Math 584: Theoretical foundations of applied math

Christopher Henderson

March 25, 2024

Contents

1. Metrics and norms

1.1. **Metrics and metric spaces.** A metric is a way to define the notion of 'distance' on a space. This will be useful because, in applied math, we often need to approximate things that are complicated: e.g., how far away from the function $f(x) = 0$ for all x is the function

$$N_t(x) = \frac{1}{\sqrt{4\pi t}} e^{-\frac{x^2}{4t}} \quad \text{where } t > 0?$$

(This function N_t is related to the heat distribution on an 'infinite' domain at time t.)

In order to begin, let us formalize a notion of 'distance'

Definition 1.1.1 (Metric space)**.** *A metric space (X, d) is a set X and endowed with a* metric *d; that is, $d : X \times X \to \mathbb{R}$ that satisfies, for all $x, y, z \in X$:*

1. *(Positive definiteness) $d(x, y) \geq 0$ and $d(x, y) = 0$ if and only if $x = y$;*
2. *(Symmetry) $d(x, y) = d(y, x)$;*
3. *(Triangle inequality) $d(x, z) \leq d(x, y) + d(y, z)$.*

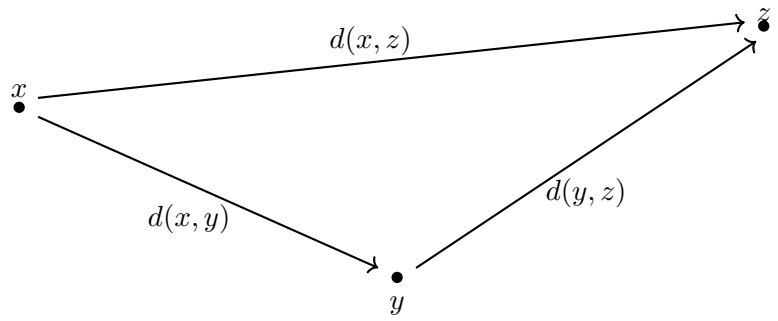


Figure 1: A cartoon showing the triangle inequality: notice that 'going' from x to z has to always be shorter than going from x to y and then from y to z.

Example 1.1.2. *(I) $(\mathbb{R}, d)$, where d is defined by* $d(x, y) = |x - y| = \begin{cases} x - y & \text{if } x \geq y, \\ y - x & \text{if } x < y. \end{cases}$

It is straightforward to check that (1) and (2) hold. Let us check (3):

$$\begin{aligned} |x - z|^2 = (x - z)^2 &= (x - y + y - z)^2 = (x - y)^2 + 2(x - y)(y - z) + (y - z)^2 \\ &\leq |x - y|^2 + 2|x - y||y - z| + |y - z|^2 \qquad \text{(because } (x - y)(y - z) \leq |x - y||y - z|) \\ &= (|x - y| + |y - z|)^2. \end{aligned}$$

By taking square roots, we get

$$|x - z| \leq |x - y| + |y - z|.$$

(II) $(\mathbb{R}^n, d_2)$ where d_2 is defined by

$$d_2(x, y) = \sqrt{(x_1 - y_1)^2 + \cdots + (x_n - y_n)^2}.$$

Exercise 1.1.1. *Show that this satisfies (1) - (3).*

(III) $(\mathbb{R}^n, d_p)$ where $p \geq 1$ and d_p is defined by

$$d_p(x, y) = \left(|x_1 - y_1|^p + \cdots + |x_n - y_n|^p\right)^{1/p}.$$

Note: (1) and (2) are still obvious. However, the triangle inequality is not as easy to prove. We will do it later in the class.

Exercise 1.1.2. *Show that this does not work for $p < 1$!*

(IV) $(\mathbb{R}^n, d_\infty)$ *where* d_∞ *is defined by*

$$d_\infty(x,y) = \max_{1\le i\le n} |x_i - y_i|.$$

Note: (1) and (2) are still obvious.

Exercise 1.1.3. *Show that the triangle inequality holds. Also, show that* $d_p(x,y) \to d_\infty(x,y)$ *as* $p \to \infty$.

(V) (X, d_{disc}) *where* X *is any set and the* discrete metric *is*

$$d_{\text{disc}}(x,y) = \begin{cases} 1 & \text{if } x \neq y, \\ 0 & \text{if } x = y. \end{cases}$$

There are many, many other examples of metric spaces. We will get into some other very important ones soon. But the above give us a nice basis to start our discussion.

Before we go on, let us dwell on the d_p metrics for a brief moment longer. First, we point out that, for any $x, y \in \mathbb{R}^n$, we have

$$\left\{ \begin{array}{ll} d_{p_1}(x,y) \le d_{p_2}(x,y) & \text{and} \\ d_{p_2}(x,y) \le n d_{p_1}(x,y) & \end{array} \right\} \qquad \text{whenever } p_2 < p_1.$$

We will prove this later, but it shows us that all of the metrics are, roughly, equivalent; that is, two points are close in one of these metrics only if they are close in *all* of these metrics. Formally, this is:

Definition 1.1.3. *Let* (X,d) *and* (X,ρ) *be metric spaces. We say that* d *and* ρ *are equivalent if there exists* $\underline{C}, \overline{C} \ge 1$ *such that, for all* $x, y \in X$,

$$\frac{1}{\underline{C}} d(x,y) \le \rho(x,y) \le \overline{C} d(x,y).$$

Exercise 1.1.4. *Is the same true of the discrete metric? If two points are 'close' in one of the* d_p *metrics, are they also close in the discrete metric? More concretely: for each* $p \in [1,\infty]$*:*

(i) Is it true that there is $\underline{C}$ *such that*

$$d_p(x,y) \le \underline{C} d_{\text{disc}}(x,y) \qquad \text{for all } x, y \in \mathbb{R}^n?$$

(ii) Is it true that there is $\overline{C}$ *such that*

$$d_{\text{disc}}(x,y) \le \overline{C} d_p(x,y) \qquad \text{for all } x, y \in \mathbb{R}^n?$$

To better illustrate what these metrics look like, let us draw the unit ball

$$B_1(0) = \{x \in \mathbb{R}^n : d_p(0,x) < 1\}$$

in each metric. We do it in two dimensions ($n = 2$) for simplicity. We end up with:

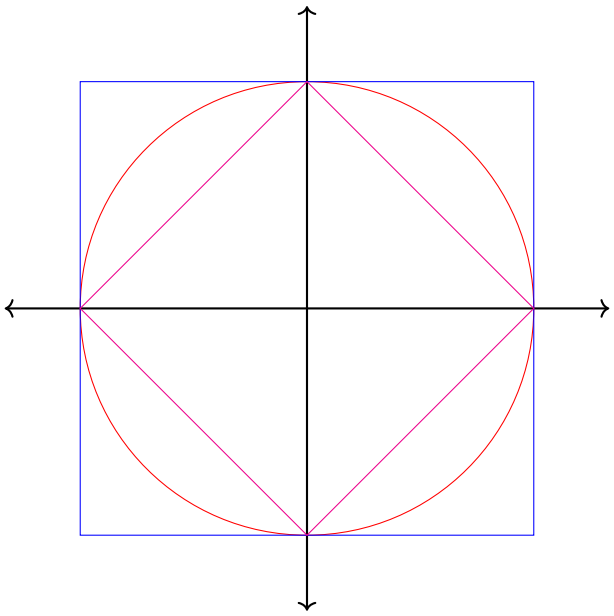


Figure 2: The unit ball corresponding to d_1, the unit ball corresponding to d_2, and the unit ball corresponding to d_∞.

1.2. **Norms and normed spaces.** Some[1] metric spaces are also vector spaces (that is, there is a notion of addition and scalar multiplication that satisfies the usual properties) and the notion of distance comes from the notion of the length of a vector. We call the 'length' of a vector its norm.

Definition 1.2.1. *If X is a vector space, then $\|\cdot\| : X \to \mathbb{R}$ is a* norm *if: for all $x, y \in X$ and $\alpha \in \mathbb{R}$:*

1. *(Positive definiteness) $\|x\| \geq 0$ and $\|x\| = 0$ if and only if $x = 0$;*
2. *(Homogeneity) $\|\alpha x\| = |\alpha|\|x\|$;*
3. *(Triangle inequality) $\|x + y\| \leq \|x\| + \|y\|$.*

To make the above comment more explicit, we see pretty quickly that, defining

$$d(x, y) = \|x - y\| \qquad \text{for all } x, y \in X,$$

(X, d) is also a metric space. Hence, all vector spaces with a norm are metric spaces. Note that, not all metric spaces come from a vector space with a norm. A metric space will fail to be a normed vector space if it is either (i) not a vector space, or (ii) if homogeneity fails.

Let us revisit our examples from above.

Example 1.2.2. *(III) $(\mathbb{R}^n, d_p)$ where $p \geq 1$ and d_p is defined by*

$$d_p(x, y) = \left(|x_1 - y_1|^p + \cdots + |x_n - y_n|^p\right)^{\frac{1}{p}}.$$

[1]Note that there are two conditions here: (1) metric space is also a vector space, and (2) metric is related to the norm.

This is clearly a vector space. And its metric comes from the norm

$$\|x\|_p = d_p(x,y) = \left(|x_1|^p + \cdots + |x_n|^p\right)^{\frac{1}{p}}, \tag{1.2.1}$$

It inherits positive definiteness from the positive definiteness of d_p since

$$\|x\|_p = d_p(0,x).$$

It inherits the triangle inequality from that of d_p:

$$\|x+y\|_p = d_p(x,-y) \leq d_p(x,0) + d_p(0,-y) = \|x\|_p + \|y\|_p.$$

On the other hand, homogeneity is obvious from (1.2.1)*:*

$$\begin{aligned}\|\alpha x\|_p &= \left(|\alpha x_1|^p + \cdots + |\alpha x_n|^p\right)^{\frac{1}{p}} = \Big(|\alpha|^p\big(|x_1|^p + \cdots + |x_n|^p\big)\Big)^{\frac{1}{p}} \\ &= |\alpha|\left(|x_1|^p + \cdots + |x_n|^p\right)^{\frac{1}{p}} = |\alpha|\|x\|_p.\end{aligned}$$

(IV) $(\mathbb{R}^n, d_\infty)$ *where* d_∞ *is defined by*

$$d_\infty(x,y) = \max_{1\leq i\leq n} |x_i - y_i|.$$

As above, this is (cleary) a vector space and its metric comes from the norm

$$\|x\|_\infty = d_\infty(x,0) = \max_{1\leq i\leq n} |x_i|.$$

Exercise 1.2.1. *Check that this is a norm.*

(V) (X, d_{disc}) *where X is any set and the* discrete metric *is*

$$d_{\text{disc}}(x,y) = \begin{cases} 1 & \text{if } x \neq y, \\ 0 & \text{if } x = y. \end{cases}$$

This is certainly not related to a normed vector space. First, X might not be a vector space (e.g., what if $X = \{$dog, cat, donkey$\}$? What is the notion of addition of vectors or scaling? Thus X is not a vector space). Even if X is a vector space, d_{disc} will not come from a norm. Indeed, suppose that $X = \mathbb{R}^n$ and d_{disc} came from a norm:

$$\|x\|_{\text{disc}} = d_{\text{disc}}(x,0).$$

However, if $x \neq 0$ then $2x \neq 0$, so that

$$\|2x\|_{\text{disc}} = 1 \neq 2 = 2\|x\|_{\text{disc}}.$$

Hence, homogeneity does not hold.

1.2.1. **Aside: how to generate non-norm metrics from norms.**

Proposition 1.2.3. *Suppose that (X, d) is a metric space and $f : [0, \infty) \to [0, \infty)$ is a C^2 function*[2] *such that*

(i) f is nondecreasing (that is, $f' \geq 0$);

(ii) $f(0) = 0$ and $f(x) > 0$ if $x > 0$;

(iii) f is concave (that is, $f'' \leq 0$).

Then

$$d_f(x, y) := f(d(x, y))$$

is a norm as well. Further, if d is given by a norm $d(x, y) = \|x - y\|$, then d_f is given by a norm if and only if there is $c > 0$ such that $f(x) = cx$ for all $x \in [0, \infty)$.

Exercise 1.2.2. *Show that last claim; that is, that d_f is only given by a norm if and only if there is $c > 0$ such that $f(x) = cx$ for all $x \in [0, \infty)$.*

Proof. Given the exercise above, all that remains is to show that d_f is a metric.

First we show that d_f is positive definite. It is nonnegative since $f \geq 0$, by assumption. Second, $d_f(x, y) = 0$ if and only if $f(d(x, y)) = 0$, which occurs if and only if $d(x, y) = 0$ (see point (ii) above). Since d is a metric, $d(x, y) = 0$ if and only if $x = y$. In summary, $d_f(x, y) = 0$ if and only if $x = y$.

Second, we note that d_f is clearly symmetric – this is inherited directly from the symmetry of d.

Third, we check the triangle inequality. This is the hardest part. We begin by stating a claim that we prove later:

Claim: $d_f(x, y) + d_f(y, z) - f(d(x, y) + d(y, z)) \geq 0$ for all $x, y \in X$.

Assuming this claim, momentarily, notice that it is equivalent to

$$f(d(x, y) + d(y, z)) \leq d_f(x, y) + d_f(y, z). \tag{1.2.2}$$

Note that, since d is a metric:

$$d(x, z) \leq d(x, y) + d(y, z).$$

As f is nondecreasing, we deduce that

$$f(d(x, z)) \leq f(d(x, y) + d(y, z)). \tag{1.2.3}$$

Thus,

$$d_f(x, z) = f(d(x, z)) \overset{(1.2.3)}{\leq} f(d(x, y) + d(y, z)) \overset{(1.2.2)}{\leq} d_f(x, y) + d_f(y, z).$$

We now prove the claim. Notice that it is obvious if either $d(x, y) = 0$ or $d(y, z) = 0$. Hence, we need only consider the case

$$d(x, y),\ d(y, z) > 0.$$

[2] $f \in C^2$ means that f, f', and f'' all exist and are continuous.

Next, fix any $\beta \geq 0$ and let $g : [0, \infty) \to \mathbb{R}$ be defined by

$$g(\theta) = f(\theta) + f(\beta) - f(\theta + \beta) - f(\theta + \beta).$$

Notice that the claim is equivalent to showing that

$$g(d(x, y)) \geq 0 \quad \text{for } \beta = d(y, z).$$

Actually, we show something stronger,

$$g(\theta) \geq 0 \qquad \text{and for all } \theta, \beta > 0. \tag{1.2.4}$$

We now prove (1.2.4). Clearly

$$g(0) = 0 \quad \text{and} \quad g'(\theta) = f'(\theta) - f'(\theta + \beta) \geq 0.$$

The inequality above follows from the fact that $\theta + \beta > \theta$ and the fact that f' is nonincreasing (recall that f is concave). Hence, g is nondecreasing. It follows that

$$g(\theta) \geq g(0) = 0.$$

Since this is true for all $\theta, \beta > 0$, we obtain exactly (1.2.4). This concludes the proof. □

Example 1.2.4. *1. We then have that, for any $p \in [1, +\infty]$ (allowing for $p = \infty$),*

$$d(x, y) = \log(1 + \|x - y\|_p)$$

is a metric on $\mathbb{R}^n$ does not define a norm. This arises from Proposition 1.2.3 with the choice $f(x) = \log(1 + x)$.

2. Another metric on $\mathbb{R}^n$ not arising from a norm is: for any $p \in [1, +\infty]$ (allowing for $p = \infty$),

$$d(x, y) = \frac{\|x - y\|_p}{1 + \|x - y\|_p}.$$

This arises from Proposition 1.2.3 with the choice $f(x) = {}^{x}/_{(1+x)}$.

1.3. **Cartesian products.** Given metric spaces (X, d_X) and (Y, d_Y) (respectively, normed spaces $(X, \|\cdot\|_X)$ and $(Y, \|\cdot\|_Y)$), one obtains a new metric space (resp. normed space) as follows:

$$(X \times Y, d_{X\times Y}) \quad \text{where } \begin{cases} X \times Y = \{(x, y) : x \in X, y \in Y\} \text{ and} \\ d((x_1, y_1), (x_2, y_2)) = d_X(x_1, x_2) + d_X(y_1, y_2). \end{cases}$$

(resp. $\|(x, y)\|_{X\times Y} = \|x\|_X + \|y\|_Y$).

Example 1.3.1. *Let d_{euc} be the Euclidean metric on $\{0, 1, 3\}$: $d_{\text{euc}}(x, y) = |x - y|$ and let d_{disc} be the discrete metric on the set $X = \{\text{a}, \text{dog}\}$. Then*

$$\mathbb{R} \times X = \{(0, \text{a}), (0, \text{dog}), (1, \text{a}), (1, \text{dog}), (3, \text{a}), (3, \text{dog})\}.$$

Then, we have, e.g.,

$$d_{\mathbb{R}\times X}((0, \text{a}), (0, \text{dog})) = d_{\text{euc}}(0, 0) + d_{\text{disc}}(\text{a}, \text{dog}) = 0 + 1 = 1,$$

$$d_{\mathbb{R}\times X}((1, \text{dog}), (3, \text{dog})) = d_{\text{euc}}(1, 3) + d_{\text{disc}}(\text{dog}, \text{dog}) = 2 + 0 = 2.$$

and

$$d_{\mathbb{R}\times X}((1, \text{a}), (3, \text{dog})) = d_{\text{euc}}(1, 3) + d_{\text{disc}}(\text{a}, \text{dog}) = 2 + 1 = 3.$$

By convention, we can take many copies of the same space X. Then we write:

$$X^2 = X \times X, \quad X^3 = X \times X \times X \quad \text{and, more generally} \quad X^n = \underbrace{X \times X \times \cdots \times X}_{n \text{ times}}.$$

Can one take infinitely many Cartesian products? We have to be a bit more careful...

