A Mathematical Concepts

This appendix provides a brief overview of some of the mathematical concepts used in this book.

A.1 Measure Spaces

Before introducing the definition of a measure space, we will first introduce the notion of a sigma-algebra over a set $\Omega$. A sigma-algebra is a collection $\Sigma$ of subsets of $\Omega$ such that:

1. $\Omega \in \Sigma$.
2. If $E \in \Sigma$, then $\Omega \setminus E \in \Sigma$ (closed under complementation).
3. If $E_1, E_2, E_3, \ldots \in \Sigma$, then $E_1 \cup E_2 \cup E_3 \ldots \in \Sigma$ (closed under countable unions).

An element $E \in \Sigma$ is called a measurable set.

A measure space is defined by a set $\Omega$, a sigma-algebra $\Sigma$, and a measure $\mu : \Omega \to \mathbb{R} \cup \{\infty\}$. For $\mu$ to be a measure, the following properties must hold:

1. If $E \in \Sigma$, then $\mu(E) \geq 0$ (non-negativity).
2. $\mu(\emptyset) = 0$.
3. If $E_1, E_2, E_3, \ldots \in \Sigma$ are pairwise disjoint, then $\mu(E_1 \cup E_2 \cup E_3 \ldots) = \mu(E_1) + \mu(E_2) + \mu(E_3) + \cdots$ (countable additivity).
A.2 Probability Spaces

A probability space is a measure space \((\Omega, \Sigma, \mu)\) with the requirement that \(\mu(\Omega) = 1\). In the context of probability spaces, \(\Omega\) is called the sample space, \(\Sigma\) is called the event space, and \(\mu\) (or more commonly \(P\)) is the probability measure. The probability axioms\(^1\) refer to the non-negativity and countable additivity properties of measure spaces together with the requirement that \(\mu(\Omega) = 1\).

A.3 Metric Spaces

A set with a metric is called a metric space. A metric \(d\), sometimes called a distance metric, is a function that maps pairs of elements in \(X\) to non-negative real numbers such that for all \(x, y, z \in X\):

1. \(d(x, y) = 0\) if and only if \(x = y\) (identity of indiscernibles).
2. \(d(x, y) = d(y, x)\) (symmetry).
3. \(d(x, y) \leq d(x, z) + d(z, y)\) (triangle inequality).

A.4 Normed Vector Spaces

A normed vector space consists of a vector space \(X\) and a norm \(\|\cdot\|\) that maps elements of \(X\) to non-negative real numbers such that for all scalars \(\alpha\) and vectors \(x, y \in X\):

1. \(\|x\| = 0\) if and only if \(x = 0\).
2. \(\|\alpha x\| = |\alpha| \|x\|\) (absolutely homogeneous).
3. \(\|x + y\| \leq \|x\| + \|y\|\) (triangle inequality).

The \(L_p\) norms are a commonly used set of norms parameterized by a scalar \(p \geq 1\). The \(L_p\) norm of vector \(x\) is

\[
\|x\|_p = \lim_{\rho \to p} \left( |x_1|^\rho + |x_2|^\rho + \cdots + |x_n|^\rho \right)^{\frac{1}{\rho}}
\]  

(A.1)

where the limit is necessary for defining the infinity norm, \(L_\infty\). Several \(L_p\) norms are shown in table A.1.

Norms can be used to induce distance metrics in vector spaces by defining the metric \(d(x, y) = \|x - y\|\). We can then, for example, use an \(L_p\) norm to define distances.

\(^1\) These axioms are sometimes called the Kolmorogov axioms. A. Kolmogorov, Foundations of the Theory of Probability, 2nd ed. Chelsea, 1956.
\(L_1: \|x\|_1 = |x_1| + |x_2| + \cdots + |x_n|\)
This metric is often referred to as the *taxicab norm*.

\(L_2: \|x\|_2 = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}\)
This metric is often referred to as the *Euclidean norm*.

\(L_\infty: \|x\|_\infty = \max(|x_1|, |x_2|, \cdots, |x_n|)\)
This metric is often referred to as the *max norm, Chebyshev norm, or chessboard norm*. The latter name comes from the minimum number of moves a chess king needs to move between two chess squares.
A.5 Positive Definiteness

A symmetric matrix $A$ is *positive definite* if $x^\top Ax$ is positive for all points other than the origin. In other words, $x^\top Ax > 0$ for all $x \neq 0$. A symmetric matrix $A$ is *positive semidefinite* if $x^\top Ax$ is always non-negative. In other words, $x^\top Ax \geq 0$ for all $x$.

