Mathematics for Machine Learning

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

June 29, 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. Indeed, it is the case that every subject presented here is covered rather minimally. We intend only 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.

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

1 About **1**

2 Notation **4**

3 Linear Algebra **5**

3.1 Vector spaces 5

3.1.1 Euclidean space 5

3.2 Metric spaces 6

3.3 Normed spaces 6

3.4 Inner product spaces 7

3.4.1 Pythagorean Theorem 8

3.4.2 Cauchy-Schwarz inequality 8

3.5 Transposition 8

3.6 Eigenthings 8

3.7 Trace 9

3.8 Determinant 9

3.9 Special kinds of matrices 9

3.9.1 Orthogonal matrices 10

3.9.2 Symmetric matrices 10

3.9.3 Positive (semi-)definite matrices 10

3.10 Singular value decomposition 10

4 Calculus and Optimization **12**

4.1 Extrema 12

4.2 Gradients 12

4.3 The Jacobian 12

4.4 The Hessian 13

4.5 Taylor's theorem 13

4.6 Matrix calculus 13

4.6.1 The chain rule 14

4.7 Convexity 14

4.7.1 Convex sets 14

4.7.2 Basics of convex functions 15

4.7.3 Consequences of convexity 15

4.7.4 Showing that a function is convex 17

4.7.5 Examples 19

5 Probability **20**

5.1 Basics 20

5.1.1 Conditional probability 21

5.1.2 Chain rule 21

5.1.3 Bayes' rule 21

5.2 Random variables 22

5.2.1 The cumulative distribution function 22

5.2.2 Discrete random variables 23

5.2.3 Continuous random variables 23

5.2.4 Other kinds of random variables 23

5.3 Joint distributions 24

5.3.1 Independence of random variables 24

5.3.2 Marginal distributions 24

5.4 Great Expectations 24

5.4.1 Properties of expected value 24

5.5 Variance 25

5.5.1 Properties of variance 25

5.5.2 Standard deviation 25

5.6 Covariance 25

5.6.1 Correlation 26

5.7 Random vectors 26

5.8 Estimation of Parameters 27

5.8.1 Maximum likelihood estimation 27

5.8.2 Maximum a posteriori estimation 27

5.9 The Gaussian distribution 28

References **29**

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
$\nabla f(\mathbf{x})$	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
$\mathbb{P}(A \mid B)$	probability of event A, given B
$p(X)$	distribution of random variable X
$p(x)$	probability density/mass function evaluated at x
A^c	complement of event A
$\mathbb{E}[X]$	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 are in bold (e.g. $\mathbf{x}$). 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 generalize it.

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 single real[1] numbers (called **scalars**). V must satisfy

1. There exists an additive identity (written $\mathbf{0}$) in V such that $\mathbf{x} + \mathbf{0} = \mathbf{x}$ for all $\mathbf{x} \in V$
2. For each $\mathbf{x} \in V$, there exists an additive inverse (written $-\mathbf{x}$) such that $\mathbf{x} + (-\mathbf{x}) = \mathbf{0}$
3. There exists a multiplicative identity (written 1) in $\mathbb{R}$ such that $1\mathbf{x} = \mathbf{x}$ for all $\mathbf{x} \in V$
4. Commutativity: $\mathbf{x} + \mathbf{y} = \mathbf{y} + \mathbf{x}$ for all $\mathbf{x}, \mathbf{y} \in V$
5. 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 $\alpha, \beta \in \mathbb{R}$
6. 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 $\alpha, \beta \in \mathbb{R}$

We won't be quoting these axioms directly, but it is worth knowing them, or at least having an intuitive feeling for what they mean. Most of them should already be familiar to the reader.

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:

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

For our purposes, it will often 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 $\mathbb{R}^n$:

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

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

[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.

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.

3.2 Metric spaces

Metrics generalize the notion of distance from Euclidean space.