To do this, let us change perspective. One can also think of X^n as being equivalent to $X^{\{1,2,\ldots,n\}}$, where we use the notation

$$X^Y = \{f : Y \to X\} \quad \text{(that is, } X^Y \text{ is the set of all functions from } Y \text{ to } X\text{)}.$$

Why are X^n and $X^{\{1,2,\ldots,n\}}$ the same? Notice that if $f \in X^{\{1,2,\ldots,n\}}$ then we can define a point

$$(f(1), f(2), \ldots, f(n)) \in X^n.$$

Likewise, if $\bar{x} \in X^n$, then $x = (x_1, x_2, \ldots, x_n)$ with each $x_i \in X$. Hence, we define

$$f \in X^{\{1,2,\ldots,n\}} \quad \text{by } f(i) = x_i.$$

Now it is easy to define an infinite Cartesian product:

$$X^{\mathbb{N}} = \{(x_1, x_2, x_3, \ldots) : x_i \in X \text{ for each } i \in \mathbb{N}\} = \{f : \mathbb{N} \to X\}.$$

(Note that one can also take $X^{\mathbb{Z}}$ or $X^{\mathbb{R}}$, which might be very different sets *a priori*).

What is not clear is what the metric on this set should be?! Indeed, using our construction above

$$d_{X^{\mathbb{N}}}(f, g) = d(f(1), g(1)) + d(f(2), g(2)) + \cdots,$$

which will usually be infinite and, therefore, lose some of its 'meaning' as a distance function. Let us introduce some important examples:

1.4. **Sequence spaces.** For any $p \geq 1$, let

$$\ell^p(\mathbb{N}) = \mathbb{R}^{\mathbb{N}} = \{\bar{x} = (x_1, x_2, \ldots) : x_i \in \mathbb{R} \text{ and } \|\bar{x}\|_{\ell^p} < \infty\},$$

where we use the norm

$$\|\bar{x}\|_{\ell^p} = \left(\sum_{i=1}^{\infty} |x_i|^p \right)^{1/p}.$$

Recall that any norm defines, also, a metric. Notice that this the infinite dimensional version of the d_p metric introduced in Example 1.1.2.

We can also define this for $p = \infty$:

$$\|\bar{x}\|_{\ell^\infty} = \sup_{i \in \mathbb{N}} |x_i|.$$

There are two slightly smaller spaces that are useful in the future (especially for finding counterexamples to false statements):

$$c_0 = \{(x_1, x_2, \ldots) \in \ell^\infty(\mathbb{N}) : \lim_{i \to \infty} x_i = 0\}$$

and

$$c = \{(x_1, x_2, \dots) \in \ell^\infty(\mathbb{N}) : \lim_{i\to\infty} x_i \text{ exists}\}.$$

Let us briefly clarify the definition of limit: we say that

$$\lim_{i\to\infty} x_i = L$$

if, for every $\varepsilon > 0$, there exists N such that

$$|L - x_n| < \varepsilon \qquad \text{whenever } n \geq N.$$

One can also write $\ell^p(\mathbb{R}, \mathbb{N})$ to emphasize that the sequences are real-valued (as opposed to, e.g., $\ell^p(\mathbb{C}, \mathbb{N})$, where the entries take complex values). Often we simply write ℓ^p instead of $\ell^p(\mathbb{N})$ out of laziness.

The space $\ell^2(\mathbb{Z}) = \mathbb{R}^{\mathbb{Z}}$ is extremely important in the theory of Fourier series, but all ℓ^p spaces will be useful for developing our understanding of norms, linear spaces, and etc.

Example 1.4.1. *(i)* $(1, -1, 1, -1, 1, -1, \dots) \in \ell^p$ *for* $p = \infty$ *only. It is not an element of* c *or* c_0*;*

(ii) $(1, 2, 3, \dots) \notin \ell^p$ *for any* p*;*

(iii) $(1, 1/2, 1/3, 1/4, \dots) \in \ell^p$ *for* $p > 1$ *only (Exercise!);*

(iv) Any element of ℓ^p *for* $p < \infty$ *is in* c_0 *and* c *(Exercise!).*

1.4.1. **Some brief reminders about series.**

Definition 1.4.2. *A series* $\sum_{n=1}^{\infty} a_n$ *is* (conditionally) convergent *if there exists* $L \in \mathbb{R}$ *such that*

$$\lim_{n\to\infty} \sum_{n=1}^{\infty} a_n = L.$$

Recall that this means: for every $\varepsilon > 0$ *there is* N_ε *such that*

$$\left| L - \sum_{n=1}^{N} a_n \right| < \varepsilon \quad \text{for all } N \geq N_\varepsilon.$$

Exercise 1.4.1. *Show that if* $\sum_{n=1}^{\infty} a_n$ *is conditionally convergent then* $\lim_{n\to\infty} a_n = 0$.

Definition 1.4.3. *A series* $\sum_{n=1}^{\infty} a_n$ *is* absolutely summable *if there exists* $C \geq 0$ *such that*

$$\sum_{n=1}^{N} |a_n| \leq C \quad \text{for all } N.$$

Exercise 1.4.2. *Suppose that* $\sum_{n=1}^{\infty} a_n$ *is absolutely summable. Show that* $\sum_{n=1}^{\infty} |a_n|$ *is convergent.*

Exercise 1.4.3. *Show that every absolutely summable series is convergent. Is the converse true?*

Proposition 1.4.4 (Comparison test)**.** *Suppose that* (a_n) *and* (b_n) *are sequences and that*

$$|a_n| \leq b_n \quad \text{for all } n.$$

If b_n *is summable then* a_n *is summable.*

Proof. Let C be such that

$$\sum_{n=1}^{\infty} b_n = \sum_{n=1}^{\infty} |b_n| \leq C \quad \text{for all } N.$$

Then, for any $N \in \mathbb{N}$, we have

$$\sum_{n=1}^{N} |a_n| \leq \sum_{n=1}^{N} b_n \leq C.$$

Thus a_n is summable. □

1.4.2. $\|\cdot\|_{\ell^\infty}$ **is a norm.**

Proof. **Positive definiteness:** Clearly $\|\bar{x}\| \geq 0$. On the other hand,

$$\|\bar{x}\|_{\ell^\infty} = 0 \iff \sup_i |x_i| = 0 \iff |x_i| = 0 \text{ for all } i \iff \bar{x} = 0.$$

Hence, $\|\cdot\|_{\ell^\infty}$ is postive definite.

Triangle inequality: Fix $\bar{x}, \bar{y} \in \ell^\infty$. For any $i = 1, 2, 3, \ldots$, we have

$$|(\bar{x} + \bar{y})_i| = |x_i + y_i| \leq |x_i| + |y_i| \leq \|\bar{x}\|_{\ell^\infty} + \|\bar{y}\|_{\ell^\infty}.$$

Since this is true for all i, we deduce

$$\|(\bar{x} + \bar{y})_i\|_{\ell^\infty} \leq \|\bar{x}\|_{\ell^\infty} + \|\bar{y}\|_{\ell^\infty}.$$

Homogeneity: Take any $\alpha \in \mathbb{R}$ and any $\bar{x} \in \ell^\infty$. Fix any $\varepsilon > 0$. Then, by definition of supremum, there is i_ε such that

$$(1 - \varepsilon)\|\bar{x}\|_{\ell^\infty} \leq |x_{i_\varepsilon}|$$

It follows that

$$|\alpha|(1 - \varepsilon)\|\bar{x}\|_{\ell^\infty} \leq |\alpha||x_{i_\varepsilon}| = |\alpha x_{i_\varepsilon}|,$$

which implies that

$$|\alpha|(1 - \varepsilon)\|\bar{x}\|_{\ell^\infty} \leq \|\alpha\bar{x}\|_{\ell^\infty}. \tag{1.4.1}$$

On the other hand, for all i,

$$|x_i| \leq \|\bar{x}\|_{\ell^\infty}.$$

Thus, for all i,

$$|\alpha x_i| = |\alpha||x_i| \leq |\alpha|\|\bar{x}\|_{\ell^\infty},$$

which implies that

$$\|\alpha\bar{x}\| = \sup_i |\alpha x_i| \leq |\alpha|\|\bar{x}\|_{\ell^\infty}. \tag{1.4.2}$$

The combination of (1.4.1) and (1.4.2) yields

$$|\alpha|(1-\varepsilon)\|\bar{x}\|_{\ell^\infty} \leq \|\alpha\bar{x}\| \leq |\alpha|\|\bar{x}\|_{\ell^\infty}.$$

Since this is true for all ε, it follows that

$$\|\alpha\bar{x}\| = |\alpha|\|x\|_{\ell^\infty},$$

that is, $\|\cdot\|_{\ell^\infty}$ is homogeneous. □

1.4.3. $\|\cdot\|_{\ell^p}$ **is a norm when** $1 \leq p < \infty$**.** Before starting, we need the following elementary inequality

Lemma 1.4.5 (Young's inequality)**.** *For any $a, b \geq 0$ and $p, q > 1$ satisfying $1 = \frac{1}{p} + \frac{1}{q}$, we have*

$$ab \leq \frac{a^p}{p} + \frac{b^q}{q}.$$

The is equality if and only if $b = \operatorname{sign}(a)|a|^{p-1}$.

Some notes:

- There are many "Young's inequalities."
- We call p and q satisfying $1 = \frac{1}{p} + \frac{1}{q}$ conjugate exponents.
- When $p = q = 2$, this is just the AM-GM inequality in the special case when $n = 2$.

Exercise 1.4.4. *Prove Young's inequality. Hint: fix a and use calculus to check that $f(x) = {}^{a^p}/_{p} + {}^{x^q}/_{q} - ax$ is nonnegative. In particular, show that $\min_{[0,\infty)} f = 0$.*

The proof of positive definiteness and homogeneity are simple and so we omit them. The main step is in proving the triangle inequality, which, in this case, goes by the special name Minkowski's inequality. We prove this and another important inequality in the following:

Theorem 1.4.6. *(I) (Hölder's inequality) For any $\bar{x} \in \ell^p$ and $\bar{y} \in \ell^q$, where*

$$1 = \frac{1}{p} + \frac{1}{q} \quad \text{where } 1 \leq p, q \leq +\infty.$$

we have

$$\sum_{i=1}^{\infty} x_i y_i \leq \|\bar{x}\|_{\ell^p}\|\bar{y}\|_{\ell^q}. \tag{1.4.3}$$

The inequality (1.4.3) *is an equality if and only if there is $c > 0$ such that*

$$y_i = c\operatorname{sign}(x_i)|x_i|^{p-1}.$$

Note that when $p = q = 2$, this also goes by the "Cauchy-Schwarz inequality."

(II) (Minkowski's Inequality) For any $\bar{x}, \bar{y} \in \ell^p$, we have

$$\|x + y\|_{\ell^p} \leq \|x\|_{\ell^p} + \|y\|_{\ell^p}.$$

Proof of Hölder's inequality. Without loss of generality we assume that $p \leq q$ (otherwise simply swap them!). The case $p = 1$ and $q = +\infty$ is obvious (convince yourself of this!). Hence, we consider only the case when $1 < p \leq q < +\infty$.

To begin, we note that the inequality is clearly true if either $\bar{x}$ or $\bar{y} = 0$. Hence, we may assume that

$$\|\bar{x}\|_{\ell^p},\ \|\bar{y}\|_{\ell^q} > 0.$$

The main tool is Young's inequality with a little trick. Fix $\varepsilon > 0$ to be chosen. Then

$$\begin{aligned} \sum_{i=1}^{\infty} x_i y_i \leq \sum_{i=1}^{\infty} |x_i||y_i| = \sum_{i=1}^{\infty} \frac{|x_i|}{\varepsilon}(\varepsilon|y_i|) \leq \sum_{i=1}^{\infty} \left(\frac{|x_i|^p}{p\varepsilon^p} + \frac{\varepsilon^q |y_i|^q}{q} \right) \\ = \frac{1}{p\varepsilon^p} \sum_{i=1}^{\infty} |x_i|^p + \frac{\varepsilon^q}{q} \sum_{i=1}^{\infty} |y_i|^q = \frac{1}{p\varepsilon^p} \|\bar{x}\|_{\ell^p}^p + \frac{\varepsilon^q}{q} \|\bar{y}\|_{\ell^q}^q. \end{aligned} \tag{1.4.4}$$

We now choose

$$\varepsilon = \frac{\|\bar{x}\|_{\ell^p}^{1-\frac{1}{p}}}{\|\bar{y}\|_{\ell^q}^{1-\frac{1}{q}}}.$$

Notice that, because p and q are conjugate exponents,

$$\varepsilon^p = \frac{\|\bar{x}\|_{\ell^p}^{p-1}}{\|\bar{y}\|_{\ell^q}} \quad \text{and} \quad \varepsilon^q = \frac{\|\bar{y}\|_{\ell^q}^{q-1}}{\|\bar{x}\|_{\ell^p}}.$$

Thus (1.4.4) becomes

$$\sum_{i=1}^{\infty} x_i y_i \leq \frac{1}{p} \|\bar{x}\|_{\ell^p} \|\bar{y}\|_{\ell^q} + \frac{1}{q} \|\bar{x}\|_{\ell^p} \|\bar{y}\|_{\ell^q} = \left(\frac{1}{p} + \frac{1}{q} \right) \|\bar{x}\|_{\ell^p} \|\bar{y}\|_{\ell^q} = \|\bar{x}\|_{\ell^p} \|\bar{y}\|_{\ell^q},$$

which concludes the proof.

Note that the only inequalities used above were (1) $x_i y_i \leq |x_i||y_i|$ and (2) Young's inequality. These are both equality if and only if the stated condition holds. □

Proof of Minkowski's inequality. The cases when $p = \infty$ was already proven above (Section 1.4.2). Hence, we consider only $1 \leq p < \infty$. Also, notice that there is nothing to prove if $\|\bar{x} + \bar{y}\|_{\ell^p} = 0$. Hence, we assume that $\|\bar{x} + \bar{y}\|_{\ell^p} > 0$.

Let us first prove the claim assuming that $x_i, y_i \geq 0$ for all i. Let $\bar{z} = (z_1, z_2, \dots)$ with

$$z_i = (x_i + y_i)^{p-1}$$

and notice that, letting q be the conjugate exponent to p,

$$\|\bar{z}\|_{\ell^q}^q = \|\bar{x} + \bar{y}\|_{\ell^p}^p,$$

so that, noting $p/q = p - 1$

$$\|\bar{z}\|_{\ell^q} = \|\bar{x} + \bar{y}\|_{\ell^p}^{p/q} = \|\bar{x} + \bar{y}\|_{\ell^p}^{p-1}.$$

Then, using Hölder's inequality, we find

$$\|x+y\|_{\ell^p}^p = \sum_{i=1}^{\infty}(x_i+y_i)^p = \sum_{i=1}^{\infty}(x_i+y_i)(x_i+y_i)^{p-1} = \sum_{i=1}^{\infty}(x_i+y_i)z_i = \sum_{i=1}^{\infty}x_iz_i + \sum_{i=1}^{\infty}y_iz_i$$
$$\leq \|\bar{x}\|_{\ell^p}\|\bar{z}\|_{\ell^q} + \|\bar{y}\|_{\ell^p}\|\bar{z}\|_{\ell^q} = \big(\|\bar{x}\|_{\ell^p} + \|\bar{y}\|_{\ell^p}\big)\|\bar{z}\|_{\ell^q} = \big(\|\bar{x}\|_{\ell^p} + \|\bar{y}\|_{\ell^p}\big)\|\bar{x}+\bar{y}\|_{\ell^p}^{p-1}.$$

The proof is then finished after dividing both sides by $\|\bar{x}+\bar{y}\|_{\ell^p}^{p-1}$.

Now we consider the general case. One can easily check that

$$|x_i + y_i| \leq \big||x_i| + |y_i|\big|.$$

Letting $\tilde{x} = (|x_1|, |x_2|, \dots)$ and $\tilde{y} = (|y_1|, |y_2|, \dots)$, we have

$$\|\bar{x}+\bar{y}\|_{\ell^p}^p = \sum_{i=1}^{\infty}|x_i+y_i|^p \leq \sum_{i=1}^{\infty}\big||x_i|+|y_i|\big|^p = \|\tilde{x}+\tilde{y}\|_{\ell^p}^p.$$

The proof is finished after applying Minkowski's inequality to $\tilde{x}$ and $\tilde{y}$ and using that $\|\bar{x}\|_{\ell^p} = \|\tilde{x}\|_{\ell^p}$ and $\|\bar{y}\|_{\ell^p} = \|\tilde{y}\|_{\ell^p}$. □

Hölder's inequality is extremely important in analysis because it tells us that ℓ^p and ℓ^q are dual to each other (when p and q are conjugate exponents). We will dig into this more later. But for now, we note that we can compute the ℓ^p norm 'by duality'; that is:

Proposition 1.4.7. *Suppose that $x \in \ell^p$, with $p \in [1,\infty]$. Then*

$$\|x\|_{\ell^p} = \sup_{\substack{y\in\ell^q,\\ \|y\|_{\ell^q}\leq 1}} \langle x, y\rangle = \sup_{\substack{y\in\ell^q,\\ \|y\|_{\ell^q}=1}} \langle x, y\rangle, \tag{1.4.5}$$

where we define

$$\langle x, y\rangle = \sum_{i=1}^{\infty} x_iy_i.$$

When $p \in (1,\infty)$, the sup *above can be replaced by a* max.

Exercise 1.4.5. *Understand why the following proof gives the second equality in* (1.4.5).

Exercise 1.4.6. *Prove the proposition in the cases where $p = 1$ and $p = +\infty$.*

Proof. We show the case $p, q \in (1,\infty)$. Also, notice that we do not need to consider the case $x = (0, 0, \dots)$ as this case is obvious. Hence, we assume that $\|x\|_{\ell^p} > 0$.

