$

Correlation also measures the linear relationship between two variables, but unlike covariance always lies between -1 and 1.

Two variables are said to be **uncorrelated** if $\text{Cov}(X, Y) = 0$ because $\text{Cov}(X, Y) = 0$ implies that $\rho(X, Y) = 0$. If two variables are independent, then they are uncorrelated, but the converse does not hold in general.

5.7 Random vectors

So far we have been talking about **univariate distributions**, that is, distributions of single variables. But we can also talk about **multivariate distributions** which give distributions of **random vectors**:

$$\mathbf{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}$$

The summarizing quantities we have discussed for single variables have natural generalizations to the multivariate case.

Expectation of a random vector is simply the expectation applied to each component:

$$\mathbb{E}[\mathbf{X}] = \begin{bmatrix} \mathbb{E}[X_1] \\ \vdots \\ \mathbb{E}[X_n] \end{bmatrix}$$

The variance is generalized by the **covariance matrix**:

$$\boldsymbol{\Sigma} = \mathbb{E}[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^\top] = \begin{bmatrix} \mathrm{Var}(X_1) & \mathrm{Cov}(X_1, X_2) & \dots & \mathrm{Cov}(X_1, X_n) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Var}(X_2) & \dots & \mathrm{Cov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_n, X_1) & \mathrm{Cov}(X_n, X_2) & \dots & \mathrm{Var}(X_n) \end{bmatrix}$$

That is, $\Sigma_{ij} = \mathrm{Cov}(X_i, X_j)$. Since covariance is symmetric in its arguments, the covariance matrix is also symmetric. It's also positive semi-definite: for any $\mathbf{x}$,

$$\mathbf{x}^\top \boldsymbol{\Sigma} \mathbf{x} = \mathbf{x}^\top \mathbb{E}[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^\top] \mathbf{x} = \mathbb{E}[\mathbf{x}^\top (\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^\top \mathbf{x}] = \mathbb{E}[((\mathbf{X} - \mathbb{E}[\mathbf{X}])^\top \mathbf{x})^2] \geq 0$$

The inverse of the covariance matrix, $\boldsymbol{\Sigma}^{-1}$, is sometimes called the **precision matrix**.

5.8 Estimation of Parameters

Now we get into some basic topics from statistics. We make some assumptions about our problem by prescribing a **parametric** model (e.g. a distribution that describes how the data were generated), then we fit the parameters of the model to the data. How do we choose the values of the parameters?

5.8.1 Maximum likelihood estimation

A common way to fit parameters is **maximum likelihood estimation** (MLE). The basic principle of MLE is to choose values that "explain" the data best by maximizing the probability/density of the data we've seen as a function of the parameters. Suppose we have random variables $X_1, \dots, X_n$ and corresponding observations $x_1, \dots, x_n$. Then

$$\hat{\theta}_{\text{MLE}} = \arg\max_\theta \mathcal{L}(\theta)$$

where $\mathcal{L}$ is the **likelihood function**

$$\mathcal{L}(\theta) = p(x_1, \dots, x_n; \theta)$$

Often, we assume that $X_1, \dots, X_n$ are i.i.d. Then we can write

$$p(x_1, \dots, x_n; \theta) = \prod_{i=1}^n p(x_i; \theta)$$

At this point, it is usually convenient to take logs, giving rise to the **log-likelihood**

$$\log \mathcal{L}(\theta) = \sum_{i=1}^n \log p(x_i; \theta)$$

This is a valid operation because the probabilities/densities are assumed to be positive, and since log is a monotonically increasing function, it preserves ordering. In other words, any maximizer of $\log \mathcal{L}$ will also maximize $\mathcal{L}$.

For some distributions, it is possible to analytically solve for the maximum likelihood estimator. If $\log \mathcal{L}$ is differentiable, setting the derivatives to zero and trying to solve for θ is a good place to start.

5.8.2 Maximum a posteriori estimation

A more Bayesian way to fit parameters is through **maximum a posteriori estimation** (MAP). In this technique we assume that the parameters are a random variable, and we specify a prior distribution $p(\theta)$. Then we can employ Bayes' rule to compute the posterior distribution of the parameters given the observed data:

$$p(\theta|x_1, \dots, x_n) \propto p(\theta)p(x_1, \dots, x_n|\theta)$$

Computing the normalizing constant is often intractable, because it involves integrating over the parameter space, which may be very high-dimensional. Fortunately, if we just want the MAP estimate, we don't care about the normalizing constant! It does not affect which values of θ maximize the posterior. So we have

$$\hat{\theta}_{\text{MAP}} = \arg\max_{\theta} p(\theta)p(x_1, \dots, x_n|\theta)$$

Again, if we assume the observations are i.i.d., then we can express this in the equivalent, and possibly friendlier, form

$$\hat{\theta}_{\text{MAP}} = \arg\max_{\theta} \left(\log p(\theta) + \sum_{i=1}^{n} \log p(x_i|\theta) \right)$$

A particularly nice case is when the prior is chosen carefully such that the posterior comes from the same family as the prior. In this case the prior is called a **conjugate prior**. For example, if the likelihood is binomial and the prior is beta, the posterior is also beta. There are many conjugate priors; the reader may find this table of conjugate priors useful.