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

1. $d(x, y) \geq 0$, with equality if and only if $x = y$
2. $d(x, y) = d(y, x)$
3. $d(x, z) \leq d(x, y) + d(y, z)$ (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

1. $\|\mathbf{x}\| \geq 0$, with equality if and only if $\mathbf{x} = \mathbf{0}$
2. $\|\alpha\mathbf{x}\| = |\alpha|\|\mathbf{x}\|$
3. $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$ (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 $\mathbb{R}^n$:

$$\begin{aligned}
\|\mathbf{x}\|_1 &= \sum_{i=1}^n |x_i| \\
\|\mathbf{x}\|_2 &= \sqrt{\sum_{i=1}^n x_i^2} \\
\|\mathbf{x}\|_p &= \left(\sum_{i=1}^n |x_i|^p\right)^{\frac{1}{p}} \qquad\qquad (p \geq 1) \\
\|\mathbf{x}\|_\infty &= \max_{1\leq i\leq n} |x_i|
\end{aligned}$$

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.

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

1. $\langle\mathbf{x},\mathbf{x}\rangle \geq 0$, with equality if and only if $\mathbf{x} = \mathbf{0}$
2. $\langle\alpha\mathbf{x} + \beta\mathbf{y},\mathbf{z}\rangle = \alpha\langle\mathbf{x},\mathbf{z}\rangle + \beta\langle\mathbf{y},\mathbf{z}\rangle$
3. $\langle\mathbf{x},\mathbf{y}\rangle = \langle\mathbf{y},\mathbf{x}\rangle$

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**.

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

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

[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 matrix notation on the righthand side (see the Transposition section if it's unfamiliar to you) 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:

1. $(\mathbf{A}^\top)^\top = \mathbf{A}$
2. $(\mathbf{A}+\mathbf{B})^\top = \mathbf{A}^\top + \mathbf{B}^\top$
3. $(\alpha\mathbf{A})^\top = \alpha\mathbf{A}^\top$
4. $(\mathbf{AB})^\top = \mathbf{B}^\top\mathbf{A}^\top$

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

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

The zero vector is excluded from this definition because $\mathbf{A0} = \mathbf{0} = \lambda\mathbf{0}$ for every λ.

The set of all eigenvectors of a matrix, each paired with its corresponding eigenvalue, is called the **eigensystem** of that matrix.

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:

1. $\operatorname{tr}(\mathbf{A} + \mathbf{B}) = \operatorname{tr}(\mathbf{A}) + \operatorname{tr}(\mathbf{B})$
2. $\operatorname{tr}(\alpha\mathbf{A}) = \alpha \operatorname{tr}(\mathbf{A})$
3. $\operatorname{tr}(\mathbf{A}^\top) = \operatorname{tr}(\mathbf{A})$
4. $\operatorname{tr}(\mathbf{ABCD}) = \operatorname{tr}(\mathbf{BCDA}) = \operatorname{tr}(\mathbf{CDAB}) = \operatorname{tr}(\mathbf{BADC})$

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):

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

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:

1. $\det(\mathbf{I}) = 1$
2. $\det(\mathbf{A}^\top) = \det(\mathbf{A})$
3. $\det(\mathbf{AB}) = \det(\mathbf{A})\det(\mathbf{B})$
4. $\det(\mathbf{A}^{-1}) = \det(\mathbf{A})^{-1}$
5. $\det(\alpha\mathbf{A}) = \alpha^n \det(\mathbf{A})$

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

$$\det(\mathbf{A}) = \prod_{i} \lambda_i$$

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

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

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

$$(\mathbf{Q}\mathbf{x})^\top(\mathbf{Q}\mathbf{y}) = \mathbf{x}^\top \mathbf{Q}^\top \mathbf{Q}\mathbf{y} = \mathbf{x}^\top \mathbf{I}\mathbf{y} = \mathbf{x}^\top \mathbf{y}$$

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

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

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 ($\mathbf{A} = \mathbf{A}^\top$). A very important property of symmetric matrices is that they can be decomposed in the following manner:

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

Here $\mathbf{Q}$ is an orthogonal matrix, and $\mathbf{\Lambda} = \operatorname{diag}(\lambda_1, \dots, \lambda_n)$, where $\lambda_1, \dots, \lambda_n$ are the eigenvalues of 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.

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.

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. 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 singular values of $\mathbf{A}$ are the square roots of the eigenvalues of $\mathbf{A}^\top\mathbf{A}$ (or equivalently, of $\mathbf{A}\mathbf{A}^\top$).

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 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. A point $\mathbf{x}$ is said to be a **local minimum** (resp. **local maximum**) of $f : \mathbb{R}^d \to \mathbb{R}$ if $f(\mathbf{x}) \leq f(\mathbf{y})$ (resp. $f(\mathbf{x}) \geq f(\mathbf{y})$) for all $\mathbf{y}$ in some neighborhood about $\mathbf{x}$. Furthermore, if $f(\mathbf{x}) \leq f(\mathbf{y})$ for all $\mathbf{y}$ in the entire domain of f, then $\mathbf{x}$ is a **global minimum** of f (similarly for global maximum).