By Hölder's inequality, we clearly have that

$$\max_{\substack{y\in\ell^q,\\ \|y\|_{\ell^q}\leq 1}} \langle x, y\rangle \leq \max_{\substack{y\in\ell^q,\\ \|y\|_{\ell^q}\leq 1}} \|x\|_{\ell^p}\|y\|_{\ell^q} \leq \|x\|_{\ell^p}.$$

Thus, we need only show the reverse inequality.

Define $\bar{y} = (y_1, y_2, \dots)$ where, for all i,

$$y_i = \frac{\operatorname{sign}(x_i)|x_i|^{p-1}}{\|x\|_{\ell^p}^{p-1}},$$

where, for any $r \in \mathbb{R}$,

$$\operatorname{sign}(r) = \begin{cases} \frac{r}{|r|} & \text{if } r \neq 0, \\ 0 & \text{if } r = 0. \end{cases}$$

First note that $q(p-1) = p$ since p and q are conjugate exponents. Hence,

$$\sum_{i=1}^{\infty} |y_i|^q = \frac{1}{\|x\|_{\ell^p}^{q(p-1)}} \sum_{i=1}^{\infty} |x_i|^{q(p-1)} = \frac{1}{\|x\|_{\ell^p}^p} \sum_{i=1}^{\infty} |x_i|^p = \frac{1}{\|x\|_{\ell^p}^p} \|x\|_{\ell^p}^p = 1.$$

Thus, $\bar{y} \in \ell^q$ and y is an admissible test sequence since $\|\bar{y}\|_{\ell^q} \leq 1$. It follows that

$$\max_{\substack{y \in \ell^q, \\ \|y\|_{\ell^q} \leq 1}} \langle x, y \rangle \geq \langle x, \bar{y} \rangle = \sum_{i=1}^{\infty} x_i y_i = \sum_{i=1}^{\infty} \frac{|x_i|^p}{\|x\|_{\ell^p}^{p-1}} = \frac{\|x\|_{\ell^p}^p}{\|x\|_{\ell^p}^{p-1}} = \|x\|_{\ell^p}.$$

□

1.5. **Function spaces.** In this section, we will define a family of norms on functions that is analogous to the family of ℓ^p-norms. This will give us a number of different ways to measure the "size" of a function.

1.5.1. **A silly example.** Take any $p \in [1, \infty)$ and any element $\bar{x} \in \ell^p(\mathbb{N})$. Recall that $\bar{x}$ is "equivalent" to a function $f_{\bar{x}} : \mathbb{N} \to \mathbb{R}$ defined by

$$f_{\bar{x}}(n) = x_n.$$

By an abuse of notation, we can consider this a function on all of $\mathbb{R}$:

$$f_{\bar{x}}(x) = \begin{cases} x_n & \text{if } x \in [n, n+1), \\ 0 & \text{if } x < 1. \end{cases}$$

With this definition, we find that

$$\|\bar{x}\|_{\ell^p}^p = \sum_{n=1}^{\infty} |x_n|^p = \sum_{n=1}^{\infty} |f_{\bar{x}}(n)|^p = \int_{\mathbb{R}} |f_{\bar{x}}(x)|^p dx.$$

This suggests a natural norm on "suitable" functions

$$\|f\|_{L^p(\mathbb{R})} = \left(\int_{\mathbb{R}} |f(x)|^p dx \right)^{\frac{1}{p}}.$$

The meaning of suitable is something that requires some work that we will ignore for now and address later.

To get functions that are not zero for $x < 1$, one could go even crazier with this and consider sequences $\bar{x} \in \ell^p(\mathbb{Z})$ and construct functions $f_{\bar{x}} : \mathbb{R} \to \mathbb{R}$ from them. Recall that

$$\begin{aligned} \ell^p(\mathbb{Z}) = \ell^p(\mathbb{R}, \mathbb{Z}) &= \left\{ f \in \mathbb{R}^{\mathbb{Z}} : \sum_{n=-\infty}^{\infty} |f(n)|^p < \infty \right\} \\ &= \left\{ (\dots, x_{-2}, x_{-1}, x_0, x_1, x_2, \dots) : x_i \in \mathbb{R} \text{ for all } i \in \mathbb{Z}, \sum_{n=-\infty}^{\infty} |x_n|^p < \infty \right\}. \end{aligned}$$

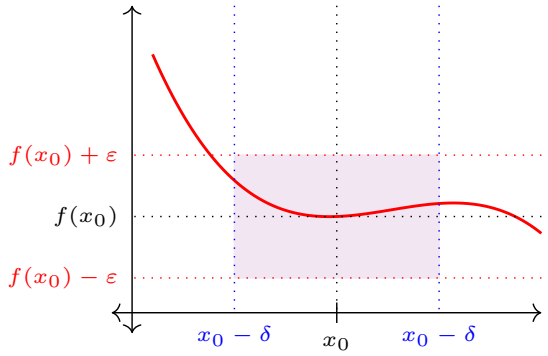


Figure 3: An illustration of continuity. All inputs x within δ of x_0 yield outputs $f(x)$ within at least ε of $f(x_0)$.

and that $\ell^p(\mathbb{Z})$ has the norm:

$$\|x_n^p\| = \left(\sum_{n=-\infty}^{\infty} |x_n|^p\right)^{\frac{1}{p}} = \left(\sum_{n=0}^{\infty} |x_{-n}|^p + \sum_{n=1}^{\infty} |x_n|^p\right)^{\frac{1}{p}}.$$

Exercise 1.5.1. *Think about how the above works when $p = \infty$.*

But all of this is just a silly example to motivate the L^p-norms and illustrate their connection to the ℓ^p-norms. So let us start to introduce those in a more serious way.

1.5.2. **Defining the L^p-norms.** For any "nice" set $\Omega \subset \mathbb{R}^n$, define the spaces

$$C(\Omega, \mathbb{R}) = \{f : \Omega \to \mathbb{R} : f \text{ is continuous}\} \qquad \text{and}$$

$C_{\text{pw}}(\Omega, \mathbb{R}) = \{f : \Omega \to \mathbb{R} : f \text{ is piecewise continuous}\}$

We almost always just write these as $C(\Omega)$ and $C_{\text{pw}}(\Omega)$. The most often used sets will be $\Omega = \mathbb{R}^n$ or $\Omega = [a_1, b_1] \times \cdots \times [a_n, b_n]$. From here, though, for simplicity, we focus on the case $n = 1$ and $\Omega = [a, b]$ for some $a < b$. Let us quick recall the definition of continuity:

Definition 1.5.1. *Let $\Omega \subset \mathbb{R}$. A function $f : \Omega \to \mathbb{R}$ is continuous at x_0 if, for every $\varepsilon > 0$, there is $\delta_{\varepsilon,x_0} > 0$ such that*

$$|f(x) - f(x_0)| < \varepsilon$$

whenever $|x - x_0| < \delta_{\varepsilon,x_0}$. See Figure 3.

We call $f : \Omega \to \mathbb{R}$ piecewise continuous if there is $n \in \mathbb{N} \cup \{0\}$ and points $x_1, \ldots, x_n \in \Omega$ such that f is continuous except at $x_1, \ldots, x_n$.

Exercise 1.5.2. *To get your bearings with continuity, try the following two exercises:*

(i) Show that $\mathbb{1}_{[0,1]} : \mathbb{R} \to \mathbb{R}$ is not continuous. This is the function such that $\mathbb{1}_{[0,1]}(x) = 1$ when $x \in [0, 1]$ and $\mathbb{1}_{[0,1]}(x) = 0$ when $x \notin [0, 1]$.

(ii) Show that $f(x) = \pi x$ is continuous.

We can put a norm on $C_{\text{pw}}([a,b])$ along the lines of the above: for $p \in [1,\infty)$ and any $f \in C_{\text{pw}}([a,b])$, let

$$\|f\|_{L^p([a,b])} = \left(\int_a^b |f(x)|^p dx \right)^{\frac{1}{p}}, \tag{1.5.1}$$

and for $p = \infty$, let

$$\|f\|_{L^\infty([a,b])} = \sup_{[a,b]} |f(x)|. \tag{1.5.2}$$

We make two important notes:

- We usually simply denote the norm by $\|\cdot\|_p$ when no confusion will arise.
- The L^∞-norm actually involves a *maximum* of $|f|$ due to the extreme value theorem. There are two reasons that we use sup in its definition. First, if we change the domain to (a,b), a maximum is no longer guaranteed. Second, in the future, we will allow $L^\infty([a,b])$ to define a normed linear space that includes "all" bounded functions, including some discontinuous ones, which may not have a maximum (but will have a supremum).

In the future, it will be useful to discuss the "L^p-spaces." We do not, however, have the theory to properly define these. Roughly, one would like to just say

$$L^p(\Omega) = \{f : \Omega \to \mathbb{R} : \|f\|_p < \infty\}. \tag{1.5.3}$$

This definition will not work.

To illustrate the issue with (1.5.3), consider the following function:

$$\mathbb{1}_{\mathbb{Q}\cap[0,1]}(x) = \begin{cases} 1 & \text{if } x \in \mathbb{Q} \cap [0,1], \\ 0 & \text{if } x \notin \mathbb{Q} \cap [0,1], \end{cases}$$

where $\mathbb{Q}$ is the set of all rational numbers. Since every interval (a,b) contains a rational number[3], one can show that the upper Riemann sum for

$$\int_{\mathbb{R}} \mathbb{1}_{\mathbb{Q}}(x) dx$$

is always 1, but the lower Riemann sum is always 0. Hence, it is not clear what value to assign $\|\mathbb{1}_{\mathbb{Q}}\|_1$. Is $\mathbb{1}_{\mathbb{Q}} \in L^1(\mathbb{R})$ or not!?!?

While in the future we will be able to resolve this mini-paradox, for the moment we will work with two different definitions, depending on what is most convenient for us. One, straightforward way to resolve it, is to take

$$L^p_{\text{prel}}(\Omega) = \{f \in C_{\text{pw}}(\Omega) : \|f\|_{L^p(\Omega)} < \infty\}.$$

Here "prel" is short for "preliminary."

Sometimes this is too restrictive to include interesting and important examples; for example, piecewise continuous functions. Another, somewhat vague approach is to use function f such that[4]

$$f \text{ is the } L^p \text{ limit of some sequence } f_n \in C(\Omega) \text{ such that } \sup_n \|f_n\|_p < \infty.$$

[3] We will show this in Section 3.1 but you probably do not find it surprising!

[4] This requires that $p < \infty$, unfortunately. However, the above "definition" for L^p spaces is good enough to then obtain L^∞ as the "limit" of the L^p spaces: if $a < b$, then $f \in L^\infty([a,b])$ if $f \in L^p([a,b])$ for all $p \in [1,\infty)$ and $\|f\|_\infty := \lim_{p\to\infty} \|f\|_{L^p([a,b])} < \infty$. This is, of course, an insane way to go about things. So simply ignore this footnote unless you find it compelling.

The "vagueness" referred to above is related to what we mean by "is the limit of," but let us ignore that for the moment.

In this way, we feel comfortable saying that $x^{-1/2} \in L^1([0,1])$. Indeed, even though this function is not continuous, it is clearly approximated in L^1 by the sequence

$$f_n(x) = \begin{cases} \frac{1}{\sqrt{x}} & \text{if } x > \frac{1}{n}, \\ x\sqrt{n} & \text{if } x \in [0, 1/n]. \end{cases}$$

Notice that for any $\varepsilon > 0$, for $n, m > \varepsilon^{-2/3}$, f_n and f_m agree everywhere except on $[0, \max\{1/n, 1/m\}]$. Hence,

$$\|f_n - f_m\|_1 \leq \int_0^{1/n} f_n(x)dx + \int_0^{1/m} f_m(x)dx = \frac{1}{2n^{3/2}} + \frac{1}{2m^{3/2}} \leq \varepsilon.$$

1.5.3. **The L^p-norms are norms.** The L^∞ case is the easiest, so we leave it as an exercise.

Exercise 1.5.3. *The L^∞-norm is a norm.*

We now check the three conditions of being a norm. Throughout the arguments, fix $p \in [1, \infty)$. Note the similarities to how the

Positive definiteness: As usual, it is obvious that $\|f\|_p \geq 0$. Next we check that $\|f\|_p = 0$ if and only if f is the zero function. Clearly if f is the zero function then $\|f\|_p = 0$. Hence, we need only check the other direction.

Suppose that f is not the zero function. Then there is some $x_0 \in [a, b]$ such that $|f(x_0)| > 0$. Since f is continuous, then there is $\delta > 0$ such that

$$|f(x_0) - f(x)| \leq \frac{|f(x_0)|}{2} \qquad \text{for all } |x - x_0| < \delta.$$

Hence, there is an interval $I \subset [a, b]$ of width at least

$$\bar{\delta} = \min\{b - a, \delta\}$$

such that, for $x \in I$,

$$|f(x)| = |f(x_0) + (f(x) - f(x_0))| \geq |f(x_0)| - |f(x) - f(x_0)| \geq |f(x_0)| - \frac{|f(x_0)|}{2} = \frac{|f(x_0)|}{2}.$$

It follows that

$$\|f\|_p^p = \int_a^b |f(x)|^p dx \geq \int_I |f(x)|^p dx \geq \int_I \frac{|f(x_0)|^p}{2^p} dx \geq \frac{|f(x_0)|^p}{2^p} \bar{\delta} > 0.$$

This concludes the proof of positive definiteness.

Homogeneity: This is obvious and, thus, omitted.

Triangle inequality: As for ℓ^p, this goes by the name **Minkowski's inequality**, which we state here.

Theorem 1.5.2 (Minkowski's inequality)**.** *Let $f, g \in C_{\text{pw}}(\Omega)$ and $p \in [1, \infty]$. Then*

$$\|f + g\|_p \leq \|f\|_p + \|g\|_p.$$

Again, as for the ℓ^p-norms, this follows from Hölder's inequality:

Theorem 1.5.3 (Hölder's inequality)**.** *Let $f, g \in C_{\text{pw}}(\Omega)$, $p \in [1, \infty]$, and q be the conjugate exponent of p. Then*

$$\int_\Omega f(x)g(x)dx \leq \|f\|_p \|g\|_q.$$

Equality holds if and only if there is $c > 0$ such that $f(x) = c\,\text{sign}(g(x))|g(x)|^{q-1}$ when $p > 1$ or $g(x) = c\,\text{sign}(f(x))$ when $p = 1$.

As we will see in future lectures, the assumption of continuity is not needed to show Minkowski's and Hölder's inequalities. Really all that is needed is that the functions f and g are *integrable*, which ends up being a thorny issue (and so is postponed). The convenience of the assumption of continuity is that we are guaranteed that f and g are (locally) integrable.

Proof of Hölder's inequality.

Exercise 1.5.4. *Prove this in the case $p = 1$ or $p = \infty$.*

The proof is exactly as in the ℓ^p case. By the exercise above, we can assume that $p, q \in (1, \infty)$. If either $\|f\|_p$ or $\|g\|_p$ is infinite, there is nothing to prove. Similarly if either is zero, there is nothing to prove. Thus, assume they are both finite and positive.

Fix $\varepsilon > 0$ to be chosen. Then, by Young's inequality (Lemma 1.4.5):

$$\begin{aligned}\int_\Omega f(x)g(x)dx &\leq \int_\Omega \frac{|f(x)|}{\varepsilon}(\varepsilon|g(x)|)dx \leq \int_\Omega \left(\frac{|f(x)|^p}{p\varepsilon^p} + \frac{\varepsilon^q|g(x)|^q}{q}\right)dx \\ &= \frac{\|f\|_p^p}{p\varepsilon^p} + \frac{\varepsilon^q\|g\|_q^q}{q}.\end{aligned} \tag{1.5.4}$$

Let

$$\varepsilon = \frac{\|f\|_p^{\frac{p-1}{p}}}{\|g\|_q^{\frac{1}{p}}} = \frac{\|f\|_p^{\frac{1}{q}}}{\|g\|_q^{\frac{q-1}{q}}}.$$

Hence, (1.5.4) becomes

$$\int_\Omega f(x)g(x)dx \leq \frac{\|f\|_p\|g\|_q}{p} + \frac{\|f\|_p\|g\|_q}{q} = \|f\|_p\|g\|_q.$$

□

Proof of Minkowski's inequality.

Exercise 1.5.5. *Prove this in the case $p = 1$ or $p = \infty$.*

Again, the proof is exactly as in the ℓ^p case. By the exercise above, we need only consider the case $p \in (1, \infty)$. Additionally, if $\|f\|_p$ or $\|g\|_p$ is infinte, then the proof is obvious. We may, thus, assume that both norms are finite. Finally, if $\|f + g\|_p = 0$, then the proof is obvious, so we may assume that $\|f + g\|_p > 0$.

Let

$$h(x) = \text{sign}(f(x) + g(x))|f(x) + g(x)|^{p-1}.$$

Then, applying Hölder's inequality,

$$\begin{aligned}\int_\Omega |f(x)+g(x)|^p dx &= \int_\Omega (f(x)+g(x))h(x)dx = \int_\Omega f(x)h(x)dx + \int_\Omega g(x)h(x)dx \\ &\leq \|f\|_p\|h\|_q + \|g\|_p\|h\|_q = (\|f\|_p + \|g\|_p)\,\|h\|_q.\end{aligned} \tag{1.5.5}$$

where $q \in (1,\infty)$ is the conjugate exponent of p.