5.9 The Gaussian distribution

There are many distributions, but one of particular importance is the **Gaussian distribution**, also known as the **normal distribution**. It is a continuous distribution, parameterized by its mean $\boldsymbol{\mu} \in \mathbb{R}^d$ and positive-definite covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$, with density

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

Note that in the special case $d = 1$, the density is written in the more recognizable form

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

We write $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ to denote that $\mathbf{X}$ is normally distributed with mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma}$.

5.9.1 The geometry of multivariate Gaussians

The geometry of the multivariate Gaussian density is intimately related to the geometry of positive definite quadratic forms, so make sure the material in that section is well-understood before tackling this section.

First observe that the p.d.f. of the multivariate Gaussian can be rewritten as

$$p(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = g(\tilde{\mathbf{x}}^\top \boldsymbol{\Sigma}^{-1} \tilde{\mathbf{x}})$$

where $\tilde{\mathbf{x}} = \mathbf{x} - \boldsymbol{\mu}$ and $g(z) = [(2\pi)^d \det(\boldsymbol{\Sigma})]^{-\frac{1}{2}} \exp\left(-\frac{z}{2}\right)$. Writing the density in this way, we see that after shifting by the mean $\boldsymbol{\mu}$, the density is really just a simple function of its precision matrix's quadratic form.

Here is a key observation: this function g is **strictly monotonically decreasing** in its argument. That is, $g(a) > g(b)$ whenever $a < b$. Therefore, small values of $\tilde{\mathbf{x}}^\top \boldsymbol{\Sigma}^{-1} \tilde{\mathbf{x}}$ (which generally correspond to points where $\tilde{\mathbf{x}}$ is closer to $\mathbf{0}$, i.e. $\mathbf{x} \approx \boldsymbol{\mu}$) have relatively high probability densities, and vice-versa. Furthermore, because g is *strictly* monotonic, it is injective, so the c-isocontours of $p(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ are the $g^{-1}(c)$-isocontours of the function $\mathbf{x} \mapsto \tilde{\mathbf{x}}^\top \boldsymbol{\Sigma}^{-1} \tilde{\mathbf{x}}$. That is, for any c,

$$\{\mathbf{x} \in \mathbb{R}^d : p(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = c\} = \{\mathbf{x} \in \mathbb{R}^d : \tilde{\mathbf{x}}^\top \boldsymbol{\Sigma}^{-1} \tilde{\mathbf{x}} = g^{-1}(c)\}$$

In words, these functions have the same isocontours but different isovalues.

Recall the executive summary of the geometry of positive definite quadratic forms: the isocontours of $f(\mathbf{x}) = \mathbf{x}^\top \mathbf{A} \mathbf{x}$ are ellipsoids such that the axes point in the directions of the eigenvectors of $\mathbf{A}$, and the lengths of these axes are proportional to the inverse square roots of the corresponding eigenvalues. Therefore in this case, the isocontours of the density are ellipsoids (centered at $\boldsymbol{\mu}$) with axis lengths proportional to the inverse square roots of the eigenvalues of $\boldsymbol{\Sigma}^{-1}$, or equivalently, the square roots of the eigenvalues of $\boldsymbol{\Sigma}$.

Acknowledgements

The author would like to thank Michael Franco for suggested clarifications.

References

[1] J. Nocedal and S. J. Wright, *Numerical Optimization*. New York: Springer Science+Business Media, 2006.

[2] J. S. Rosenthal, *A First Look at Rigorous Probability Theory (Second Edition)*. Singapore: World Scientific Publishing, 2006.

[3] J. Pitman, *Probability*. New York: Springer-Verlag, 1993.

[4] S. Axler, *Linear Algebra Done Right (Third Edition)*. Springer International Publishing, 2015.

[5] S. Boyd and L. Vandenberghe, *Convex Optimization*. New York: Cambridge University Press, 2009.

[6] J. A. Rice, *Mathematical Statistics and Data Analysis*. Belmont, California: Thomson Brooks/Cole, 2007.

[7] G. B. Folland, *Real Analysis: Modern Techniques and Their Applications (Second Edition)*. New York: John Wiley & Sons, 1999.