Mathematics for Machine Learning

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

January 18, 2017

1 About

Machine learning uses tools from a variety of mathematical fields. This document is intended to summarize the mathematical background needed for an introductory class in machine learning, which at UC Berkeley is known as CS 189. We will cover topics in linear algebra, optimization, and probability. 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.

You are free to distribute this document as you wish. 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
$\nabla f(\mathbf{x})$	gradient of the function f evaluated at $\mathbf{x}$
$\nabla^2 f(\mathbf{x})$	Hessian of the function f evaluated at $\mathbf{x}$
$A^\top$	transpose of the matrix 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).
- We do not provide proofs of most of the theorems/statements given in this document. The proofs are not the point – our primary goal is to review important definitions and concepts.

3 Linear Algebra

We begin by discussing 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 using these axioms directly, but it is worth knowing them, or at least having an intuitive feeling for what they mean.

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 $x, y, z \in S$.

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

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

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

3.6 Eigenthings

For a square matrix $A \in \mathbb{R}^{n \times n}$, there may be vectors which, when 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 A corresponding to **eigenvalue** λ if

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

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

3.7 Trace

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

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

The trace has several nice algebraic properties, most of which can be easily verified from the definition:

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

This last property is known as **invariance under cyclic permutations**. Note that the matrices cannot be reordered arbitrarily, for example $\operatorname{tr}(ABCD) \neq \operatorname{tr}(BACD)$ in general.

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

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

3.8 Determinant

The **determinant** of a 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. $\det(I) = 1$
2. $\det\bigl(A^\top\bigr) = \det(A)$
3. $\det(AB) = \det(A)\det(B)$
4. $\det\bigl(A^{-1}\bigr) = \det(A)^{-1}$
5. $\det(\alpha A) = \alpha^n \det(A)$

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

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

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

$$Q^\top Q = QQ^\top = I$$

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

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

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

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

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

3.9.2 Symmetric Matrices

A matrix $A \in \mathbb{R}^{n\times n}$ is said to be **symmetric** if it is equal to its own transpose ($A = A^\top$). A very important property of symmetric matrices is that they can be decomposed in the following manner:

$$A = Q\Lambda Q^\top$$

Here Q is an orthogonal matrix, and $\Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_n)$, where $\lambda_1, \dots, \lambda_n$ are the eigenvalues of A. This is referred to as the **eigendecomposition** or **spectral decomposition** of A.

3.9.3 Positive (Semi-)Definite Matrices

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

A symmetric matrix A is **positive semi-definite** if for all $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{x}^\top A\mathbf{x} \geq 0$. Sometimes people write $A \succeq 0$ to indicate that 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 $A^\top A$ is positive semi-definite for any $A \in \mathbb{R}^{m\times n}$, since

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

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* $A \in \mathbb{R}^{m\times n}$ has an SVD (even non-square matrices)! The decomposition goes as follows:

$$A = U\Sigma V^\top$$

where $U \in \mathbb{R}^{m \times m}$ and $V \in \mathbb{R}^{n \times n}$ are orthogonal matrices and $\Sigma \in \mathbb{R}^{m \times n}$ is a diagonal matrix[4] with the **singular values** of 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 A.

The singular values of A are the square roots of the eigenvalues of $A^\top A$ (or equivalently, of $AA^\top$).

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

There is another useful way to write the SVD:

$$A = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$$

[4] Some would protest that a diagonal matrix must be square. We simply mean that all the off-diagonal entries are zero.

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. We will not give specific examples of cost functions here (go to class for these!) but we will assume that our cost function has the form $f : \mathbb{R}^n \to \mathbb{R}$ and is sufficiently differentiable.