On the other hand, a direct computation yields

$$\|h\|_q^q = \int |f(x)+g(x)|^{q(p-1)}dx = \int |f(x)+g(x)|^p dx = \|f+g\|_p^p, \tag{1.5.6}$$

The second equality above holds because, due to being conjugate exponents, p and q satisfy

$$p(q-1) = p.$$

Since, also, $p/q = p-1$, then (1.5.6) implies that

$$\|h\|_q = \|f+g\|_p^{p-1}.$$

Plugging this into (1.5.5), implies that

$$\|f+g\|_p^p = \int_\Omega |f(x)+g(x)|^p dx \leq (\|f\|_p + \|g\|_p)\,\|h\|_q = (\|f\|_p + \|g\|_p)\,\|f+g\|_p^{p-1}.$$

Since $\|f+g\|_p > 0$, by assumption, we may divide both sides by $\|f+g\|_p^{p-1}$ to conclude the proof. □

Just as before, we can also show that using "duality," we can characterize the L^p-norm in an alternate way:

Exercise 1.5.6. *Fix $p \in [1,\infty]$ and let q be its conjugate exponent. Fix any $a,b \in [-\infty,+\infty]$ with $a<b$, and let $f \in C_{\mathrm{pw}}([a,b])$. Show that the the L^p-norm of f can be obtained by "testing" with all function of L^q-norm bounded by one:*

$$\|f\|_p = \sup_{\substack{g\in C_{\mathrm{pw}}([a,b]),\\ \|g\|_q\leq 1}} \int_a^b f(x)g(x)dx = \sup_{\substack{g\in C_{\mathrm{pw}}([a,b]),\\ \|g\|_q= 1}} \int_a^b f(x)g(x)dx.$$

(Sometimes we simply use $\langle f,g\rangle$ to denote the "dot product" $\int_a^b f(x)g(x)dx$.)

Let us use these inequalities to obtain a few simple bounds.

Example 1.5.4. *(i) Using Hölder's inequality with $p=q=2$, we find*

$$\int_0^1 \sqrt{x(1-x)}dx \leq \left(\int_0^1 x dx\right)^{1/2}\left(\int_0^1 (1-x)dx\right)^{1/2} = \frac{1}{2}. \tag{1.5.7}$$

Another estimate you can get is

$$\begin{aligned}\int_0^1 \sqrt{x(1-x)}dx &\leq \left(\int_0^1 x(1-x)dx\right)^{1/2}\left(\int_0^1 1\,dx\right)^{1/2} \\ &= \left(\frac{1}{2}-\frac{1}{3}\right)^{1/2}\cdot 1 = \frac{1}{\sqrt{6}} \approx .4082...\end{aligned} \tag{1.5.8}$$

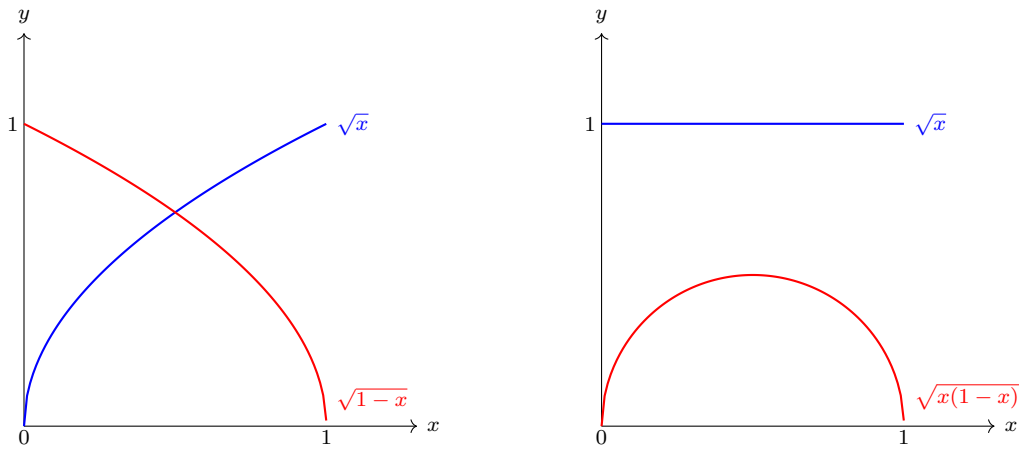


Figure 4: The functions used in Example 1.5.4.(i).

(The actual value is $\pi/8 \approx .39276...$). Why is the second estimate better than the first? Recall that Hölder's inequality (for $p = q = 2$) is sharp when f and g are the same up to a constant. In (1.5.7), the f and g are $\sqrt{x}$ and $\sqrt{1-x}$, respectively, which are not particularly similar functions. On the other hand, in (1.5.8), the f and g are $\sqrt{x(1-x)}$ and 1, respectively, which are both mostly flat functions. Hence, we are closer to the "sharp" (equality) case.

To put a finer point on this, we want to think of $\langle f, g\rangle$ as an inner product (dot product), which will be maximized up to $\|f\|_{L^p}\|g\|_{L^2}$ when f and g are most aligned; that is, they are the same up to multiplication by a constant. Roughly, for functions this means they have the same wiggles in the same places (even if one is much larger than the other). In Figure 4 it is heuristically clear that f and g are more "similar" in the second case than the first.

(ii) Let us find a lower bound of $\int_0^\pi (\sin(x)/x)^3 dx$. Notice that the conjugate exponent of $p = 3$ is $q = 3/2$. Then:

$$2 = \int_0^\pi \sin x dx = \int_0^\pi \left(\frac{\sin x}{x}\right) x dx \leq \left(\int_0^\pi \left(\frac{\sin x}{x}\right)^3 dx\right)^{\frac{1}{3}} \left(\int_0^\pi x^{3/2} dx\right)^{\frac{2}{3}}$$
$$= \left(\int_0^\pi \left(\frac{\sin x}{x}\right)^3 dx\right)^{\frac{1}{3}} \left(\frac{2}{5}\pi^{\frac{5}{2}}\right)^{\frac{2}{3}} = \left(\int_0^\pi \left(\frac{\sin x}{x}\right)^3 dx\right)^{\frac{1}{3}} \left(\frac{2}{5}\right)^{\frac{2}{3}} \pi^{\frac{5}{3}}.$$

Hence,

$$\frac{50}{\pi^5} = \left(2\pi^{\frac{-5}{3}} \left(\frac{2}{5}\right)^{\frac{-2}{3}}\right)^3 \leq \int_0^\pi \left(\frac{\sin x}{x}\right)^3 dx.$$

(iii) Let us see "how continuous" a function f can be if we know something about the L^p-norm of its derivative f'. Let $p > 1$ and q be its conjugate exponent. Then:

$$|f(x) - f(y)| = \left|\int_x^y f'(z)dz\right| \leq \left(\int_x^y 1^q dz\right)^{\frac{1}{q}} \left(\int_x^y |f'(z)|^p dz\right)^{\frac{1}{p}}$$
$$\leq \left(\int_x^y 1^q dz\right)^{\frac{1}{q}} \left(\int_a^b |f'(z)|^p dz\right)^{\frac{1}{p}} \leq \left(\int_x^y 1^q dz\right)^{\frac{1}{q}} \|f'\|_p = |x-y|^{1/q}\|f'\|_p$$

We point out that, as $p \to +\infty$, $1/q \to 1$, so that this gets "more continuous" (that is, the modulus of continuity[5] gets smaller).

Note: This is a basic version of an extremely useful class of inequalities called "Sobolev inequalities." What we have deduced from the statement above is that, if $f, f' \in L^p([a,b])$ then $f \in C^{1/q}([a,b])$ and that

$$[f]_{C^{1/q}([a,b])} \leq \|f'\|_p,$$

where, for any $\alpha \in (0,1)$, the α-Hölder semi-norm is defined by

$$[f]_{C^\alpha([a,b])} := \sup_{x \neq y} \frac{|f(x) - f(y)|}{|x-y|^\alpha}$$

and α-Hölder space is defined by

$$C^\alpha([a,b]) = \{f \in C([a,b]) : [f]_{C^\alpha} < +\infty\}.$$

These Hölder spaces are also extremely important to the theory of partial differential equations (PDE), but let us not discuss them further.

Exercise 1.5.7. *Show that, if $f' \in L^2(\mathbb{R})$ then $f \in C^{1/2}(\mathbb{R})$.*

Exercise 1.5.8. *Suppose that $u, u' \in C(\mathbb{R})$ satisfies*

$$\lim_{x \to -\infty} u(x) = 1, \quad \lim_{x \to +\infty} u(x) = 0, \quad \text{and} \quad \|u\|_\infty = 1.$$

Suppose that $f \in C([0,1])$ is nonnegative function with the property that there is $\alpha > 0$ and $z_1, z_2 \in (0,1)$ with $z_1 < z_2$ such that $f(z) \geq \alpha$ for all $z \in [z_1, z_2]$. Find an explicit lower bound on

$$\int_{\mathbb{R}} f(u(x))dx.$$

This lower bound plays an important role in understanding PDE models for flame propagation.

When we are working on a finite interval $[a,b]$ with $-\infty < a < b < \infty$, we have a nice inclusion amongst L^p spaces. To get a handle on why this should be true, think about the inclusion $L^\infty([a,b]) \subset L^1([a,b])$. Indeed,

$$\|f\|_{L^1([a,b])} = \int_a^b |f(x)|dx \leq \int_a^b \max_{x \in [a,b]} |f(x)|dx = \int_a^b \|f\|_{L^\infty([a,b])}dx = |a-b|\|f\|_{L^\infty([a,b])}.$$

On finite domains, the larger the p, the more the L^p-norm is about boundedness. In general we get the following inequality directly from Hölder's inequality:

Corollary 1.5.5. *Given $a, b \in \mathbb{R}$ with $a < b$ and $f \in L^r([a,b])$ with $r > s \geq 1$, then $f \in L^s([a,b])$ and*

$$\|f\|_s \leq |b-a|^{\frac{1}{s} - \frac{1}{r}} \|f\|_r. \tag{1.5.9}$$

Proof. Let $p = r/s$ and $q = \frac{r}{r-s}$. It is easy to check that p and q are conjugate exponents. An application of Hölder's inequality yields

$$\|f\|_s^s = \int_a^b |f(x)|^s \cdot 1 \, dx \leq \left(\int_a^b (|f(x)|^s)^{\frac{r}{s}} \, dx \right)^{\frac{s}{r}} \left(\int_a^b 1^{\frac{r}{r-s}} \, dx \right)^{\frac{r-s}{r}} \leq \|f\|_r^s |b-a|^{1-\frac{s}{r}}.$$

The proof is then finished by taking the $1/s$ power of each side. □

[5]The modulus of continuity is, roughly, the largest δ that can be taken for each ε in the definition of continuity.

Before continuing, we make a few notes about Corollary 1.5.5:

- The quantitative statement (1.5.9) encodes the qualitative statement $L^r([a,b]) \subset L^s([a,b])$. This is an example of an "embedding" theorem. We say L^r is embedded in L^s.
- This is the *opposite* embedding from the ℓ^p spaces, which gives $\ell^s \subset \ell^r$ when $s < r$. One can ask if $L^s([a,b]) \subset L^s([a,b])$, but this is, unfortunately, not true. Indeed: fix $\alpha > 0$ such that $\alpha r < 1$ and $\alpha s > 1$. Then, letting

  $$f(x) = (x-a)^{-\alpha},$$

  it is easy to see that $f \in L^r([a,b]) \setminus L^s([a,b])$.
- This proof strategy will clearly not hold on an infinite domain $(-\infty, b)$, (a, ∞), or $(-\infty, \infty)$ since this corresponds to the limit $|b-a| \to \infty$. Actually, the inclusion is not even true!

  Exercise 1.5.9. *Under the conditions of Corollary 1.5.5 but with either $a = -\infty$ or $b = \infty$, find $f \in L^s \setminus L^p$.*

Important aside: scaling.

There is an important "scaling" property in Theorem 1.5.3 and Corollary 1.5.5. Actually, this helps inform the necessity of having conjugate exponents. Let us consider Theorem 1.5.3 on $\mathbb{R}$ at first. Fix any $f \in L^p(\mathbb{R})$ and $g \in L^q(\mathbb{R})$ and let $\lambda > 0$. Define

$$f_\lambda(x) = \lambda^{\frac{1}{p}} f(\lambda x) \quad \text{and} \quad g_\lambda(x) = \lambda^{\frac{1}{q}} g(\lambda x).$$

An easy computation (u-substitution!) shows that

$$\|f_\lambda\|_p = \|f\|_p \quad \text{and} \quad \|g_\lambda\|_q = \|g\|_q.$$

Suppose that p and q **were not** conjugate exponents, so that

$$\frac{1}{p} + \frac{1}{q} \neq 1,$$

but that, for whatever reason, we happened to know that Theorem 1.5.3 held regardless for this choice of p and q. Then:

$$\lambda^{\frac{1}{p}+\frac{1}{q}-1} \int f(x)g(x)dx = \int f_\lambda(x) g_\lambda(x) dx \leq \|f_\lambda\|_p \|g_\lambda\|_q = \|f\|_p \|g\|_q.$$

As long as $\langle f, g \rangle = \int f(x)g(x)dx > 0$ (which is certainly true for some choice of f and g), we can derive a contradiction. Indeed,

- if $\frac{1}{p} + \frac{1}{q} > 1$, then we can take $\lambda \to \infty$;
- if $\frac{1}{p} + \frac{1}{q} < 1$, then we can take $\lambda \to 0$.

In both cases, the left hand side tends to infinite, but the right hand side stays constant. This is a contradiction. Hence, the conjugate exponent condition is "forced" on us by scaling. **This is an important way to test if an inequality that you are trying to prove has any hope of working!**

Exercise 1.5.10. *Perform a scaling analysis to understand the $|a-b|^{\frac{1}{s}-\frac{1}{r}}$ factor in Corollary 1.5.5.*

Back to L^p spaces...

Sometimes it is useful to take intersections of normed linear spaces. There is a general process to doing this that we define here. Below we specialize to the L^p-spaces and explore some examples.

Definition 1.5.6. *Given two normed linear spaces $(X, \|\cdot\|_X)$ and $(Y, \|\cdot\|_Y)$ such that $X, Y \subset Z$ for some linear space Z, we define new space:*

$$(X \cap Y, \|\cdot\|_{X\cap Y}) \qquad \text{where } \|v\|_{X\cap Y} = \|v\|_X + \|v\|_Y.$$

It is straightforward to check that $\|v\|_{X\cap Y}$ is a norm. Let us think

Exercise 1.5.11. *Does the definition $\|v\|_{X\cap Y} = \min\{\|v\|_X, \|v\|_Y\}$ work as well (i.e., is this a norm)?*

Example 1.5.7. *Consider the spaces $L^1([0,\infty))$ and $L^7([0,\infty))$. Call $X = L^1 \cap L^7$ and let us think about X for a bit. Let $p \in [1,7]$ and we claim that:*

(i) $L^p \not\subset X$ and

(ii) $X \subset L^p$ with the stronger claim that there is $C_p > 0$, depending only on p, such that if $f \in X$ then

$$\|f\|_p \leq C_p \|f\|_X. \tag{1.5.10}$$

Let us first look at (i). If $p < 7$, then

$$f(x) = x^{-\frac{2}{p+7}} e^{-x},$$

is clearly in L^p but is not in L^7 and, hence, X. On the other hand, when $p = 7$,

$$f(x) = \frac{1}{1 + x^{\frac{2}{7}}},$$

is an element of $L^p = L^7$, but is not an element of L^1 and, hence, X.

Now we investigate (ii). Our main tools will be Hölder's and Young's inequalities. We begin by applying Hölder's inequality with conjugate exponents r, s satisfying:

$$\frac{1}{r} + \frac{1}{s} = 1 \qquad \text{and} \qquad \frac{1}{r} + \frac{7}{s} = p.$$

In fact, we have $r = \frac{6}{7-p}$ and $s = \frac{6}{p-1}$. Then:

$$\begin{aligned}\|f\|_p^p = \int_0^\infty |f(x)|^p dx &= \int_0^\infty |f(x)|^{\frac{1}{r}} |f(x)|^{\frac{7}{s}} dx \leq \||f|^{1/r}\|_r \||f|^{7/s}\|_s \\ &= \left(\int_0^\infty |f(x)| dx\right)^{1/r} \left(\left(\int_0^\infty |f(x)|^7 dx\right)^{\frac{1}{7}}\right)^{\frac{7}{s}} = \|f\|_1^{\frac{1}{r}} \|f\|_7^{\frac{7}{s}} = \|f\|_1^{\frac{7-p}{6}} \|f\|_7^{\frac{7(p-1)}{6}}.\end{aligned}$$

We used the explicit forms of r and s in the last inequality. Rearranging this, we find

$$\|f\|_p \leq \|f\|_1^{\frac{7-p}{6p}} \|f\|_7^{\frac{7(p-1)}{6p}},$$

which is close to what we want. In fact, it implies the inclusion $X \subset L^p$, but it is not precisely of the form (1.5.10)*. As usual, when we have a product and we want a sum (or vice versa), we appeal to Young's inequality. Notice that miraculously*[6] $6p/(7-p)$ *and* $6p/7(p-1)$ *are conjugate exponents. Let $\varepsilon > 0$ be a constant to be chosen and then Young's inequality gives*

$$\|f\|_p \leq \frac{\varepsilon^{\frac{6p}{7-p}}\|f\|_1}{\frac{6p}{7-p}} + \frac{\|f\|_7}{\frac{6p}{7(p-1)}\varepsilon^{\frac{6p}{7(p-1)}}}. \tag{1.5.11}$$

At this point we are finished by taking $\varepsilon = 1$ and letting

$$C_p = 2\max\left\{\frac{7-p}{6p}, \frac{7(p-1)}{6p}\right\}.$$

Exercise 1.5.12. *Choose the best ε to find the smallest C_p that follows from* (1.5.11)*. Is this estimate sharp? That is, for each p, can you find f such that* (1.5.11)*, with the C_p you computed, is actually and equality?*

Exercise 1.5.13. *Suppose we fix $p \in [1,7]$ and any f that is an element of L^p* and *X. Will it be true that the converse of* (1.5.10) *holds? That is, is there a constant $C_p > 0$ such that*

$$\|f\|_X \leq C_p\|f\|_p?$$

If the answer is yes, prove it. If the answer is no, establish this by finding a sequence $f_n \in X$ such that $\|f_n\|_X \geq n\|f_n\|_p$ for each n.