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.

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 $f : \mathbb{R}^d \to \mathbb{R}$ at $\mathbf{x} \in \mathbb{R}^d$, denoted $\nabla f(\mathbf{x})$, is given by

$$\nabla f(\mathbf{x}) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(\mathbf{x}) \\ \vdots \\ \frac{\partial f}{\partial x_n}(\mathbf{x}) \end{bmatrix}$$

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

$$\mathbf{J}_f(\mathbf{x}) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1}(\mathbf{x}) & \dots & \frac{\partial f_1}{\partial x_n}(\mathbf{x}) \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1}(\mathbf{x}) & \dots & \frac{\partial f_m}{\partial x_n}(\mathbf{x}) \end{bmatrix}$$

i.e.

$$[\mathbf{J}_f(\mathbf{x})]_{ij} = \frac{\partial f_i}{\partial x_j}(\mathbf{x})$$

4.4 The Hessian

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

$$\nabla^2 f(\mathbf{x}) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2}(\mathbf{x}) & \dots & \frac{\partial^2 f}{\partial x_1 \partial x_n}(\mathbf{x}) \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1}(\mathbf{x}) & \dots & \frac{\partial^2 f}{\partial x_n^2}(\mathbf{x}) \end{bmatrix}$$

i.e.

$$[\nabla^2 f(\mathbf{x})]_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j}(\mathbf{x})$$

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 Taylor's theorem

Taylor's theorem has natural generalizations to functions of more than one variable. One version states

$$f(\mathbf{x} + \mathbf{y}) = f(\mathbf{x}) + \nabla f(\mathbf{x} + \alpha\mathbf{y})^\top \mathbf{y}$$

for some $\alpha \in (0, 1)$. Furthermore, if f is twice-differentiable, we have

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

for some $\alpha \in (0, 1)$.

This theorem is used in proofs about necessary and sufficient conditions for local optima. We don't reproduce the proofs here, but the interested reader can consult [**?**].

Here is an example of such a useful fact: if f is differentiable, then for any extremum $\mathbf{x}$, $\nabla f(\mathbf{x}) = \mathbf{0}$. Note that the converse does not hold in general, that is, $\nabla f(\mathbf{x}) = \mathbf{0}$ does not necessarily imply that $\mathbf{x}$ is an extremum. (It could be a **saddle point** of f.)

4.6 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 $2\mathbf{A}\mathbf{x}$.

4.6.1 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. We can generalize this to the multivariate case.

Proposition 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*

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

Proof. Observe that each x_i potentially affects every output element of g, so we must sum the effects across all of these elements. The derivative of $f \circ g$ with respect to x_i is given by

$$\frac{\partial(f \circ g)}{\partial x_i}(\mathbf{x}) = \sum_{j=1}^{m} \frac{\partial(f \circ g)}{\partial g_j} \frac{\partial g_j}{\partial x_i}(\mathbf{x}) = \nabla f(g(\mathbf{x}))^\top \frac{\partial g}{\partial x_i}(\mathbf{x})$$

where

$$\frac{\partial g}{\partial x_i}(\mathbf{x}) = \begin{bmatrix} \frac{\partial g_1}{\partial x_i}(\mathbf{x}) \\ \vdots \\ \frac{\partial g_m}{\partial x_i}(\mathbf{x}) \end{bmatrix}$$

Stacking these partial derivatives and recalling the definitions of the Jacobian matrix and matrix multiplication, we obtain the desired result. □

4.7 Convexity

Convexity is a term that is used to describe both sets and functions. There is a whole branch of mathematics called **convex analysis** devoted to studying convexity of sets and functions, but we will just focus on the bits relevant to optimization, particularly convexity as it pertains to 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.

