A Mathematical Concepts

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 Ω. A sigma-algebra is a collection Σ of subsets of Ω 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 $E_1 \cup E_2 \cup E_2 \ldots \in \Sigma$ (*closed under countable unions*).

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

A *measure space* is defined by a set Ω, a sigma-algebra Σ, and a *measure* $\mu : \Omega \to \mathbb{R} \cup \{\infty\}$. For μ 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 $\mu(E_1 \cup E_2 \cup E_2 \ldots) = \mu(E_1) + \mu(E_2) + \mu(E_2) + \cdots$ (*countable additivity*).

A.2 Probability Spaces

A *probability space* is a measure space (Ω, Σ, μ) with the requirement that $\mu(\Omega) = 1$. In the context of probability spaces, Ω is called the *sample space*, Σ is called the *event space*, and μ (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 $\mu(\Omega) = 1$.

[1] These axioms are sometimes called the *Kolmorogov axioms*. A. Kolmogorov, *Foundations of the Theory of Probability*, 2nd ed. Chelsea, 1956.

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 $\|\cdot\|$, which maps elements of X to non-negative real numbers such that for all scalars α and vectors $\mathbf{x}, \mathbf{y} \in X$:

1. $\|\mathbf{x}\| = 0$ if and only if $\mathbf{x} = \mathbf{0}$.
2. $\|\alpha\mathbf{x}\| = |\alpha|\|\mathbf{x}\|$ (*absolutely homogeneous*).
3. $\|\mathbf{x} + \mathbf{y}\| \leq \|x\| + \|y\|$ (*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 $\mathbf{x}$ is

$$\|\mathbf{x}\|_p = \lim_{\rho \to p} (|x_1|^\rho + |x_2|^\rho + \cdots + |x_n|^\rho)^{\frac{1}{\rho}} \tag{A.1}$$

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

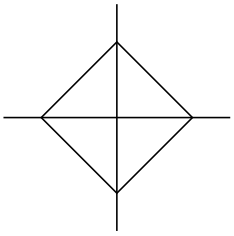
Norms can be used to induce distance metrics in vector spaces by defining the metric $d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|$. We can then, for example, use an L_p norm to define distances.

A.5 Positive Definiteness

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

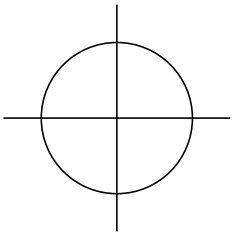
L_1: $\|\mathbf{x}\|_1 = |x_1| + |x_2| + \cdots + |x_n|$

This metric is often referred to as the *taxicab norm*.



L_2: $\|\mathbf{x}\|_2 = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}$

This metric is often referred to as the *Euclidean norm*.



L_∞: $\|\mathbf{x}\|_\infty = \max(|x_1|, |x_2|, \cdots, |x_n|)$

This metric is often referred to as the *max norm*, *Chebyschev norm*, or *chessboard norm*. The latter name comes from the minimum number of moves a chess king needs to move between two chess squares.

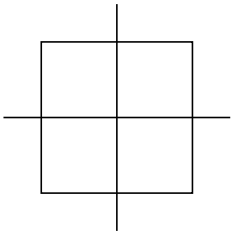


Table A.1. Common L_p norms. The illustrations show the shape of the norm contours in two dimensions. All points on the contour are equidistant from the origin under that norm.

A.6 Convexity

A *convex combination* of two vectors $\mathbf{x}$ and $\mathbf{y}$ is the result of

$$\alpha\mathbf{x} + (1-\alpha)\mathbf{y} \tag{A.2}$$

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

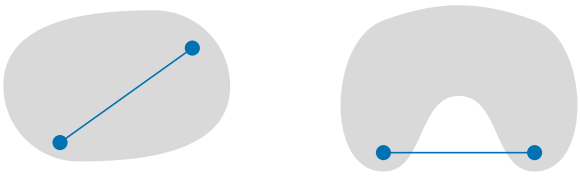
$$w_1\mathbf{v}^{(1)} + w_2\mathbf{v}^{(2)} + \cdots + w_m\mathbf{v}^{(m)} \tag{A.3}$$

with nonnegative weights $\mathbf{w}$ 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 $\mathcal{S}$ is convex if we have

$$\alpha\mathbf{x} + (1-\alpha)\mathbf{y} \in \mathcal{S}. \tag{A.4}$$

for all $\mathbf{x}, \mathbf{y}$ in $\mathcal{S}$ and for all α 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 nonconvex 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 $\mathcal{S}$ if, for all $\mathbf{x}, \mathbf{y}$ in $\mathcal{S}$ and for all α in $[0,1]$,

$$f(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}) \leq \alpha f(\mathbf{x}) + (1-\alpha)f(\mathbf{y}) \tag{A.5}$$