2. Continuity and dual spaces

2.1. **The ε-δ definition of continuity and sequential continuity.** We begin by giving the definition of continuity in a metric space (X, d).

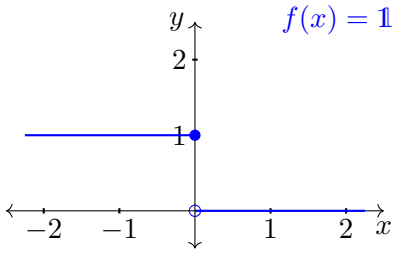
Definition 2.1.1. *Let (X, d_X) and (Y, d_Y) be metric spaces. Fix $x_0 \in X$. We say that a function $f : X \to Y$ is* continuous at x_0 *if, for every $\varepsilon > 0$ there is $\delta > 0$ such that*

$$d_Y(f(x), f(x_0)) < \varepsilon \qquad \text{whenever } d_X(x, x_0) < \delta.$$

We say that f is continuous *if it is continuous at every point $x_0 \in X$.*

Exercise 2.1.1. *Take the metric spaces X and Y above to be $\mathbb{R}$ with the metric induced by the absolute value (i.e., the usual Euclidean metric) and think about how this definition matches up with the $\varepsilon - \delta$ definition for continuity that you previously learned.*

Example 2.1.2. *Let us illustrate this definition by considering the very simple example of $f : \mathbb{R} \to \mathbb{R}$, where $f(x) = \mathbb{1}_{(-\infty,0]}$.*



[6] There are no miracles in math, this is simple due to the importance of the conjugate exponents in the computations above.

While we can all agree that this is not continuous at $x = 0$, we need to prove it with our new definition. Since the jump at $x = 0$ has size 1, we choose $\varepsilon = 1/2$, which is smaller than 1. Then, for any $\delta > 0$, we have that

$$|\delta/2 - 0| < \delta \qquad but \qquad |f(\delta/2) - f(0)| = |0 - 1| = 1 > 1/2 = \varepsilon.$$

Hence, f is not continuous.

There is, actually, another "version" of continuity using sequences that we define now.

Definition 2.1.3. *Suppose that $(x_n)_n$ is a sequence of points in X and $x_\infty \in X$. We say that*

$$\lim_{n\to\infty} x_n = x_\infty,$$

or, equivalently, that x_n converges to x_∞ as $n \to \infty$ *if, for every $\varepsilon > 0$, there is N_ε such that*

$$d(x_n, x_\infty) < \varepsilon \qquad whenever\ n \geq N_\varepsilon.$$

Example 2.1.4. *Let us see this definition in action. Consider the sequence of functions f_1, f_2, f_3, ... defined by*

$$f_n(x) = n^{1/4}\mathbb{1}_{[0,1/n]}.$$

In a "normal" sense, f_n does not converge to any function f since $f_n(0) = n^{1/4} \to \infty$ as $n \to \infty$. But what about in L^p spaces? First let us check L^2:

$$\|f_n - 0\|_2 = \left(\int (f_n(x) - 0)^2 dx\right)^{1/2} = \left(\int f_n(x)^2 dx\right)^{1/2} = \left(\int_0^{1/n} n^{1/2} dx\right)^{1/2} = \frac{1}{n^{1/4}}. \quad (2.1.1)$$

For any $\varepsilon > 0$, let N_ε be the smallest integer that is greater than $1/\varepsilon^4$. Then, $N_\varepsilon^{-1/4} < \varepsilon$. If $n \geq N_\varepsilon$, we have, from (2.1.1)*,*

$$\|f_n - 0\|_2 \leq n^{-1/4} \leq N_\varepsilon^{-1/4} < \varepsilon.$$

Thus, $f_n \to 0$ as $n \to \infty$ in L^2.

Exercise 2.1.2. *Show that $f_n \to 0$ as $n \to \infty$ in L^p for any $p \in [1, 4)$.*

Exercise 2.1.3. *Show that if $x_n \to x_\infty$ and y is another point in the metric space (X, d) then $d(x_n, y) \to d(x_\infty, y)$. Use this to show that, in the previous example, $f_n \not\to f$ for any $f \in L^p$ when $p \geq 4$.*

Note that the previous examples and exercises show that the *metric* you work with matters a lot in determining the convergence of the sequence.

Definition 2.1.5. *Let (X, d_X) and (Y, d_Y) be metric spaces. Fix $x_0 \in X$. We say that a function $f : X \to Y$ is* sequentially continuous at x_0 *if, for every sequence $(x_n)_n$,*

$$\lim_{n\to\infty} f(x_n) = f(x_\infty) \qquad whenever \lim_{n\to\infty} x_n = x_\infty.$$

We say that f is sequentially continuous *if it is continuous at every point $x_\infty \in X$.*

We see below that this is equivalent to continuity as defined above *because we are in a metric space.* Note that there is a weaker notion of spaces that allow "convergence" called topological spaces in which sequential continuity is not equivalent to continuity.

Theorem 2.1.6. *Suppose that (X, d) and (Y, ρ) are two metric spaces, $f : X \to Y$, and $x_\infty \in X$. Then f is continuous at x_∞ if and only if f is sequentially continuous at x_∞.*

Before beginning the proof we note the importance of this theorem. Often, it is easier or more natural to work with sequences instead of shuffling around epsilons and delta. This theorem tells us that that is OK.

Proof that continuity implies sequential continuity. Fix any sequence $(x_n)_n$ such that

$$\lim_{n\to\infty} x_n = x_\infty. \tag{2.1.2}$$

Fix any $\varepsilon > 0$. Since f is continuous at x_∞, there is $\delta > 0$ such that

$$d(f(x_\infty), f(x)) < \varepsilon \qquad \text{whenever } d(x_\infty, x) < \delta. \tag{2.1.3}$$

Due to (2.1.2), there is N_δ such that, whenever $n \geq N_\delta$, we have

$$d(x_\infty, x_n) < \delta.$$

Putting this together with (2.1.3), we see that, whenever $n \geq N_\delta$,

$$d(f(x_\infty), d(f(x))) < \varepsilon.$$

This is precisely what we need to show in order to conclude that

$$\lim_{n\to\infty} f(x_n) = f(x_\infty).$$

Hence, the proof is complete. □

Proof that sequential continuity implies continuity. Before beginning, we note that this proof is difficult (impossible?) to do directly, and it is much easier to give a proof of the contrapositive. In other words, we show that if f is not continuous at x_∞ then f is not sequentially continuous at x_∞. This is logically equivalent to what we are trying to prove.

Suppose that f is not continuous at x_∞. Then, for every n, there is $\varepsilon > 0$ and x_n such that

$$d(f(x_\infty), f(x_n)) > \varepsilon \qquad \text{and} \qquad d(x_\infty, x_n) < \frac{1}{n}.$$

Here we are just taking $\delta = 1/n$ for each n and taking the negation of the definition of continuity.

It is clearly true that

$$\lim_{n\to\infty} x_n = x_\infty.$$

On the other hand, it is also clearly true that

$$\lim_{n\to\infty} f(x_n) \neq f(x_\infty).$$

Hence, f is not sequentially continuous at x_∞. □

Example 2.1.7. *1. Fix a constant $c \in \mathbb{R}$ and take the function $g : \mathbb{R} \to \mathbb{R}$ defined by $g(x) = cx$ for all x. We know from our Calculus that, for any convergent sequence,*

$$\lim_{n\to\infty} cx_n = c \lim_{n\to\infty} x_n.$$

Hence, g is sequentially continuous. It follows from Theorem 2.1.6 that g is continuous as well.

2. *Let $f : \mathbb{R} \to \mathbb{R}$ be defined by $f(x) = e^x$. Fix any $\theta \in \mathbb{R}$ and we will show that f is continuous at θ. Notice that*

$$f(\theta) - f(x) = e^\theta - e^y = e^\theta \left(1 - e^{y-\theta}\right).$$

Fix any $\varepsilon > 0$ and let

$$\delta = \min \left\{ \log(1 + \varepsilon e^{-\theta}), \log\left(\frac{1}{1 - \varepsilon e^{-\theta}}\right) \right\}. \tag{2.1.4}$$

Consider any $y \in \mathbb{R}$ such that

$$|y - \theta| < \delta.$$

If $y \geq \theta$, then

$$0 \leq y - \theta < \delta \leq \log(1 + \varepsilon e^{-\theta}).$$

Hence

$$|f(\theta) - f(x)| = e^\theta \left(e^{y-\theta} - 1\right) < e^\theta \left(e^{\log(1+\varepsilon e^{-\theta})} - 1\right) = \varepsilon,$$

which is exactly what we want to show.

Exercise 2.1.4. *Check the case $y < \theta$ on your own! This is where the other alternative in* (2.1.4) *comes from.*

Notice here that δ depends on θ, the point at which we wish to show continuity.

3. *Define a function*[7] *$\mathcal{L} : L^1(\mathbb{R}_+) \to \mathbb{R}_+$ by*

$$\mathcal{L}(f) = \int_0^\infty \sin(x) f(x) dx.$$

Is this continuous at $f(x) = e^{-x}$? Fix any $\varepsilon > 0$ and let

$$\delta = \varepsilon.$$

Notice that, due to Hölder's inequality and the fact that $\| \sin \|_{L^\infty} = 1$, we have, for any $h \in L^1$,

$$|\mathcal{L}(h)| \leq \|h\|_{L^1}. \tag{2.1.5}$$

Fix any $g \in L^1$ such that $\|g - f\|_1 < \delta$. Notice that

$$\mathcal{L}(f) - \mathcal{L}(g) = \int_0^\infty \sin(x) f(x) dx - \int_0^\infty \sin(x) g(x) dx = \int_0^\infty \sin(x)(f(x) - g(x)) dx = \mathcal{L}(f - g).$$

In view of (2.1.5), *we have*

$$|\mathcal{L}(f) - \mathcal{L}(g)| \leq \|f - g\|_{L^1} < \delta = \varepsilon,$$

as desired. Thus $\mathcal{L}$ is continuous. Notice that δ does not depend on f, the point at which we want to show continuity.

[7] Recall that $\mathbb{R}_+ = [0, \infty)$

4. Define a function[8] $\mathcal{N} : L^1(\mathbb{R}_+) \to \mathbb{R}_+$ *by*

$$\mathcal{N}(f) = \left(\int_0^\infty \sin(x) f(x) dx\right)^2 = (\mathcal{L}(f))^2.$$

Is this continuous at $f(x) = \pi e^{-x}$*? Notice that* $\|f\|_1 = 2\pi$*. Fix any* $\varepsilon > 0$ *and let*

$$\delta = \min\left\{1, \varepsilon/2\|f\|_{L^1} + 1\right\}$$

Fix any $g \in L^1$ *such that* $\|g - f\|_1 < \delta$*. Notice that*

$$\mathcal{N}(f) - \mathcal{N}(g) = \left(\int_0^\infty \sin(x)(f(x) + g(x))dx\right)\left(\int_0^\infty \sin(x)(f(x) - g(x))dx\right) = \mathcal{L}(f+g)\mathcal{L}(f-g),$$

and that

$$\begin{aligned} \mathcal{L}(f+g) &\leq \|f+g\|_{L^1} = \|2f + (g-f)\|_{L^1} \\ &\leq 2\|f\|_{L^1} + \|g-f\|_{L^1} < 2\|f\|_{L^1} + \delta \leq 2\|f\|_{L^1} + 1. \end{aligned} \tag{2.1.6}$$

The first inequality is due to (2.1.5)*, the second inequality is due to the triangle inequality, and the third inequality is due to the fact that* $\delta \leq 1$*. Note that this is why I have the* 1 *in the definition of* δ *(indeed,* ε *could be anything, so if I just used* $\varepsilon/100$*, then* $4 + \varepsilon/100$ *could be* $10^{10^{10^{10^{\cdots}}}}$*). Additionally, getting the optimal* δ *here does* not *help – no matter how small* δ *is,* $\|f+g\|_{L^1}$ *will still, roughly, be* $2\|f\|_{L^1}$*. Hence, all of the "smallness" should come from the* $\mathcal{L}(f-g)$ *term, not the* $\mathcal{L}(f+g)$ *term.*

Using (2.1.6) *and also* (2.1.5) *again, we find*

$$|\mathcal{N}(f) - \mathcal{N}(g)| \leq (2\|f\|_{L^1} + 1)\|f-g\|_{L^1} < (2\|f\|_{L^1} + 1)\delta \leq \varepsilon.$$

Hence, $\mathcal{N}$ *is continuous.*

A procedural comment: our δ *worked out perfectly here. The reason is that, in practice, we leave* δ *unchosen until the last step and then go back and figure out what the correct* δ *should be.*

Definition 2.1.8. *A function* $f : (X, d) \to (Y, \rho)$ *is* uniformly continuous *if, for all* $\varepsilon > 0$*, there is* $\delta > 0$ *such that*

$$\rho(f(x_1), f(x_2)) < \varepsilon$$

whenever $d(x_1, x_2) < \delta$*.*

Notice that the same δ works for all x_1 and x_2. In the definition of continuity, the δ may depend on the base point x_0. Hence, uniform continuity is a stronger property than continuity: every uniformly continuous function is continuous, but not every continuous function is uniformly continuous.

To understand this distinction further, let us return to our our examples from Example 2.1.7 and see which are uniformly continuous as well.

[8] Recall that $\mathbb{R}_+ = [0, \infty)$

Example 2.1.9. *1. Fix a constant $c \in \mathbb{R}$ and take the function $g : \mathbb{R} \to \mathbb{R}$ defined by $g(x) = cx$ for all x. Fix any $\varepsilon > 0$ and let*

$$\delta = \frac{\varepsilon}{1 + |c|} > 0.$$

Notice that this depends only on f and ε. Suppose that

$$|x - y| < \delta.$$

Then

$$|g(x) - g(y)| = |c||x - y| \leq |c| \frac{\varepsilon}{1 + |c|} < \varepsilon.$$

Hence g is uniformly continuous.

2. Let $f : \mathbb{R} \to \mathbb{R}$ be defined by $f(x) = e^x$. In Example 2.1.7.(ii), our δ depended on the point θ at which we were establishing continuity. This suggests that f is not *uniformly continuous; however, it is not a proof. Let us show this. Fix any $\varepsilon > 0$. Suppose that f were uniformly continuous so that there is $\delta > 0$ such that*

$$|f(\theta) - f(\beta)| < \varepsilon \quad \text{whenever } |\theta - \beta| < \delta.$$

Then

$$\varepsilon > |f(n) - f(n + \delta/2)| = e^n \left(e^{\delta/2} - 1 \right).$$

Since $\exp\{\delta/2\} - 1 > 0$ and $e^n \to \infty$ as $n \to \infty$, this cannot possibly hold. Thus f is not uniformly continuous.

3. Define a function $\mathcal{N} : L^1(\mathbb{R}_+) \to \mathbb{R}_+$ by

$$\mathcal{N}(f) = \left(\int \sin(x) f(x) dx \right)^2.$$

As in the previous example, we show that $\mathcal{N}$ is not uniformly continuous and we show this by contradiction. Suppose that $\varepsilon > 0$ and $\delta > 0$ is such that

$$|\mathcal{N}(f) - \mathcal{N}(g)| < \varepsilon \quad \text{whenever } \|f - g\|_1 < \delta. \tag{2.1.7}$$

For any $A > 0$, let $f_A(x) = A \mathbb{1}_{[0,\pi/4]}$. Notice that $\|f_A\|_1 = A\pi/4$ and $f_A + f_B = f_{A+B}$ for any $A, B > 0$. Then,

$$\|f_{A+\delta/2} - f_A\|_1 = \|f_{\delta/2}\|_1 = \frac{\delta \pi}{8} < \delta,$$

but

$$\begin{aligned}
\mathcal{N}(f_{A+\delta/2}) - \mathcal{N}(f_A) &= \left(\int \sin(x)(f_{A+\delta/2} + f_A) dx \right) \left(\int \sin(x)(f_{A+\delta/2} - f_A) dx \right) \\
&= \left(\int \sin(x) f_{2A+\delta/2} dx \right) \left(\int \sin(x) f_{\delta/2} dx \right) \\
&= \left(\int_0^{\pi/4} \sin(x) \left(2A + \frac{\delta}{2} \right) dx \right) \left(\int_0^{\pi/4} \sin(x) \frac{\delta}{2} dx \right) \\
&\geq \left(2A + \frac{\delta}{2} \right) \frac{\delta}{2} \left(\int_0^{\pi/4} \sin(\pi/4)\, dx \right) \left(\int_0^{\pi/4} \sin(\pi/4)\, dx \right) \\
&= \left(2A + \frac{\delta}{2} \right) \frac{\delta}{2} \left(\frac{\pi}{4} \right)^2 \sin(\pi/4)^2.
\end{aligned}$$

Each step is a direct calculation until the first inequality, which uses that sin *is decreasing on* $[0, \pi/4]$. *The above can be made larger than* ε *by simply increasing* A *sufficiently. This violates* (2.1.7). *Hence* $\mathcal{N}$ *is not uniformly continuous.*

Exercise 2.1.5. *Define* $f : [0, \infty) \to [0, \infty)$ *by* $f(x) = \sqrt{x}$. *Show that* f *is uniformly continuous. Hint: you will need to choose* δ *to be* ε^2 *(possibly divided by a large constant).*

How does this work on the ℓ^p spaces?