A.6 Convexity

A *convex combination* of two vectors $x$ and $y$ is the result of

$$ax + (1 - \alpha)y$$

(A.2)

for some $\alpha \in [0, 1]$. Convex combinations can be made from $m$ vectors,

$$w_1v^{(1)} + w_2v^{(2)} + \cdots + w_mv^{(m)}$$

(A.3)

with nonnegative weights $w$ that sum to one.

A *convex set* is a set for which a line drawn between any two points in the set is entirely within the set. Mathematically, a set $S$ is convex if we have

$$ax + (1 - \alpha)y \in S.$$  

(A.4)

for all $x, y$ in $S$ and for all $\alpha$ in $[0, 1]$. A convex and a nonconvex set are shown in figure A.1.

A convex set

Not a convex set

Figure A.1. Convex and non-convex sets.

A *convex function* is a *bowl-shaped* function whose domain is a convex set. By bowl-shaped, we mean it is a function such that any line drawn between two points in its domain does not lie below the function. A function $f$ is convex over a convex set $S$ if, for all $x, y$ in $S$ and for all $\alpha$ in $[0, 1]$,

$$f(ax + (1 - \alpha)y) \leq \alpha f(x) + (1 - \alpha)f(y)$$

(A.5)
Convex and concave regions of a function are shown in figure A.2.

A function $f$ is strictly convex over a convex set $S$ if, for all $x, y$ in $S$ and $\alpha$ in $(0, 1)$,

$$f(\alpha x + (1 - \alpha) y) < \alpha f(x) + (1 - \alpha) f(y)$$  \hspace{1cm} (A.6)

Strictly convex functions have at most one minimum, whereas a convex function can have flat regions.\(^2\) Examples of strict and nonstrict convexity are shown in figure A.3.

A function $f$ is concave if $-f$ is convex. Furthermore, $f$ is strictly concave if $-f$ is strictly convex.

### A.7 Information Content

If we have a discrete distribution that assigns probability $P(x)$ to value $x$, the information content\(^3\) of observing $x$ is given by

$$I(x) = -\log P(x)$$  \hspace{1cm} (A.7)


The unit of information content depends on the base of the logarithm. We generally assume natural logarithms (with base $e$), making the unit *nat*, short for *natural*. In information theoretic contexts, the base is often 2, making the unit *bit*. We can think of this quantity as the number of bits required to transmit the value $x$ according to an optimal message encoding when the distribution over messages follows the specified distribution.

### A.8 Entropy

*Entropy* is an information theoretic measure of uncertainty. The entropy associated with a discrete random variable $X$ is the expected information content:

$$H(X) = \mathbb{E}_X[I(x)] = \sum_x P(x)I(x) = -\sum_x P(x) \log P(x) \quad (A.8)$$

where $P(x)$ is the mass assigned to $x$.

For a continuous distribution where $p(x)$ is the density assigned to $x$, the *differential entropy* or *continuous entropy* is defined to be

$$h(X) = \int p(x)I(x) \, dx = -\int p(x) \log p(x) \, dx \quad (A.9)$$

### A.9 Cross Entropy

The *cross entropy* of one distribution relative to another can be defined in terms of expected information content. If we have one discrete distribution with mass function $P(x)$ and another with mass function $Q(x)$, then the cross entropy of $P$ relative to $Q$ is given by

$$H(P, Q) = -\mathbb{E}_{X \sim P}[\log Q(x)] = -\sum_x P(x) \log Q(x) \quad (A.10)$$

For continuous distributions with density functions $p(x)$ and $q(x)$, we have

$$H(p, q) = -\int p(x) \log q(x) \, dx \quad (A.11)$$
A.10 Relative Entropy

The relative entropy or Kullback–Leibler (KL) divergence is a measure of how one probability distribution is different from a reference distribution.\(^4\) If \(P(x)\) and \(Q(x)\) are mass functions, then the KL divergence from \(Q\) to \(P\) is the expectation of the logarithmic differences, with the expectation using \(P\):

\[
D_{\text{KL}}(P \mid\mid Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)} = -\sum_x P(x) \log \frac{Q(x)}{P(x)} \tag{A.12}
\]

This quantity is only defined if the support of \(P\) is a subset of that of \(Q\). The summation is over the support of \(P\) to avoid division by zero.

For continuous distributions with density functions \(p(x)\) and \(q(x)\), we have

\[
D_{\text{KL}}(p \mid\mid q) = \int p(x) \log \frac{p(x)}{q(x)} \, dx = -\int p(x) \log \frac{q(x)}{p(x)} \, dx \tag{A.13}
\]

Again, this quantity is only defined if the support of \(p\) is a subset of that of \(q\). The integral is over the support of \(p\) to avoid division by zero.