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

A function f is *strictly convex* over a convex set $\mathcal{S}$ if, for all $\mathbf{x}, \mathbf{y}$ in $\mathcal{S}$ and α in $(0,1)$,

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

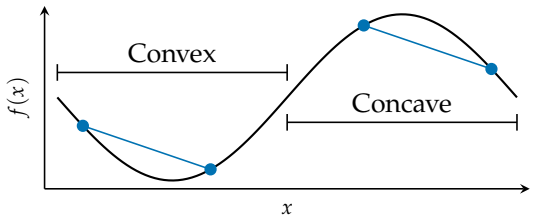
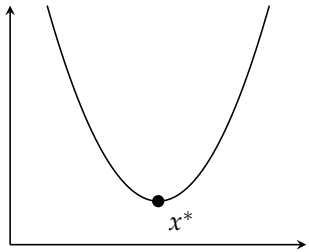
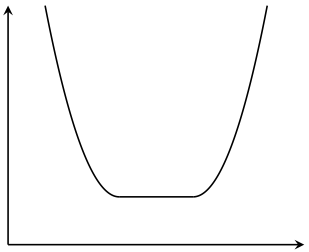


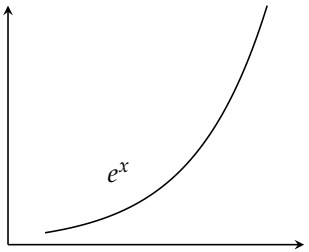
Figure A.2. Convex and nonconvex portions of a function.



Strictly convex function with one global minimum.



Convex function without a unique global minimum.



Strictly convex function without a global minimum.

Figure A.3. Not all convex functions have single global minima.

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.

[2] Optimization of convex functions is the subject of the textbook by S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

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

$$I(x) = -\log P(x) \tag{A.7}$$

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.

[3] Sometimes information content is referred to as *Shannon information*, in honor of the founder of the field of information theory. C. E. Shannon, "A Mathematical Theory of Communication," *Bell System Technical Journal*, vol. 27, no. 4, pp. 623–656, 1948.

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:

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

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

$$h(X) = \int p(x)I(x)\,dx = -\int p(x)\log p(x)\,\mathrm{d}x \tag{A.9}$$

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

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

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

$$H(p, q) = -\int p(x)\log q(x)\,\mathrm{d}x \tag{A.11}$$

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:

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

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.

[4] Named for the two American mathematicians who introduced this measure, Solomon Kullback (1907–1994) and Richard A. Liebler (1914–2003). S. Kullback, *Information Theory and Statistics*. Wiley, 1959.

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

$$D_{\text{KL}}(p \mid\mid q) = \int p(x) \log \frac{p(x)}{q(x)} \, \mathrm{d}x \tag{A.13}$$

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 $f(\mathbf{x})$ when f is a differentiable function. We begin at a point $\mathbf{x}$ and iteratively apply the following update rule:

$$\mathbf{x} \leftarrow \mathbf{x} + \alpha \nabla f(\mathbf{x}) \tag{A.14}$$

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 α will generally require more iterations to come close to a local maximum. Large values for α will often result in bouncing around the local optimum without quite reaching it. If α is constant over iterations, it is sometimes called a *learning rate*. Many applications involve a *decaying step factor* where, in addition to updating $\mathbf{x}$ at each iteration, we also update α 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

$$f(x + h) = f(x) + \int_0^h f'(x + a) \, \mathrm{d}a \tag{A.16}$$

[5] The first fundamental theorem of calculus relates a function to the integral of its derivative:

$$f(b) - f(a) = \int_a^b f'(x) \, \mathrm{d}x$$

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

$$f(x+h) = f(x) + \int_0^h \left(f'(x) + \int_0^a f''(x+b)\,\mathrm{d}b \right) \mathrm{d}a \tag{A.17}$$

$$= f(x) + f'(x)h + \int_0^h \int_0^a f''(x+b)\,\mathrm{d}b\,\mathrm{d}a \tag{A.18}$$

$$= f(x) + f'(x)h + \int_0^h \int_0^a \left(f''(x) + \int_0^b f'''(x+c)\,\mathrm{d}c \right) \mathrm{d}b\,\mathrm{d}a \tag{A.19}$$

$$= f(x) + f'(x)h + \frac{f''(x)}{2!}h^2 + \int_0^h \int_0^a \int_0^b f'''(x+c)\,\mathrm{d}c\,\mathrm{d}b\,\mathrm{d}a \tag{A.20}$$

$$\vdots \tag{A.21}$$

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

$$= \sum_{n=0}^{\infty} \frac{f^{(n)}(x)}{n!}h^n \tag{A.23}$$

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:

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

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:

$$f(x) \approx f(a) + f'(a)(x-a) \tag{A.25}$$

A quadratic Taylor approximation uses the first three terms:

$$f(x) \approx f(a) + f'(a)(x-a) + \frac{1}{2}f''(a)(x-a)^2 \tag{A.26}$$

and so on.