Proposition 2.1.10. *For any* $p, q \in [1, \infty]$, *define*

$$\mathrm{id} : \ell^p \to \ell^q \qquad \text{by } \mathrm{id}(\bar{x}) = \bar{x}.$$

If $p \leq q$, *this is well-defined and uniformly continuous.*

Proof. We showed on Homework 1 that $\ell^p \subset \ell^q$. All of you showed, as a part of that exercise, that

$$\|\bar{x}\|_{\ell^q} \leq \|\bar{x}\|_{\ell^p}.$$

With this in hand, we show continuity. Suppose that $\varepsilon > 0$. Let $\delta = \varepsilon$ and suppose that

$$\|\bar{x} - \bar{y}\|_{\ell^p} < \delta = \varepsilon. \tag{2.1.8}$$

Then, by (2.1.8), we have

$$\|\mathrm{id}(\bar{x}) - \mathrm{id}(\bar{y})\|_{\ell^q} = \|\bar{x} - \bar{y}\|_{\ell^q} \leq \|\bar{x} - \bar{y}\|_{\ell^p} < \varepsilon.$$

This concludes the proof. □

Exercise 2.1.6. *If* $p > q$, *the above map is not well-defined; however, if we look at*

$$\mathrm{id} : \ell^p \cap \ell^q \to \ell^q \qquad \text{by } \mathrm{id}(\bar{x}) = \bar{x},$$

we can ask if this new map is continuous (it is clearly well-defined). Show that it is not.

Proposition 2.1.11. *For any* $p, q \in [1, \infty]$, *define*

$$f : \ell^p \to \ell^q \qquad \text{by } f(\bar{x}) = \left(|x_1|^{p/q}, |x_2|^{p/q}, |x_3|^{p/q}, \dots \right).$$

If $p > q$, *this is well-defined and continuous.*

Proof. To show that it is well-defined, we simply note that if $\bar{x} \in \ell^p$, then

$$\|f(\bar{x})\|_{\ell^q}^q = \sum_{i=1}^{\infty} \left| |x_i|^{p/q} \right|^q = \sum_{i=1}^{\infty} |x_i|^p = \|\bar{x}\|_{\ell^p}^p < \infty,$$

so that $f(\bar{x})$ is an element of ℓ^q.

We now look at continuity. Fix $\varepsilon > 0$ and $\bar{x} \in \ell^p$. We observe that, for any $a, b \geq 0$, there is a constant $C > 0$ such that

$$|a^{p/q} - b^{p/q}| \leq C(|a| + |b|)^{p/q - 1} |a - b|.$$

Exercise 2.1.7. *Prove this!*

Define

$$\delta = \min\left\{1, \frac{\varepsilon}{C^{\frac{1}{q}}(2\|\bar{x}\|_{\ell^q}+1)^{\frac{p-q}{q}}}\right\}.$$

Then if $\bar{x}, \bar{y} \in \ell^p$, $\|\bar{x}-\bar{y}\|_{\ell^p} < \delta$, we have

$$\|\bar{y}\|_{\ell^p} \leq \|\bar{x}\|_{\ell^p} + 1$$

and

$$\begin{aligned}\|f(\bar{x})-f(\bar{y})\|_{\ell^q}^q &= \sum_{i=1}^{\infty}\left||x_i|^{p/q}-|y_i|^{p/q}\right|^q \leq C\sum_{i=1}^{\infty}\left|(|x_i|+|y_i|)^{\frac{p-q}{q}}(|x_i-y_i|)\right|^q \\ &\leq C\sum_{i=1}^{\infty}(|x_i|+|y_i|)^{p-q}|x_i-y_i|^q \leq C\|(|x_i|+|y_i|)_i\|_{\ell^p}^{p-q}\|\bar{x}-\bar{y}\|_{\ell^p}^q,\end{aligned} \tag{2.1.9}$$

where we use Hölder's inequality with conjugate exponents p/q and $p/(p-q)$ in the last step and we are using $(|x_i|+|y_i|)_i$ to denote the sequence $(|x_1|+|y_1|, |x_2|+|y_2|, \dots)$. Notice that

$$\|(|x_i|+|y_i|)_i\|_{\ell^p} \leq \|(|x_i|)_i\|_{\ell^p} + \|(|y_i|)_i\|_{\ell^[} = \|\bar{x}\|_{\ell^p} + \|\bar{y}\|_{\ell^p} \leq 2\|\bar{x}\|_{\ell^p} + \delta \leq 2\|\bar{x}\|_{\ell^p} + 1.$$

Hence, (2.1.9) becomes

$$\|f(\bar{x})-f(\bar{y})\|_{\ell^q} < C^{\frac{1}{q}}(2\|\bar{x}\|_{\ell^q}+1)^{\frac{p-q}{q}}\delta = \varepsilon.$$

□

Exercise 2.1.8. *Is the above map uniformly continuous?*

Let's look at a few examples where continuity fails in order to get a sense of what can go wrong.

Example 2.1.12. *Consider the space*

$$X = \{\bar{x} \in \ell^2 : \lim_{n\to\infty} n^9 x_n = 0\}$$

with the ℓ^2-norm. Define the function:

$$T : (X, \ell^2) \to (\ell^2, \ell^2) \qquad \text{where } T\bar{x} = (x_1, 2x_2, 3x_3, \dots, nx_n, \dots).$$

At first glance this does not seem to be such a bad function – each coordinate function $f_i(\bar{x}) = ix_i$ is clearly continuous. But they are getting steeper as $i \to \infty$. This is what will cause us problems.

Suppose that T is continuous is at 0. Then there is $\delta > 0$ such that, if $\|\bar{x}-0\|_{\ell^2} < \delta$, then $\|T\bar{x}\|_2 = \|T\bar{x}-T0\|_2 < 1$. Fix $N > 10/\delta$. Then, define

$$\bar{x} = \frac{\delta}{2}e_N = (\underbrace{0, \dots, 0}_{\substack{\text{first } N \\ \text{entries}}}, \tfrac{\delta}{2}, 0, 0, \dots).$$

It is clear that $\|\bar{x}\|_2 = \delta/2$ and

$$T\bar{x} = (\underbrace{0, \dots, 0}_{\substack{\text{first } N \\ \text{entries}}}, \tfrac{N\delta}{2}, 0, 0, \dots).$$

Hence,

$$\|T\bar{x}\|_2 = \frac{N\delta}{2} > \frac{10}{\delta}\frac{\delta}{2} = 5 > 1.$$

Hence T is not continuous!

Notice that this lack of continuity boils down to the fact that there is no C such that

$$\|T\bar{x}\|_{\ell^2} \leq C\|\bar{x}\|_{\ell^2} \quad \text{for all } \bar{x}. \tag{2.1.10}$$

Indeed, we showed above that

$$\|Te_n\|_{\ell^2} = n\|e_n\|_{\ell^2}.$$

We will see later that inequalities of the form (2.1.10) *are very important for linear operators.*

Note: the above example is secretly using differentiation in L^2 with Fourier series. This is not at all obvious. But, given this context, we should not be surprised that it is not continuous – differentiation takes "smooth" objects and makes the "rougher." There is no reason that a "small" object has to have a "small" derivative. Indeed, one can check that $\varepsilon \sin(nx)$ has $L^2([0,1])$ norm that is roughly $\varepsilon/2$, but its derivative $\varepsilon n \cos(nx)$ has L^2-norm that is roughly $\varepsilon n/2$!

In a moment, we will see that, actually, T is continuous *nowhere*.

2.2. **Linear operators.** Throughout this section we use $(X, \|\cdot\|_X)$ and $(Y, \|\cdot\|_Y)$ to denote arbitrary normed linear spaces.

Definition 2.2.1. *A function $T : X \to Y$ is linear if, for every $x_1, x_2 \in X$ and $\alpha, \beta \in \mathbb{R}$,*

$$T(\alpha x_1 + \beta x_2) = \alpha T(x_1) + \beta T(x_2).$$

We often simply write Tx instead of $T(x)$.

The case $Y = \mathbb{R}$ plays an extremely special role in the theory that we will develop. As such, we give a special name to linear maps $T : X \to \mathbb{R}$. We call them linear functionals.

Let us point out how T acts on the zero vector. To make this discussion clear, we use some excessive notation: let $\bar{0}_X$ and $\bar{0}_Y$ be the zero vectors in, respectively, X and Y and let 0 to denote the scalar $0 \in \mathbb{R}$. Then, we have, by linearity,

$$T(\bar{0}_X) = T(0\bar{0}_X) = 0T(\bar{0}_X) = \bar{0}_Y.$$

The last equality follows because $0y = \bar{0}_Y$ for any $y \in Y$. OK, hopefully we never have to resort to such overwrought notation ever again...

Example 2.2.2. *When the domain and codomain are $\mathbb{R}^n$ and $\mathbb{R}^m$, respectively, all linear maps are given by $m \times n$ matrices. Why?*

Let $e_1, \ldots, e_n$ be the standard basis vectors[9] *in $\mathbb{R}^n$ and $f_1, \ldots, f_m$ be those for $\mathbb{R}^m$. Since, for any i, $T(e_i) \in \mathbb{R}^m$, then there are $\alpha_{1i}, \ldots, \alpha_{mi}$ such that*

$$T(e_i) = \alpha_{1i}f_1 + \alpha_{2i}f_2 + \cdots + \alpha_{mi}f_m.$$

[9]The vector e_i is the one with all zeros entries except for a 1 in the ith entry.

One can readily check that

$$T(\bar{x}) = \begin{bmatrix} \alpha_{11} & \alpha_{21} & \alpha_{13} & \cdots & \alpha_{1n} \\ \alpha_{21} & \alpha_{22} & & & \\ \vdots & & & \ddots & \\ \alpha_{m1} & \alpha m2 & \cdots & & \alpha_{mn} \end{bmatrix} \bar{x}.$$

It is clearly true when $\bar{x} = e_i$, by construction. After that it is easy to check by simply decomposing any $\bar{x}$ as $\bar{x} = \beta_1 e_1 + \cdots + \beta_n e_n$ and using the linearity of T.

Exercise 2.2.1. *Do this!*

It is not too hard to convince yourself, at least heuristically, that T has to be continuous. We will revisit this later more rigorously.

This way of thinking has its limits. An example is Example 2.1.12, where we could think of the operator as the infinitely large matrix

$$\begin{bmatrix} 1 & 0 & 0 & \cdots & & \\ 0 & 2 & 0 & \cdots & & \\ 0 & 0 & 3 & \cdots & & \\ & & & \ddots & & \\ \vdots & & & 0 & n & 0 \cdots \\ \vdots & & & & & \ddots \end{bmatrix}$$

However, this coordinate-based thinking might make you think that this matrix corresponds with a continuous operator, which, as we showed above, it does not.

Let's go over a few examples in "infinite dimensional" spaces to get used to this idea.

Example 2.2.3. *(i) The shift operator*

$$S : \ell^p \to \ell^p$$

defined by

$$S(x_1, x_2, \dots) = (0, x_1, x_2, \dots)$$

is linear. Indeed,

$$\begin{aligned} S(\alpha\bar{x} + \beta\bar{y}) &= S((\alpha x_1, \alpha x_2, \dots) + (\beta y_1, \beta y_2, \dots)) = S(\alpha x_1 + \beta y_1, \alpha x_2 + \beta y_2, \dots) \\ &= (0, \alpha x_1 + \beta y_1, \alpha x_2 + \beta y_2, \dots) = (0, \alpha x_1, \alpha x_2, \dots) + (0, \beta y_1, \beta y_2, \dots) \\ &= \alpha(0, x_1, x_2, \dots) + \beta(0, y_1, y_2, \dots) = \alpha S(\bar{x}) + \beta S(\bar{y}). \end{aligned}$$

We point out that

$$\|S(\bar{x})\|_{\ell^p} = \|\bar{x}\|_{\ell^p}.$$

(ii) Let $\iota : \ell^1 \to \ell^2$ be the inclusion map $\iota(\bar{x}) = \bar{x}$. It is obvious that ι is linear.

Notice that

$$\|\iota(\bar{x})\|_{\ell^2} \leq \|\bar{x}\|_{\ell^1}. \tag{2.2.1}$$

We showed this on a previous homework assignment. Moreover, we showed in Proposition 2.1.10 that ι is continuous.

Of course $\ell^2 \subsetneq \ell^1$ so that there is a space $X = \iota(\ell^1)$ in ℓ^2. One can check that $\iota(\ell^1)$ is a linear space and then we can look at:

$$\iota^{-1} : (X, \|\cdot\|_{\ell^2}) \to (\ell^1, \|\cdot\|_{\ell^1}),$$

which is given by $\iota^{-1}(\bar{x}) = \bar{x}$. (MOVE THIS) Is this function continuous and does an equation of the form (2.2.1) hold? For the latter, we mean, is there a constant $C > 0$ such that, for every $\bar{x} \in X$,

$$\|\iota^{-1}(\bar{x})\|_{\ell^2} \leq C\|\bar{x}\|_{\ell^1}?$$

This is, of course, not true. Take, for example, the elements

$$\bar{x}_N = \Big(\underbrace{\frac{1}{\sqrt{N}}, \dots, \frac{1}{\sqrt{N}}}_{N \text{ terms}}, 0, 0, \dots\Big)$$

It is easy to see that

$$\|\bar{x}_N\|_{\ell^2} = 1,$$

but

$$\|\iota^{-1}(\bar{x}_N)\|_{\ell^1} = \|\bar{x}_N\|_{\ell^1} = \sqrt{N}.$$

(iii) Define the map

$$T : \ell^3 \to \ell^p$$

by

$$T(x_1, x_2, \dots) = \Big(x_1, \frac{x_2}{\sqrt{2}}, \frac{x_3}{\sqrt{3}}, \dots, \frac{x_n}{\sqrt{n}}, \dots\Big).$$

This is a linear operator and is well-defined when $p = 3$. Indeed,

$$\begin{aligned} T(\alpha\bar{x} + \beta\bar{y}) &= \Big(\alpha x_1 + \beta y_1, \frac{\alpha x_2}{\sqrt{2}} + \frac{\beta y_2}{\sqrt{2}}, \dots\Big) = \Big(\alpha x_1, \frac{\alpha x_2}{\sqrt{2}}, \dots\Big) + \Big(\beta y_1, \frac{\beta y_2}{\sqrt{2}}, \dots\Big) \\ &= \alpha\Big(x_1, \frac{x_2}{\sqrt{2}}, \dots\Big) + \beta\Big(y_1, \frac{y_2}{\sqrt{2}}, \dots\Big) = \alpha T(\bar{x}) + \beta T(\bar{y}), \end{aligned}$$

and

$$\|T(\bar{x})\|_{\ell^3}^3 = \sum_{i=1}^{\infty} \frac{|x_i|^3}{i^{3/2}} \leq \sum_{i=1}^{\infty} |x_i|^3 \leq \|\bar{x}\|_{\ell^3}^3.$$

Exercise 2.2.2. *For which other p is T well-defined?*

(iv) Define the map

$$D : (C^\infty([0,1]), \|\cdot\|_1) \to (C^\infty([0,1]), \|\cdot\|_1)$$

where

$$Df = f'.$$

This is clearly well-defined, and it is linear:

$$D(\alpha f + \beta g) = (\alpha f + \beta g)' = \alpha f' + \beta g' = \alpha Df + \beta Dg.$$

(v) Define the map

$$\delta_{\frac{1}{2}} : (C([0,1]), \|\cdot\|_p) \to \mathbb{R}$$

by

$$\delta_{\frac{1}{2}}(f) = f(1/2).$$

This is clearly linear.

We will investigate the continuity of these maps further in a moment. Before we do this, though, we state one of the most important facts about linear operators.

Definition 2.2.4. *Suppose that $T : (X, \|\cdot\|_X) \to (Y, \|\cdot\|_Y)$ is a linear map between normed linear spaces. We say that T is* bounded *if there is a constant C such that*

$$\|Tx\|_Y \leq C\|x\|_X.$$

We denote by $\mathcal{B}(X,Y)$ the set of bounded linear operators from $X \to Y$ and we define the norm[10]

$$\|T\|_{\mathcal{B}(X,Y)} = \sup_{x \neq 0} \frac{\|Tx\|_Y}{\|x\|_X}.$$

Note: often we drop the $\mathcal{B}(X,Y)$ subscript.

Exercise 2.2.3. *Show that*

$$\|T\| = \sup_{\substack{x \in X \setminus \{0\}, \\ \|x\|_X \leq 1}} \|Tx\|_Y = \sup_{\substack{x \in X, \\ \|x\|_X = 1}} \|Tx\|_Y.$$

From a philosophical point of view, why are bounded linear operators important? Let us change perspective a bit and think about a function $T : X \to Y$ as describing the "similarity" of its domain X and codomain Y. When we discuss normed linear spaces, the important parts of the space are *only* the linear structure and the norms. If T is linear as an operator, then it necessarily preserves the linear structure of X, which shows that, at least on the range of T, the linear structure of X and Y are the same. If T is bounded, then it does not distort the norms "too much," telling us that there is some "similarity" of the norms $\|\cdot\|_X$ and $\|\cdot\|_Y$.

Theorem 2.2.5. *Suppose that $T : (X, \|\cdot\|_X) \to (Y, \|\cdot\|_Y)$ is a linear map between normed linear spaces and $\bar{x} \in X$. Then the following are equivalent:*

(a) T is bounded;

(b) T is uniformly continuous;

(c) T is continuous at $\bar{x}$;

(d) T is sequentially continuous at $\bar{x}$.

[10] We will show this is a norm later.

Proof. From the definition of continuity and Theorem 2.1.6, we have immediately that (b) implies (c), and (c) implies (d). Hence, we need only establish that (a) implies (b) and (d) implies (a).

(a) implies (b): Fix $\varepsilon > 0$ and let

$$\delta = \frac{\varepsilon}{1 + \|T\|}.$$

Then, for any $x_1, x_2 \in X$, if $\|x_1 - x_2\|_X < \delta$, we have

$$\|T(x_1) - T(x_2)\|_Y = \|T(x_1 - x_2)\|_Y \leq \|T\|\|x_1 - x_2\|_X \leq \|T\|\delta \leq \|T\|\frac{\varepsilon}{1 + \|T\|} < \varepsilon.$$

Thus T is uniformly continuous.