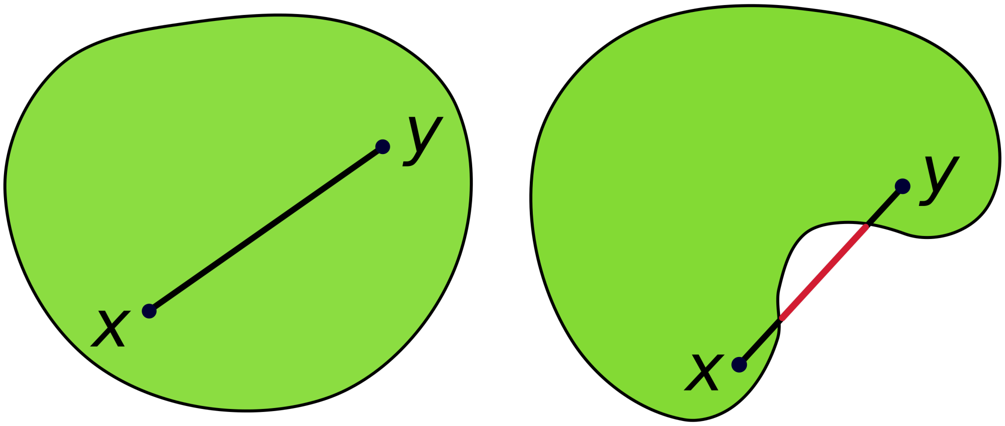
4.7.1 Convex sets

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

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

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

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



(a) A convex set

(b) A non-convex set

Figure 1: What convex sets look like

4.7.2 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 $t \in [0, 1]$.

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.)

4.7.3 Consequences of convexity

Why do we care about convexity?

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 2. *If f is convex, then any local minimizer of f is a global minimizer of f.*

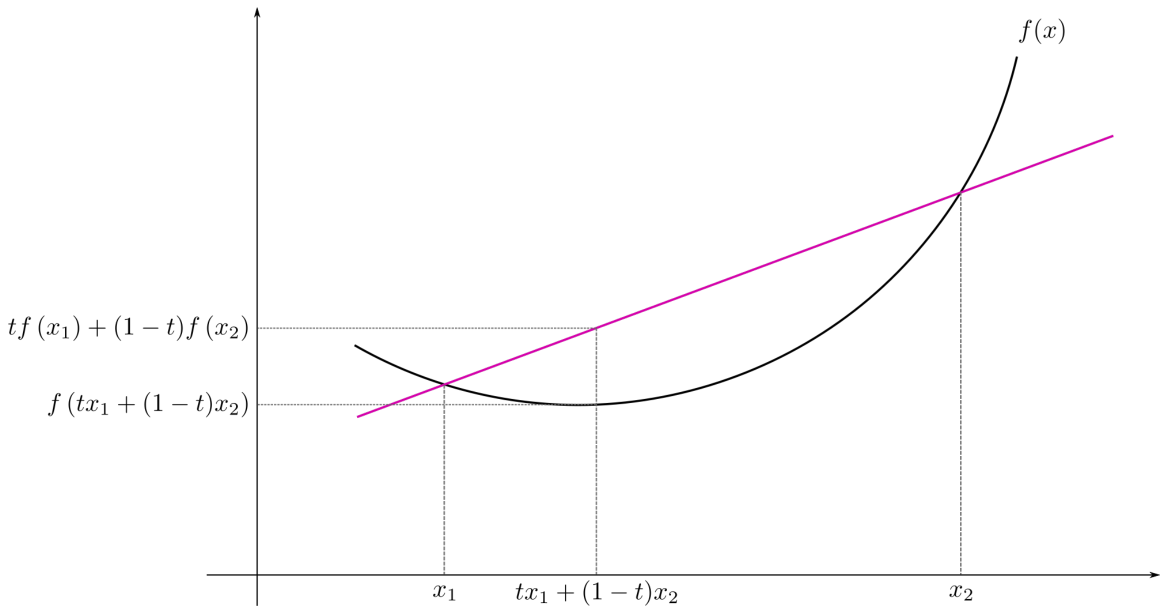


Figure 2: What convex functions look like

Proof. Suppose f is convex, and let $\mathbf{x}^*$ be a local minimizer of f. Then for some neighborhood $\mathcal{N}$ about $\mathbf{x}^*$,

$$f(\mathbf{x}) \geq f(\mathbf{x}^*) \qquad \forall \mathbf{x} \in \mathcal{N}$$

Suppose towards a contradiction that there exists $\tilde{\mathbf{x}} \in \operatorname{dom} f$ such that $f(\tilde{\mathbf{x}}) < f(\mathbf{x}^*)$. Consider the line segment $\mathbf{x}(t) = t\mathbf{x}^* + (1-t)\tilde{\mathbf{x}},\ t \in [0,1]$. 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 $\mathbf{x} \in \operatorname{dom} f$, so $\mathbf{x}^*$ is a global minimizer of f. □

Proposition 3. *If f is strictly convex, then there exists at most one local minimizer of f. Consequently, if it exists it is the unique global minimizer of f.*

Proof. The second sentence follows immediately from the first, so all we must show is that if a local minimizer exists it is unique.