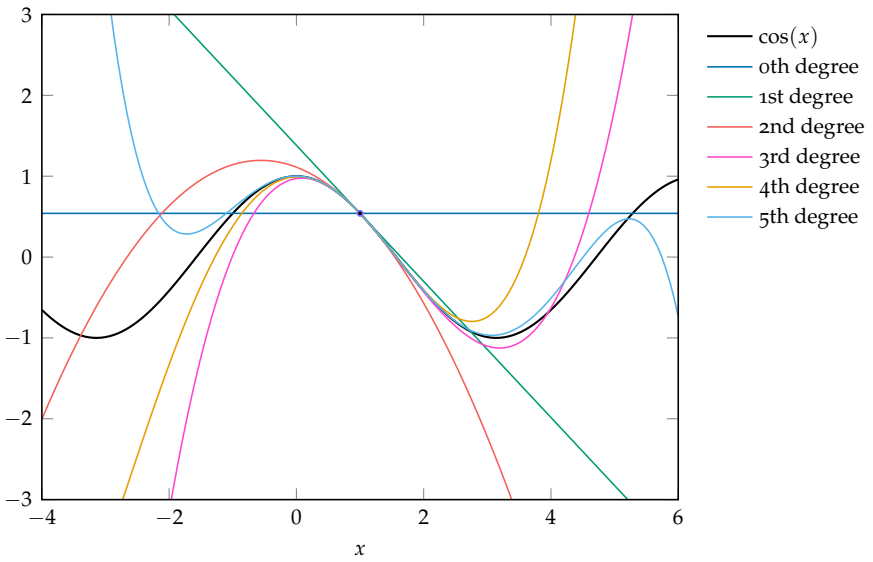


Figure A.4. Successive approximations of $\cos(x)$ about 1 based on the first n terms of the Taylor expansion.

In multiple dimensions, the Taylor expansion about $\mathbf{a}$ generalizes to

$$f(\mathbf{x}) = f(\mathbf{a}) + \nabla f(\mathbf{a})^\top (\mathbf{x} - \mathbf{a}) + \frac{1}{2}(\mathbf{x} - \mathbf{a})^\top \nabla^2 f(\mathbf{a})(\mathbf{x} - \mathbf{a}) + \ldots \tag{A.27}$$

The first two terms form the tangent plane at $\mathbf{a}$. 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:

$$\mathbb{E}_{x\sim p}[f(x)] = \int f(x)p(x)\,\mathrm{d}x \approx \frac{1}{n}\sum_i f(x^{(i)}) \tag{A.28}$$

where $x^{(1)}, \ldots, x^{(n)}$ are drawn from p. The variance of the estimate is equal to $\mathrm{Var}_{x\sim p}[f(x)]/n$.

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:

$$\mathbb{E}_{x\sim p}[f(x)] = \int f(x)p(x)\,\mathrm{d}x \tag{A.29}$$

$$= \int f(x)p(x)\frac{q(x)}{q(x)}\,\mathrm{d}x \tag{A.30}$$

$$= \int f(x)\frac{p(x)}{q(x)}q(x)\,\mathrm{d}x \tag{A.31}$$

$$= \mathbb{E}_{x\sim q}\left[f(x)\frac{p(x)}{q(x)}\right] \tag{A.32}$$

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

$$\mathbb{E}_{x\sim p}[f(x)] = \mathbb{E}_{x\sim q}\left[f(x)\frac{p(x)}{q(x)}\right] \approx \frac{1}{n}\sum_i f(x^{(i)})\frac{p(x^{(i)})}{q(x^{(i)})} \tag{A.33}$$

A.15 Contraction Mappings

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

$$d(f(x), f(y)) \leq \alpha d(x, y) \tag{A.34}$$

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,[6] 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.

[6] A complete metric space is one where every Cauchy sequence in that space converges to a point in that space. A sequence $x_1, x_2, \ldots$ is Cauchy, if for every positive real number $\epsilon > 0$ there is a positive integer n such that for all positive integers $i, j > n$, we have $d(x_i, x_j) < \epsilon$.

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.

Example A.1. An example contraction mapping for $\mathbb{R}^2$.

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

$$\begin{aligned} d(\mathbf{f}(\mathbf{x}), \mathbf{f}(\mathbf{y})) &= \|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y})\|_2 \\ &= \|[x_2/2 + 1, x_1/2 + 1/2] - [y_2/2 + 1, y_1/2 + 1/2]\|_2 \\ &= \|[\frac{1}{2}(x_2 - y_2), \frac{1}{2}(x_1 - y_1)]\|_2 \\ &= \frac{1}{2}\|[(x_2 - y_2), (x_1 - y_1)]\|_2 \\ &= \frac{1}{2} d(\mathbf{x}, \mathbf{y}) \end{aligned}$$

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