(d) implies (a): We show this by contrapositive (that is, if (a) does not hold, then (d) does not hold). Suppose that T is not bounded. Then there are x_n such that

$$\frac{\|Tx_n\|_Y}{\|x_n\|_X} \to \infty \qquad \text{as } n \to \infty. \tag{2.2.2}$$

Let

$$\tilde{x}_n = \bar{x} + \frac{x_n}{\|Tx_n\|_Y}.$$

By (2.2.2) and the homogeneity of $\|\cdot\|_Y$, we see that

$$\|\tilde{x}_n - \bar{x}\|_X = \frac{\|x_n\|_X}{\|Tx_n\|_Y} \to 0 \qquad \text{as } n \to \infty.$$

Hence $\tilde{x}_n \to \bar{x}$ as $n \to \infty$. On the other hand,

$$\begin{aligned}\|T\tilde{x}_n - T\bar{x}\|_Y &= \left\|T\bar{x} + T\frac{x_n}{\|Tx_n\|_Y} - T\bar{x}\right\|_Y = \left\|T\frac{x_n}{\|Tx_n\|_Y}\right\|_Y \\ &= \left\|\frac{1}{\|Tx_n\|_Y}Tx_n\right\|_Y = \frac{\|Tx_n\|_Y}{\|Tx_n\|_Y} = 1.\end{aligned}$$

Hence $T\tilde{x}_n \not\to T\bar{x}$ as $n \to \infty$. Thus T is not sequentially continuous at $\bar{x}$. □

Let's revisit Example 2.2.3 to see which of these maps are continuous using Theorem 2.2.5.

Example 2.2.6. *(i) The shift operator*

$$S : \ell^p \to \ell^p$$

defined by

$$S(x_1, x_2, \dots) = (0, x_1, x_2, \dots)$$

is linear. We showed above that S is an isometry *(see Definition 2.2.7), that is,*

$$\|S\bar{x}\|_{\ell^p} = \|\bar{x}\|_{\ell^p} \quad \text{for all } \bar{x} \in \ell^p.$$

It follows that $\|S\| = 1$. Hence, S is (unsurprisingly) continuous.

(ii) Let $\iota : \ell^1 \to \ell^2$ be the inclusion map $\iota(\bar{x}) = \bar{x}$. It is obvious that ι is linear. As we noted above,

$$\|\iota\bar{x}\|_{\ell^2} \leq \|\bar{x}\|_{\ell^1} \qquad \text{for all } \bar{x} \in \ell^p.$$

Hence, we have that $\|\iota\| \leq 1$. This is enough to deduce that ι is continuous. On the other hand, it is easy to see that $\|\iota e_1\|_{\ell^2} = \|e_1\|_{\ell^1}$. Hence, $\|\iota\| = 1$.

(iii) Define the map

$$T : \ell^3 \to \ell^3$$

by

$$T(x_1, x_2, \dots) = \left(x_1, \frac{x_2}{\sqrt{2}}, \frac{x_3}{\sqrt{3}}, \dots, \frac{x_n}{\sqrt{n}}, \dots\right).$$

Exercise 2.2.4. *The exact same analysis as in (ii) lets us conclude that $\|T\| = 1$ and T is continuous. Show this!*

(iv) Define the map

$$D : (C^\infty([0,1]), \|\cdot\|_1) \to (C^\infty([0,1]), \|\cdot\|_1)$$

where

$$Df = f'.$$

The L^1-norm of f captures how big f is, which is unrelated to how much f' wiggles, as captured by the L^1-norm of f'. Hence, we suspect that D will not be a bounded linear operator. We show this now by constructing a sequence of functions f_n such that $\|Df_n\|_1/\|f_n\|_1 \to \infty$. Let

$$f_n(x) = ne^{-nx}.$$

Notice that $\|f_n\|_1 = 1 - e^{-n}$. On the other hand,

$$\|Df_n\|_1 = \int \left|-ne^{-nx}\right| dx = n \int f_n(x)dx = n\|f_n\|_1.$$

Thus, as $n \to \infty$,

$$\frac{\|Df_n\|_1}{\|f_n\|_1} = n \to \infty, \qquad \text{which implies that } \|D\| = \sup_{f \in L^1} \frac{\|Df\|_1}{\|f\|_1} = \infty.$$

So D is not continuous.

(v) Define the map

$$\delta_{\frac{1}{2}} : (C([0,1]), \|\cdot\|_p) \to \mathbb{R}$$

by

$$\delta_{\frac{1}{2}}(f) = f(1/2).$$

Let us consider the case $p < \infty$ first. In this case, we have to ask ourselves, does the L^p-norm of a function control its size at a single point? This is, of course, not true, since one can imagine f that has a maximum of 1 but is zero nearly everywhere else so that its L^p-norm is small. We construct such an f now. For $n \in \mathbb{N}$, let

$$f_n(x) = e^{-n|x-1/2|}.$$

Then, $\delta_{\frac{1}{2}}(f_n) = 1$, but as $n \to \infty$,

$$\|f_n\|_p = \left(\int e^{-np|x-1/2|}dx\right)^{\frac{1}{p}} = \left(\frac{2}{np}\left(1 - e^{-np/2}\right)\right)^{\frac{1}{p}} \to 0.$$

Hence,

$$\|\delta_{\frac{1}{2}}\| \geq \sup_n \frac{|\delta_{\frac{1}{2}} f_n|}{\|f_n\|_p} = \sup_n \left(\frac{np}{2\left(1 - e^{-np/2}\right)}\right)^{\frac{1}{p}} = \infty.$$

Hence $\delta_{\frac{1}{2}}$ is not bounded for $p < \infty$.

Exercise 2.2.5. *Show that $\delta_{\frac{1}{2}}$ is continuous when $p = \infty$.*

There is a special type of bounded linear transformation that is important in the study of duality. We briefly introduce it here.

Definition 2.2.7. *We say that $T \in \mathcal{B}(X, Y)$ is an* **isometry** *if, for all $x \in X$,*

$$\|T(x)\|_Y = \|x\|_X.$$

Why is this important? Recall the discussion above about how a linear operator $T : X \to Y$ tells us how "similar" to normed spaces are. If T is an isometry, then it tells us that the norms $\|\cdot\|_X$ and $\|\cdot\|_Y$ (at least when restricted to the range of T, which may not be all of Y), are exactly the same. Taking this a bit further, it tells us that

$$T(X) = \{T(x) : x \in X\}$$

is an "exact copy" of X that lives inside of Y.

Exercise 2.2.6. *(i) Show that any isometry is injective.*

(ii) Show that $\iota : (\mathbb{R}^n, \ell^p) \to \ell^p$ given by $\iota(\bar{x}) = (x_1, \dots, x_n, 0, \dots)$ is an isometry.

(iii) Show that $\iota : \ell^1 \to \ell^\infty$ given by $\iota(\bar{x}) = \bar{x}$ is not *an isometry.*

2.2.1. **The space $\mathcal{B}(X, Y)$ is a normed linear space.** Notice that, given $T_1, T_2 \in \mathcal{B}(X, Y)$ and $\alpha, \beta \in \mathbb{R}$, we can define a new normed linear operator $(\alpha T_1 + \beta T_2)$, defined by

$$(\alpha T_1 + \beta T_2)(x) = \alpha T_1(x) + \beta T_2(x).$$

This is clearly well-defined, and, by the triangle inequality, we get that

$$\|(\alpha T_1 + \beta T_2)(x)\|_Y \leq \|\alpha T_1(x)\|_Y + \|\beta T_2(x)\|_Y \leq |\alpha|\|T_1\|\|x\|_X + |\beta|\|T_2\|\|x\|_X.$$

Hence, $(\alpha T_1 + \beta T_2)$ is bounded and, thus, an element of $\mathcal{B}(X, Y)$. We conclude that $\mathcal{B}(X, Y)$ is a linear space.

Now we check that $\|\cdot\|_{\mathcal{B}(X,Y)}$ is, indeed, a norm, as our notation suggests.

- (Positive definiteness) It is obvious that $\|T\|$ is non-negative. If $T \neq 0$, then $Tx_1 \neq 0$ for some $x_1 \in X \setminus \{0\}$. It follows that

  $$\|T\| = \sup_{x \neq 0} \frac{\|Tx\|_Y}{\|x\|_X} \geq \frac{\|Tx_1\|_Y}{\|x_1\|_X} > 0,$$

  where the last inequality uses the positive definiteness of $\|\cdot\|_X$ and $\|\cdot\|_Y$. On the other hand, if $T = 0$, it is clear that $\|T\| = 0$.

- (Homogeneity) Fix any $\alpha\mathbb{R}$. To show that $\|\alpha T\| = |\alpha|\|T\|$, we first show that $\|\alpha T\| \leq |\alpha|\|T\|$ and then we show that $\|\alpha T\| \geq |\alpha|\|T\|$.

  Notice that, for any $x \in X$, we have

  $$\|(\alpha T)(x)\|_Y = |\alpha|\|Tx\|_Y \leq |\alpha|\|T\|\|x\|_X,$$

which implies that

$$\|\alpha T\| \leq |\alpha|\|T\|.$$

On the other hand, let $x_n \in X$ be a sequence of points such that $\|x_n\|_X = 1$ for all n and

$$\|T\| = \lim_{n\to\infty} \|Tx_n\|.$$

Then

$$\begin{aligned}\|\alpha T\| &\geq \limsup_{n\to\infty} \|(\alpha T)(x_n)\|_Y = \limsup_{n\to\infty} \|\alpha T(x_n)\|_Y \\ &= \limsup_{n\to\infty} |\alpha|\|T(x_n)\|_Y = |\alpha| \limsup_{n\to\infty} \|T(x_n)\|_Y = |\alpha|\|T\|.\end{aligned}$$

The second equality above uses the homogeneity of $\|\cdot\|_Y$.

- (Triangle inequality) Take any $T, S \in \mathcal{B}(X,Y)$ and any $x \in X$. Then we have

$$\begin{aligned}\|(S+T)(x)\|_Y &= \|Sx + Tx\|_Y \leq \|Sx\|_Y + \|Tx\|_Y \\ &\leq \|S\|\|x\|_X + \|T\|\|x\|_Y = (\|S\| + \|T\|)\,\|x\|_Y,\end{aligned}$$

where the first inequality is the triangle inequality in Y and the second inequality uses the definition of $\|S\|$ and $\|T\|$. Since this holds for all x, we have that

$$\|S+T\| \leq \|S\| + \|T\|.$$

2.2.2. **The dual space: X^*.**

Definition 2.2.8. *Given any normed linear space X, a very special case of $\mathcal{B}(X,Y)$ is* **the dual space***:*

$$X^* := \mathcal{B}(X, \mathbb{R}).$$

This is also called the space of **bounded linear functionals.**

Example 2.2.9. *(i) Recall that $c \subset \ell^\infty$ is the set of all sequences $\bar{x}$ such that $\lim_{n\to\infty} x_n$ exists. Define*

$$\lambda : c \to \mathbb{R} \qquad \text{by } \lambda(\bar{x}) = \lim_{n\to\infty} x_n.$$

Clearly λ is well-defined and linear. Additionally,

$$|\lambda(x)| = |\lim_{n\to\infty} x_n| \leq \sup_n |x_n| = \|\bar{x}\|_{\ell^\infty}.$$

Hence, λ is bounded, and, as a result, is an element of the dual space c^*.

Exercise 2.2.7. *Show that there is no $\bar{y}$ such that*

$$\lambda(\bar{x}) = \sum_{i=1}^{\infty} x_i y_i \qquad \text{for all } \bar{x} \in c.$$

Note: using some fancy functional analysis techniques, we can utilize this λ to show that $(\ell^\infty)^* \neq \ell^1$.

(ii) Consider $C^1([0,1])$ with the norm

$$\|f\|_{C^1} = \|f\|_\infty + \|f'\|_\infty.$$

Then

$$\lambda : C^1 \to \mathbb{R} \qquad \text{defined by } \lambda(f) = f(0) - 7f'(2/3)$$

is a bounded linear functional: $\lambda \in (C^1)^*$. Indeed, it is clearly well-defined and linear, while

$$|\lambda(f)| = |f(0) - 7f(2/3)| \leq |f(0)| + 7|f(2/3)| \leq 7\|f\|_\infty + 7\|f'\|_\infty = \|f\|_{C^1}.$$

The quintessential examples of a dual space are (where p and q are conjugate exponents):

- $(\mathbb{R}^n, \|\cdot\|_p)^* = (\mathbb{R}^n, \|\cdot\|_q)$;
- $(\ell^p)^* = \ell^q$ when $p < \infty$ and $(\ell^\infty)^* \supsetneq \ell^1$;
- $(L^p)^* = L^q$ when $p < \infty$ and $(L^\infty)^* \supsetneq L^1$ (see important footnote[11]).

This is a little poorly defined because we need to write down a way of identifying bounded linear functionals with elements of the space on the right hand side of the equality. This is exactly where the notion of an isometry comes in. If we can define an isometry $\iota : X \to Y^*$, then it tells us that there is an exact copy of X living inside of Y. If ι is a surjection (and, thus, a bijection by Exercise 2.2.6), that exact copy of X is all of Y^*. In other words, Y^* is an exact copy of X. We usually abuse terminology in this case and simply say that X *is* Y^*.

Let us first see this in a the case of $\mathbb{R}^n$:

Lemma 2.2.10. *The dual space of $(\mathbb{R}^n, \|\cdot\|_p)$ is $(\mathbb{R}^n, \|\cdot\|_q)$, where p and q are conjugate exponents. By this, we mean that there is a linear map*

$$\iota : (\mathbb{R}^n, \|\cdot\|_q) \to (\mathbb{R}^n, \|\cdot\|_p)^*$$

that is bijective (one-to-one and onto) and is defined in the following natural way: for any $\bar{x} \in (\mathbb{R}^n, \|\cdot\|_q)$, $\iota(\bar{x})$ is the element in $(\mathbb{R}^n, \|\cdot\|_p)^*$ defined by:

$$\iota(\bar{x})(\bar{y}) = \sum_{i=1}^n x_i y_i \qquad \text{for all } \bar{y} \in \mathbb{R}^n.$$

Moreover, ι is an isometry (recall Definition 2.2.7).

Proof. It is easy to see that ι is linear. Next we check that it is one-to-one. Notice that, for any $\bar{x}$ and $\bar{z}$, we have

$$\iota(\bar{x})(\bar{x} - \bar{z}) - \iota(\bar{z})(\bar{x} - \bar{z}) = \iota(\bar{x} - \bar{z})(\bar{x} - \bar{z}) = \sum_{i=1}^n |x_i - z_i|^2.$$

This is zero if and only if $\bar{x} = \bar{z}$. Hence, $\iota(\bar{x}) = \iota(\bar{z})$ if and only if $\bar{x} = \bar{z}$; that is, ι is one-to-one.

[11] We have been a bit vague in these notes about what L^∞ means, which, up to now, really did not matter. It does, however, play a huge role in determining the dual space. Actually, $(C[0,1], L^\infty)^*$ is "nearly" $L^1([0,1])$ (in a sense that we will not make more precise in these notes), but $(L^\infty([0,1]))^* \neq L^1([0,1])$. To properly make sense of this, we require measure theory, which we develop in a later section.

We now check that ι is onto. Fix any $\lambda \in (\mathbb{R}^n, \|\cdot\|_p)^*$. Let e_i be the standard basis vectors (e_i is a vector of all zeros, except in the ith coordinate where it is one). Let

$$\bar{x} = (\lambda(e_1), \lambda(e_2), \ldots, \lambda(e_n)) \in \mathbb{R}^n.$$

By linearity, we have, for any $\bar{y} \in R^n$,

$$\lambda(\bar{y}) = \lambda(y_1 e_1 + y_2 e_2 + \cdots + y_n e_n) = \sum_{i=1}^{n} y_i \lambda(e_i) = \sum_{i=1}^{n} y_i x_i = \iota(\bar{x})(\bar{y}).$$

Since this is true for all $\bar{y}$, we have that $\lambda = \iota(\bar{x})$.

Finally, we check the equality of the norms. We do this for $p > 1$. Fix any $\bar{x} \in (\mathbb{R}^n, \|\cdot\|_q)$. First, by Hölder's inequality, we have, for all $\bar{y}$.

$$|\iota(\bar{x})(\bar{y})| \leq \|\bar{x}\|_q \|\bar{y}\| p,$$

which implies that

$$\|\iota(\bar{x})\| \leq \|\bar{x}\|_q. \tag{2.2.3}$$

On the other hand, let $\bar{z}$ have coordinates

$$z_i = \operatorname{sign}(x_i)|x_i|^{q-1}.$$

It is easy to check that

$$\|\bar{z}\|_p = \|\bar{x}\|_q^{q-1}.$$

Then

$$|\iota(\bar{x})(\bar{z})| = \sum_{i=1}^{n} |x_i|^q = \|\bar{x}\|_q^q = \|\bar{x}\|_q \|\bar{x}\|_p.$$

Thus, we have

$$\|\iota(\bar{x})\| \geq \|\bar{x}\|_q. \tag{2.2.4}$$

The combination of (2.2.3) and (2.2.4) yield the claim in the case $p > 1$.