A.11 Gradient Ascent

Gradient ascent is a general approach for attempting to maximize a function \(f(x)\) when \(f\) is a differentiable function. We begin at a point \(x\) and iteratively apply the following update rule:

\[
x \leftarrow x + \alpha \nabla f(x) \tag{A.14}
\]

where \(\alpha > 0\) is called a step factor. The idea of this optimization approach is that we take steps in the direction of the gradient until reaching a local maximum. There is no guarantee that we will find a global maximum using this method. Small values for \(\alpha\) will generally require more iterations to come close to a local maximum. Large values for \(\alpha\) will often result in bouncing around the local optimum without quite reaching it. If \(\alpha\) is constant over iterations, it is sometimes called a learning rate. Many applications involve a decaying step factor where, in addition to updating \(x\) at each iteration, we also update \(\alpha\) according to

\[
\alpha \leftarrow \gamma \alpha \tag{A.15}
\]

where \(0 < \gamma < 1\) is the decay factor.

A.12 Taylor Expansion

The Taylor expansion, also called the Taylor series, of a function is important to many approximations used in this book. From the first fundamental theorem of calculus\(^5\), we know that

\[
f(x + h) = f(x) + \int_0^h f'(x + a) \, da \quad (A.16)
\]

Nesting this definition produces the Taylor expansion of \(f\) about \(x\):

\[
f(x + h) = f(x) + \int_0^h \left( f'(x) + \int_0^a f''(x + b) \, db \right) \, da \quad (A.17)
\]

\[
= f(x) + f'(x)h + \int_0^h \int_0^a f''(x + b) \, db \, da \quad (A.18)
\]

\[
= f(x) + f'(x)h + \int_0^h \left( f''(x) + \int_0^b f'''(x + c) \, dc \right) \, db \, da \quad (A.19)
\]

\[
= f(x) + f'(x)h + \frac{f''(x)}{2!}h^2 + \int_0^h \int_0^a \int_0^b f'''(x + c) \, dc \, db \, da \quad (A.20)
\]

\[
\vdots \quad (A.21)
\]

\[
= f(x) + \frac{f'(x)}{1!}h + \frac{f''(x)}{2!}h^2 + \frac{f'''(x)}{3!}h^3 + \ldots \quad (A.22)
\]

\[
= \sum_{n=0}^{\infty} \frac{f^{(n)}(x)}{n!}h^n \quad (A.23)
\]

In the formulation above, \(x\) is typically fixed and the function is evaluated in terms of \(h\). It is often more convenient to write the Taylor expansion of \(f(x)\) about a point \(a\) such that it remains a function of \(x\):

\[
f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!}(x - a)^n \quad (A.24)
\]

The Taylor expansion represents a function as an infinite sum of polynomial terms based on repeated derivatives at a single point. Any analytic function can be represented by its Taylor expansion within a local neighborhood.

A function can be locally approximated by using the first few terms of the Taylor expansion. Figure A.4 shows increasingly better approximations for \(\cos(x)\) about \(x = 1\). Including more terms increases the accuracy of the local approximation, but error still accumulates as one moves away from the expansion point.
A linear Taylor approximation uses the first two terms of the Taylor expansion:

\[ f(x) \approx f(a) + f'(a)(x - a) \]  \hspace{1cm} (A.25)

A quadratic Taylor approximation uses the first three terms:

\[ f(x) \approx f(a) + f'(a)(x - a) + \frac{1}{2}f''(a)(x - a)^2 \]  \hspace{1cm} (A.26)

and so on.

In multiple dimensions, the Taylor expansion about \( a \) generalizes to

\[ f(x) = f(a) + \nabla f(a)^\top (x - a) + \frac{1}{2}(x - a)^\top \nabla^2 f(a)(x - a) + \ldots \]  \hspace{1cm} (A.27)

The first two terms form the tangent plane at \( a \). The third term incorporates local curvature. This text will use only the first three terms shown here.

### A.13 Monte Carlo Estimation

Monte Carlo estimation allows us to evaluate the expectation of a function \( f \) when its input \( x \) follows a probability density function \( p \):

\[ \mathbb{E}_{x \sim p}[f(x)] = \int f(x)p(x) \, dx \approx \frac{1}{n} \sum_{i} f(x^{(i)}) \]  \hspace{1cm} (A.28)
where $x^{(1)}, \ldots, x^{(n)}$ are drawn from $p$. The variance of the estimate is equal to $\text{Var}_{x \sim p}[f(x)]/n$.