Suppose f has a local minimum, and let $\mathbf{x}^*$ be the minimizer. Suppose towards a contradiction that there exists a local minimizer $\tilde{\mathbf{x}} \in \operatorname{dom} f$ 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 minimizers of f by the previous result. Hence $f(\mathbf{x}^*) = f(\tilde{\mathbf{x}})$. Consider the line segment $\mathbf{x}(t) = t\mathbf{x}^* + (1-t)\tilde{\mathbf{x}},\ t \in [0,1]$. 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 minimizer.

It follows that if $\tilde{\mathbf{x}}$ is a local minimizer of f, then $\tilde{\mathbf{x}} = \mathbf{x}^*$, so $\mathbf{x}^*$ is the only local minimizer. □

4.7.4 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.

Proposition 4. *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 5. *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 6. *Suppose f is twice differentiable and $\operatorname{dom} f$ is convex and open. 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 7. *If f is convex and $\alpha \geq 0$, then αf is convex.*

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

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

so αf is convex. □

Proposition 8. *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 $\mathbf{x}, \mathbf{y} \in \operatorname{dom}(f+g) = \operatorname{dom} f \cap \operatorname{dom} g$,

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

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 9. *If $f_1, \dots, f_n$ are convex and $\alpha_1, \dots, \alpha_n \geq 0$, then*

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

is convex.

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

Proposition 10. *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 11. *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. □

4.7.5 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 $\mathcal{F}$.[4]

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**.[5]

If $\mathbb{P}(A) = 1$, we say that A occurs **almost surely** (often abbreviated a.s.).[6] Conversely if $\mathbb{P}(A) = 0$, we say that A occurs **almost never**.

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

Proposition 12. *If A is an event, then $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$.*

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

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

which proves the result. □

Proposition 13. *Let A be an event. Then*

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

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

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

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

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

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

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

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

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

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

Proposition 14. *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 15. *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 $\mathbb{P}(B) > 0$.[7]

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**:

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

[7] 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[8] $X : \Omega \to \mathbb{R}$.[9]

We denote the range of X by $X(\Omega) = \{X(\omega) : \omega \in \Omega\}$. To give a concrete example (taken from [**?**]), 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:

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

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

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

[8] The function must be measurable.

[9] 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**[10], 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.

[10] 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 $X_1, \dots, X_n$, the joint distribution is written $p(X_1, \dots, X_n)$ 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.

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

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:

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

where $X_1, \dots, X_n$ 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 $\operatorname{Var}(X)$ of a random variable X by

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

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

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

5.5.1 Properties of variance

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

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

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 $X_1, \dots, X_n$ are uncorrelated[11], then

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

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 $\sqrt{\operatorname{Var}(X)}$. The standard deviation of X has the same units as X.

5.6 Covariance

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

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

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

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

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

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**:

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

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**:

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

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

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

The inverse of the covariance matrix, Σ^{-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

$$\hat{\theta}_{\text{MLE}} = \operatorname*{arg\,max}_{\theta} \ell(\theta)$$

where ℓ is the **likelihood function**

$$\ell(\theta) = p(x_1, \dots, x_n; \theta)$$

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**

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

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 $\log \ell$ will also maximize ℓ.

For some distributions, it is possible to analytically solve for the maximum likelihood estimator. If $\log \ell$ 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

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

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 the mean μ and variance σ^2. In the single-variable case, the Gaussian distribution has the following density:

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

We write $X \sim \mathcal{N}(\mu, \sigma^2)$ to denote that X is normally distributed with mean μ and variance σ^2. As one might expect, this implies $\mathbb{E}[X] = \mu$ and $\text{Var}(X) = \sigma^2$.

In the multivariate case, the mean becomes a vector $\boldsymbol{\mu}$ and the variance is generalized by the covariance matrix $\boldsymbol{\Sigma}$. The density is

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

where d is the dimension.

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.