4.1 Extrema

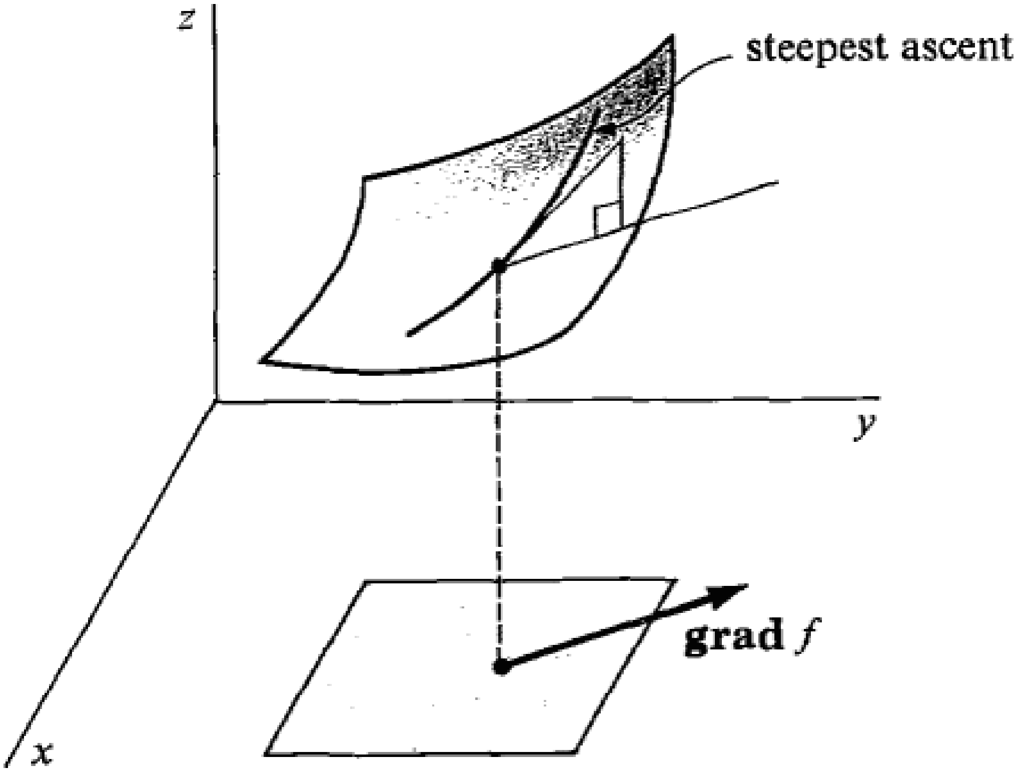
Optimization is about finding **extrema**, which depending on the application could be minima or maxima. A point $\mathbf{x}$ is said to be a **local minimum** (resp. **local maximum**) of f if $f(\mathbf{x}) \leq f(\mathbf{y})$ (resp. $f(\mathbf{x}) \geq f(\mathbf{y})$) for all $\mathbf{y}$ in some 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).

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, denoted ∇f, is given by

$$\nabla f = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix}$$

Gradients have the following very important property: $\nabla f(\mathbf{x})$ points in the direction of **steepest ascent** from $\mathbf{x}$:



We will use this fact frequently when iteratively minimizing a function via **gradient descent**.

4.3 Hessians

The **Hessian** is a matrix of second-order partial derivatives:

$$\nabla^2 f = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix}$$

i.e.

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

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.

4.4 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 [4].

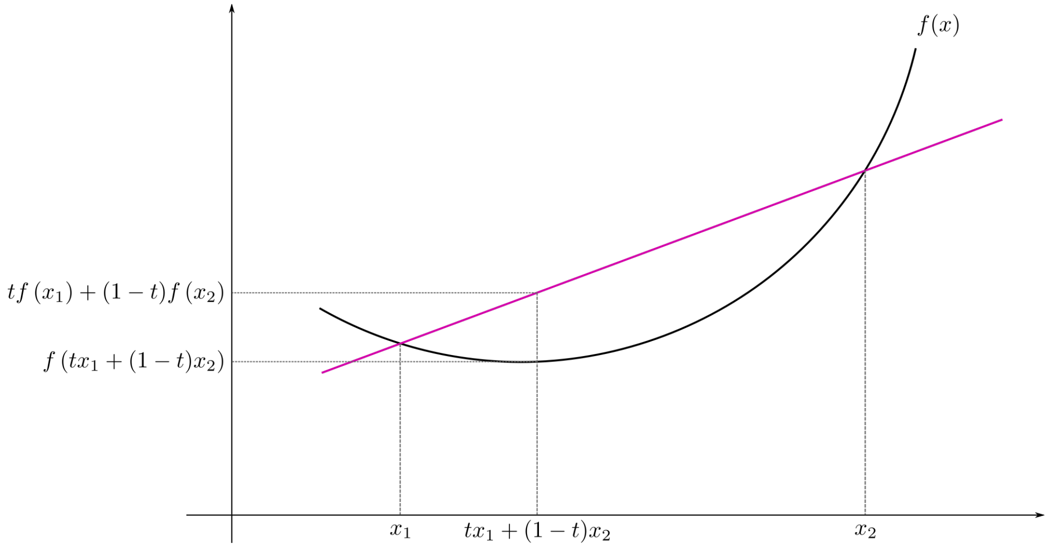
Here is an example of such a useful fact: if f is differentiable, then for any extremum $\mathbf{x}$, $\nabla f(\mathbf{x}) = \mathbf{0}$. Note that the converse does not hold in general, that is, $\nabla f(\mathbf{x}) = \mathbf{0}$ does not necessarily imply that $\mathbf{x}$ is an extremum. (It could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

4.5 Convexity

A function f is said to be **convex** if

$$f(\alpha\mathbf{x} + (1 - \alpha)\mathbf{y}) \le \alpha f(\mathbf{x}) + (1 - \alpha) f(\mathbf{y})$$

for all $\mathbf{x}, \mathbf{y} \in \operatorname{dom} f$ and any $\alpha \in [0, 1]$. Geometrically, this means that the line segment between two points on the graph of f lies on or above the graph itself:



We say that a function is **concave** if its negation is convex.