### A.14 Importance Sampling

Importance sampling allows us to compute $\mathbb{E}_{x \sim p}[f(x)]$ from samples drawn from a different distribution $q$:

$$
\mathbb{E}_{x \sim p}[f(x)] = \int f(x) p(x) \, dx
$$

(A.29)

$$
= \int f(x) p(x) \frac{q(x)}{q(x)} \, dx
$$

(A.30)

$$
= \int f(x) \frac{p(x)}{q(x)} q(x) \, dx
$$

(A.31)

$$
= \mathbb{E}_{x \sim q}\left[f(x) \frac{p(x)}{q(x)}\right]
$$

(A.32)

The equation above can be approximated using samples $x^{(1)}, \ldots, x^{(n)}$ drawn from $q$:

$$
\mathbb{E}_{x \sim p}[f(x)] = \mathbb{E}_{x \sim q}\left[f(x) \frac{p(x)}{q(x)}\right] \approx \frac{1}{n} \sum_{i} f(x^{(i)}) \frac{p(x^{(i)})}{q(x^{(i)})}
$$

(A.33)

### A.15 Contraction Mappings

A contraction mapping $f$ is defined with respect to a function over a metric space such that

$$
d(f(x), f(y)) \leq \alpha d(x, y)
$$

(A.34)

where $d$ is the distance metric associated with the metric space and $0 \leq \alpha < 1$. A contraction mapping thus reduces the distance between any two members of a set. Such a function is sometimes referred to as a contraction or contractor.

A consequence of repeatedly applying a contraction mapping is that the distance between any two members of the set is driven to 0. The contraction mapping theorem or the Banach fixed-point theorem states that every contraction mapping on a complete, non-empty metric space has a unique fixed point. Furthermore, for any element $x$ in that set, repeated application of a contraction mapping to that element results in convergence to that fixed point.

© 2022 Massachusetts Institute of Technology, shared under a under a Creative Commons CC-BY-NC-ND license.
2021-09-11 23:35:49-07:00, revision db8cdea, comments to bugs@algorithmsbook.com
Showing that a function \( f \) is a contraction mapping on a metric space is useful in various convergence proofs associated with the concepts presented earlier. For example, we can show that the Bellman operator is a contraction mapping on the space of value functions with the max-norm. Application of the contraction mapping theorem allows us to prove that repeated application of the Bellman operator results in convergence to a unique value function. Example A.1 shows a simple example of a contraction mapping.

Consider the function \( f(x) = [x_2/2 + 1, x_1/2 + 1/2] \). We can show \( f \) is a contraction mapping for the set \( \mathbb{R}^2 \) and the Euclidean distance function:

\[
d(f(x), f(y)) = \|f(x) - f(y)\|_2
\]

\[
= \|[x_2/2 + 1, x_1/2 + 1/2] - [y_2/2 + 1, y_1/2 + 1/2]\|_2
\]

\[
= \|[1/2(x_2 - y_2), 1/2(x_1 - y_1)]\|_2
\]

\[
= \frac{1}{2} \|[x_2 - y_2, x_1 - y_1]\|_2
\]

\[
= \frac{1}{2} d(x, y)
\]

We can plot the effect of repeated applications of \( f \) to points in \( \mathbb{R}^2 \) and show how they converge toward \([5/3, 4/3] \).
A.16 Graphs

A graph $G = (V, E)$ is defined by a set of nodes (also called vertices) $V$ and edges $E$. Figure A.5 shows an example of a graph. An edge $e \in E$ is a pair of nodes $(v_i, v_j)$. We focus primarily on directed graphs, where edges are directed and define parent-child relationships. An edge $e = (v_i, v_j)$ is often represented graphically as an arrow from $v_i$ to $v_j$ with $v_i$ as the parent and $v_j$ as the child. If there is an edge connecting $v_i$ and $v_j$, then we say that $v_i$ and $v_j$ are neighbors. The set of all parents of a node $v_i$ is denoted $\text{Pa}(v_i)$.

A path from node $v_i$ to node $v_j$ is a sequence of edges connecting $v_i$ to $v_j$. If this path can be followed from node to node along the direction of the edges, then we can call this path a directed path. A directed path is a path without regard to the direction of the edges. A node $v_j$ is a descendant of $v_i$ if there exists a directed path from $v_i$ to $v_j$. A cycle is a directed path from a node to itself. If a graph does not contain any cycles, it is acyclic.