There are a number of ways to show that a function is convex:

(i) If f is differentiable, then it is convex on an interval if and only if its derivative is non-decreasing on that interval.

(ii) If f is continuously differentiable, then it is convex on an interval if and only if

$$f(x) \ge f(y) + f'(y)(x - y)$$

for all x, y in the interval.

(iii) If f is twice differentiable, then it is convex on some convex set if and only if the Hessian $\nabla^2 f$ is positive semi-definite on the interior of that set.

(iv) If f is convex and $c \geq 0$, then cf is convex.

(v) If f and g are convex, then $f + g$ is convex.

(vi) If f is convex, then $g(\mathbf{x}) \equiv f(A\mathbf{x} + \mathbf{b})$ is convex for any A and $\mathbf{b}$.

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

Convexity sounds at first like a very mysterious property, but it has some wonderful implications. A particularly nice consequence is that any local minimum of a convex function f is a global minimum of f! Generally speaking, the minimization of convex functions is much better understood than minimization of general nonlinear functions.

Unfortunately, many loss functions that we would like to minimize are non-convex (e.g. for training neural networks). This means whatever local minima our optimization algorithm finds may not be globally optimal.

5 Probability

Probability theory provides powerful tools for modeling and dealing with uncertainty. In machine learning, we will use it extensively, 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. We call this set the **sample space** and denote it Ω. Any subset $A \subseteq \Omega$ is called an **event**. A **probability distribution** over Ω specifies how likely each event is to occur. We write $\mathbb{P}(A)$ for the probability of event A.

Here are the basic axioms of probability:

1. For any event $A \subseteq \Omega$, $\mathbb{P}(A) \geq 0$
2. $\mathbb{P}(\Omega) = 1$
3. If $A_1, \dots, A_n$ are **mutually exclusive**, i.e. $A_i \cap A_j = \varnothing$ for $i \neq j$, then

$$\mathbb{P}(A_1 \cup \dots \cup A_n) = \mathbb{P}(A_1) + \dots + \mathbb{P}(A_n)$$

From these axioms, a number of useful rules can be derived (see [1]):

1. $\mathbb{P}(\varnothing) = 0$
2. If $A \subseteq B$, then $\mathbb{P}(A) \leq \mathbb{P}(B)$
3. $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$
4. $\mathbb{P}(A) + \mathbb{P}(A^c) = 1$

5.1.1 Conditional Probability

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

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

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

5.1.2 Chain Rule

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

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

5.1.3 Bayes' Rule

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

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

It is sometimes beneficial to omit the normalizing constant and write

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

Under this formulation, $\mathbb{P}(A)$ is often referred to as the **prior** and $\mathbb{P}(B \mid A)$ 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** (r.v.) is some uncertain quantity with an associated probability distribution over the values it can assume. We write $X \sim p(\cdot)$ to indicate that the random variable X is distributed according to some probability mass/density function p (more on these functions below).

Formally, a random variable is a function from the sample space Ω to some other set of values, which we denote $X(\Omega) = \{X(\omega) \mid \omega \in \Omega\}$. To give a concrete example (taken from [1]), 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. $X = X(\omega)$. Moreover, we see the fundamental way that a random variable relates back to its underlying sample space:

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

Typically we work only with random variables and don't concern ourselves with the original sample space Ω, but it's worth knowing that it really is there behind the scenes.

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 Discrete Random Variables

Discrete random variables are usually specified by a nonnegative **probability mass function** (p.m.f.) p 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.:

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

5.2.2 Continuous Random Variables

Continuous random variables are usually specified by a nonnegative **probability density function** (p.d.f.) p which satisfies

$$\int_{X(\Omega)} p(x)\,\mathrm{d}x = 1$$

For a continuous X, the probability of a particular value is zero ($\forall x \in X(\Omega)$, $\mathbb{P}(X = x) = 0$), so to get a positive probability we must integrate the p.d.f. over some range of values:

$$\mathbb{P}(a \le X \le b) = \int_a^b p(x)\,\mathrm{d}x$$

5.2.3 The Cumulative Distribution Function

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

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

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

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

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.

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

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

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

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 $X_1, \ldots, X_n$ are uncorrelated[5], 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

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

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

5.6.1 Correlation

Normalizing the covariance gives the **correlation**:

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

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 $\operatorname{Cov}(X, Y) = 0$ because $\operatorname{Cov}(X, Y) = 0$ 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**:

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

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

The metrics 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**:

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

We see $\Sigma_{ij} = \operatorname{Cov}(X_i, X_j)$. 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 of the data we've seen, conditioned on 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:

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

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

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

Again, if we assume $x_1, \dots, x_n$ are i.i.d., then we can express this in the equivalent, and possibly friendlier, form

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

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.

References

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

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

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

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

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