Exercise 2.2.8. *Complete the proof in the case $p = 1$.*

□

We point out that it is not so difficult to prove the sequence version of this:

Exercise 2.2.9. *Show that $(\ell^p)^* = \ell^q$ where p and q are conjugate exponents with $p < \infty$. Here, use the identification that, for a given $\bar{x} \in \ell^q$,*

$$\lambda_{\bar{x}}(\bar{y}) = \sum_{i=1}^{\infty} x_i y_i \qquad \text{for any } \bar{y} \in \ell^p.$$

How does this work for L^p? We identify a function $f \in L^q$ with a bounded linear functional $\lambda_f \in L^q$ as follows:

$$\lambda_f : L^p \to \mathbb{R}, \quad \text{defined by, for all } g \in L^p \quad \lambda_f(g) = \int f(x)g(x)dx. \tag{2.2.5}$$

Then the content of the statement $(L^p)^* = L^q$ is to say the following: $\lambda_f \in (L^p)^*$ and, given any $\lambda \in (L^p)^*$, there is $f \in L^q$ such that $\lambda = \lambda_f$ and $\|\lambda_f\| = \|f\|_q$. Let us prove *most* of this.

Proposition 2.2.11. *Let $p, q \in [1, \infty]$ be conjugate exponents. Let $f \in L^q$, and define λ_f as in* (2.2.5)*. Then $\lambda_f \in (L^p)^*$. Moreover,*

$$\|\lambda_f\| = \|f\|_q. \tag{2.2.6}$$

Proof. It is clear that λ_f is well-defined and linear. The fact that λ_f is bounded follows directly from Hölder's inequality: for any $g \in L^p$,

$$|\lambda_f(g)| = \left| \int f(x)g(x)dx \right| \leq \|f\|_q \|g\|_p. \tag{2.2.7}$$

To show (2.2.6), we first notice that (2.2.7) implies that $\|\lambda_f\| \leq \|f\|q$. The other inequality follows directly from Proposition 1.4.7. □

2.2.3. **Aside: matrix norms.** Given a norm $\|\cdot\|$ on $\mathbb{R}^n$, a norm $|||\cdot|||$ on $\mathbb{R}^m$, and a $m \times n$ matrix A, we can think of:

$$A : (\mathbb{R}^n, \|\cdot\|) \to (\mathbb{R}^m, |||\cdot|||).$$

Is A bounded? Of course! We recall the following lemma that is a homework problem for you:

Lemma 2.2.12. *There exists $C > 0$ such that*

$$\frac{1}{C}\|x\|_2 \leq \|x\| \leq C\|x\|_2,$$

where $\|\cdot\|_2$ is the usual Euclidean norm and $\|\cdot\|$ is any norm.

Exercise 2.2.10. *Prove this!*

Now we show that $A \in \mathcal{B}((\mathbb{R}^n, \|\cdot\|), (\mathbb{R}^m, |||\cdot|||))$. Let $e_1, \ldots, e_n$ be the standard basis vectors in $\mathbb{R}^n$. Then, for any $\bar{x} \in \mathbb{R}^n$, we have

$$\bar{x} = x_1 e_1 + \cdots x_n e_n,$$

so that

$$A\bar{x} = x_1 Ae_1 + \cdot + x_n Ae_n.$$

Thus, by the Cauchy-Schwarz inequality and Lemma 2.2.12,

$$\begin{aligned} |||A\bar{x}||| &\leq \sum_{i=1}^n |x_i| \, |||Ae_i||| \leq \left(\sum_{i=1}^n |x_i|^2 \right)^{\frac{1}{2}} \left(\sum_{i=1}^n |||Ae_i|||^2 \right)^{\frac{1}{2}} \\ &= \|\bar{x}\|_2 \left(\sum_{i=1}^n |||Ae_i|||^2 \right)^{\frac{1}{2}} \leq C \left(\sum_{i=1}^n |||Ae_i|||^2 \right)^{\frac{1}{2}} \|\bar{x}\|. \end{aligned}$$

Hence,

$$\|A\|_{\mathcal{B}((\mathbb{R}^n, \|\cdot\|), (\mathbb{R}^m, |||\cdot|||))} \leq C \left(\sum_{i=1}^n |||Ae_i|||^2 \right)^{\frac{1}{2}} < \infty.$$

Of course, the choice of $\|\cdot\|$ and $|||\cdot|||$ determine what norm of A we get, and, depending on what we want to measure (i.e., depending on the application), certain norms will be better than others. Some standard choices are:

Example 2.2.13. *(i) When $\|\cdot\| = \|\cdot\|_\infty$ and $|||\cdot||| = \|\cdot\|_\infty$, we find $\|A\|_{\infty,\infty}$ as follows. Let $v_1, \ldots, v_m$ be the row vectors of A. In other words,*

$$A = \begin{bmatrix} - & v_1 & - \\ - & v_2 & - \\ & \vdots & \\ - & v_m & - \end{bmatrix} \qquad \text{so that, for all } x \in \mathbb{R}^m, \quad Ax = \begin{pmatrix} x \cdot v_1 \\ x \cdot v_2 \\ \vdots \\ x \cdot v_m \end{pmatrix}.$$

Then

$$\|Ax\|_\infty = \max\{|x \cdot v_1|, \ldots, |x \cdot v_n|\} \leq \max\{\|x\|_\infty \|v_1\|_1, \ldots, \|x\|_\infty \|v_n\|_1\},$$

which implies that

$$\|A\|_{\infty,\infty} \leq \max\{\|v_1\|_1, \ldots, \|v_n\|_1\}.$$

Is this sharp? Of course it is! Take $k \in \{1, \ldots, n\}$ to be the index of the maximal $\|v_i\|_1$. Take $\hat{v}_k$ to be such that $\hat{v}_k \cdot v_k = \|\hat{v}_k\|_\infty \|v_i\|_1$. Then, it is easy to check that $\|A\hat{v}_k\|_\infty = \|\hat{v}_k\|_\infty \|v_k\|_1$. Hence

$$\|A\|_{\infty,\infty} = \max\{\|v_1\|_1, \ldots, \|v_n\|_1\}$$

(ii) We consider $A : (\mathbb{R}^n, \|\cdot\|_2) \to (\mathbb{R}^n, \|\cdot\|_2)$ for any A symmetric. Then by the spectral theorem, there are eigenvalues

$$\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$$

associated to orthonormal eigenvectors $v_1, \ldots, v_n$ (that is $v_i \cdot v_j = 0$ if $i \neq j$ and $\|v_i\|_2 = 1$). We claim that $\|A\| = \max\{|\lambda_i| : i = 1, \ldots, n\}$. Indeed, for any $x \in \mathbb{R}^n$, we have

$$x = \alpha_1 v_1 + \cdots + \alpha_n v_n \qquad \text{where } \|x\|_2^2 = \alpha_1^2 + \cdots + \alpha_n^2.$$

Then

$$Ax = A(\alpha_1 v_1 + \cdots + \alpha_n v_n) = \alpha_1 \lambda_1 v_1 + \cdots \alpha_n \lambda_n v_n.$$

Using then the orthonormality of $v_1, \ldots, v_n$, we have

$$\begin{aligned} \|Ax\|_2^2 &= \alpha_1^2 \lambda_1^2 + \cdots + \alpha_n^2 \lambda_n^2 \leq \max\{|\lambda_i| : i = 1, \ldots, n\}^2 (\alpha_1^2 + \cdots + \alpha_n^2) \\ &\leq \max\{|\lambda_i| : i = 1, \ldots, n\}^2 \|x\|_2^2. \end{aligned}$$

Hence $\|A\|_{2,2} \leq \max\{|\lambda_i| : i = 1, \ldots, n\}$.

The opposite inequality follows by evaluating Av_k where k is such that $|\lambda_k| = \max\{|\lambda_i| : i = 1, \ldots, n\}$.

(iii)

Exercise 2.2.11. *When $A : (\mathbb{R}^n, \|\cdot\|_2) \to (\mathbb{R}^m, \|\cdot\|_2)$, show that $\|A\|_{2,2} = \sqrt{\max\{\lambda_i(A^T A)\}}$.*

Let us note that there are many many other norms for matrices that do not come from thinking of them as operators. For example, $\|A\|_\infty = \max_{ij} |a_{ij}|$. Each norm will be more or less useful depending on the application.

Aside: condition number of a matrix. Suppose we are seeking a solution of the problem

$$Ax = b$$

where A is an $m \times n$ matrix and we are working with the norms such that

$$A : (\mathbb{R}^n, \|\cdot\|) \to (\mathbb{R}^m, |||\cdot|||).$$

Suppose that we can only measure b up to error $\varepsilon|||b|||$ so that we can only work with $\tilde{b} \in \mathbb{R}^m$ with

$$|||\tilde{b} - b||| \leq \varepsilon|||b|||.$$

(For example, perhaps $\tilde{b}$ is found by some inexact algorithm). How bad will our estimate for x be? That is, if we let $\tilde{x} = A^{-1}\tilde{b}$, what estimate do we have on

$$\|x - \tilde{x}\|?$$

Notice that the relative error is given by

$$\begin{aligned}\frac{\|x - \tilde{x}\|}{\|x\|} = \frac{\|A^{-1}(b - \tilde{b})\|}{\|x\|} &\leq \frac{\|A^{-1}\|\,|||b - \tilde{b}|||}{\|x\|} \\ &\leq \frac{\varepsilon\|A^{-1}\|\,|||b|||}{\|x\|} = \frac{\varepsilon\|A^{-1}\|\,|||Ax|||}{\|x\|} \leq \varepsilon\|A^{-1}\|\,|||A|||.\end{aligned}$$

Hence the product $r(A) = \|A^{-1}\|\,|||A|||$ (which is necessarily at least 1!) measures the amplification of the relative error. It is called the **condition number**.

Note, the condition number is *scaling invariant*. Indeed, consider the condition number associated to αA for any $\alpha > 0$. Notice that $(\alpha A)^{-1} = \alpha^{-1}A^{-1}$ so that

$$\|(\alpha A)^{-1}\|\,|||\alpha A||| = \frac{1}{\alpha}\|A^{-1}\|\alpha\|A|||| = \|A^{-1}\|\,|||A|||.$$

Example 2.2.14. *(i)* $A = \begin{bmatrix} \frac{1}{1000} & \frac{1}{500} \\ 0 & \frac{1}{2000} \end{bmatrix}$. *At first, these numbers look terrible! But, due to scaling invariance, it is enough to consider the simpler matrix*

$$B = \begin{bmatrix} 2 & 4 \\ 0 & 1 \end{bmatrix} \qquad \text{since } A = \frac{1}{2000}B.$$

Now we are more optimistic that these more "reasonable" numbers will not give us a terrible condition number...

It is easy to see that $B^{-1} = \begin{bmatrix} 1/2 & -2 \\ 0 & 1 \end{bmatrix}$. *Thus*

$$r_{\infty,\infty}(A) = r_{\infty,\infty}(B) = \|B\|_{\infty,\infty}\|B^{-1}\|_{\infty,\infty} = \frac{5}{2} \cdot 6 = 15,$$

which is not terrible.

(ii)

Exercise 2.2.12. *Find the condition number for the matrix A of the previous example but using the $\|\cdot\|_{2,2}$ norm.*

$M = \begin{bmatrix} 1 & 1.999 \\ 1/2 & 1 \end{bmatrix}$. *These numbers look more reasonable, but, it turns out that*

$$r_{\infty,\infty}(M) = 2.999 \cdot 5998 (\approx 18000!).$$

Roughly what is going on in the previous problems is that M is much "closer" to non-invertible matrices than A is. Indeed,

$$M + \begin{bmatrix} 0 & .001 \\ 0 & 0 \end{bmatrix}$$

is not invertible. Hence, it is more likely that we add a huge vector to x and have it change b by only a little bit. This obstructs our ability to find approximate x's given approximate b's. Hence, the condition number has to be large.

3. Topology

3.1. Dense sets and separable metric spaces.

3.1.1. **How much information does a continuous function have?.** Let us begin by noticing the following silly oversight from our entire mathematical background... how is $\sqrt{2}$ defined? It is an irrational number, so we cannot "just write it down." Instead, we have to define it in a roundabout way. Let $f : \mathbb{R} \to \mathbb{R}$ be defined by $f(x) = x^2$. This is a continuous function and $f(0) = 0$, while $f(2) = 4$. Hence, the intermediate value theorem implies that there exists some number $\theta \in (0, 2)$ such that $f(\theta) = 2$. We call $\sqrt{2} = \theta$.

If we only have the natural numbers at our disposal, the "closest" we can come to $\sqrt{2}$ is to note that $\sqrt{2} \in (1, 2)$ since

$$f(1) = 1 < 2 < 4 = f(2).$$

What if we add the half-integers $\mathbb{Z}/2$? Then the best we can do is

$$f(1) = 1 < 2 < \frac{9}{4} = f(3/2).$$

This is still not "good enough," so we need to add *more* elements. Of course, $\mathbb{Z}/4$, $\mathbb{Z}/8$, etc. will never get us "close enough." But if we use $\mathbb{Q}$, we will be all set. Why? $\mathbb{Q}$ is dense in $\mathbb{R}$.

Definition 3.1.1. *Given a metric space (X, d) and a subset $Y \subset X$, we say that Y is dense in X if, for every $x \in X$, there is a sequence $(y_n)_n$ of elements in Y such that*

$$\lim_{n\to\infty} y_n = x.$$

To push this a bit further, what if we only know $f(x) = x^2$ for $x \in \mathbb{Q}$ and that f is continuous... is that enough to determine $f(x)$ for every $x \in \mathbb{R}$? Yes! Indeed, take any $x \in \mathbb{R}$ and let x_n be a sequence of rational numbers such that $x_n \to x$. Then, we must have that

$$f(x) = \lim_{n\to\infty} f(x_n).$$

Going about this backwards, if we "restrict" f to $\mathbb{Q}$ (that is, we consider the function $f : \mathbb{Q} \to \mathbb{R}$ defined by $f(x) = x^2$), it has enough information to extend it to a *continuous* function $\tilde{f} : \mathbb{R} \to \mathbb{R}$ that agrees with f on $\mathbb{Q}$ (that is, $f(x) = \tilde{f}(x)$ for all $x \in \mathbb{Q}$).

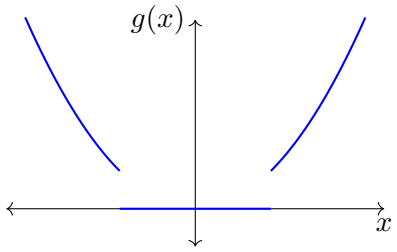
Exercise 3.1.1. *Show this in generality: suppose $f, g : (X, d_X) \to (Y, d_Y)$ are continuous functions that are equal on $\tilde{X} \subset X$, which is dense, then $f = g$.*

Will that always be the case? Starting with a continuous function on $\mathbb{Q}$, can we extend it to a continuous function on $\mathbb{R}$? Naïvely, we might expect that the answer is yes. But, as mathematicians, we should be suspicious of heuristics unless we can come up with some mathematics to back it up...

Consider the following example: let $g : \mathbb{Q} \to \mathbb{R}$ be defined by

$$g(x) = \begin{cases} x^2 & \text{if } x^2 > 2, \\ 0 & \text{if } x^2 < 2. \end{cases}$$

If we graph this, it might "seem" discontinuous. Roughly it looks like the following:



But let us think about this more carefully. Fix any $x_0 \in \mathbb{Q}$ and we will show that g is continuous at x_0. Let $\varepsilon > 0$.

(i) If $x_0^2 < 2$, let $\delta = (2 - x_0^2)/10$. Then, if $|x - x_0| < \delta$,

$$x^2 = x_0^2 + (x - x_0)(x + x_0) \leq 2 - 10\delta + (x - x_0)(x + x_0) \leq 2 - \delta + \delta(|x| + |x_0|). \tag{3.1.1}$$

Since $x_0^2 < 2$, then $|x_0| < 2$. Since $\delta = (2 - x_0^2)/10$ and $x_0^2 \geq 0$, then $\delta < 1$. By the triangle inequality, we have that $|x| < |x_0| + 1 < 3$. Returning to (3.1.1), we find

$$x^2 \leq 2 - 10\delta + \delta(|x| + |x_0|) < 2 - 10\delta + \delta \cdot 5 < 2 - 5\delta.$$

It follows that $g(x) = 0$. It follows that if $|x - x_0| < \delta$ then $|g(x) - g(x_0)| = 0 < \varepsilon$.

(ii) If $x_0^2 > 2$, let

$$\delta = \min\left\{1, \frac{x_0^2 - 2}{100(2|x_0| + 1)}, \frac{\varepsilon}{100(2|x_0| + 1)}\right\}.$$

Let $x \in \mathbb{Q}$ be such that $|x - x_0| < \delta$. First we show that $x^2 > 2$.

$$\begin{aligned} x^2 &= x_0^2 + (x - x_0)(x + x_0) \geq 100(2|x_0| + 1)\delta + 2 + (x - x_0)(x + x_0) \\ &\geq 100(2|x_0| + 1)\delta + 2 - \delta(|x| + |x_0|) \geq 100(2|x_0| + 1)\delta + 2 - \delta(2|x_0| + \delta) \\ &\geq 100(2|x_0| + 1)\delta + 2 - \delta(2|x_0| + 1) \geq 100(2|x_0| + 1)\delta + 2 - \delta(2|x_0| + 1) \\ &> 2. \end{aligned}$$

Hence, we have $g(x) = x^2$. On the other hand, the usual proof works to show continuity in this regime:

$$|x^2 - x_0^2| = |(x - x_0)(x + x_0)| \leq \delta(2|x_0| + 1) \leq \frac{\varepsilon}{100(2|x_0| + 1)}(2|x_0| + 1) < \varepsilon.$$

Hence g is continuous.

Now let us see if there is $\tilde{g} : \mathbb{R} \to \mathbb{R}$ that is continuous and agrees with g on $\mathbb{Q}$. Suppose this were true. Take $x_n \searrow \sqrt{2}$ and $y_n \nearrow \sqrt{2}$ such that $x_n, y_n \in \mathbb{Q}$. Then, by continuity and the fact that $\tilde{g}(x_n) = g(x_n)$ and $\tilde{g}(y_n) = g(y_n)$, by assumption, we have

$$\tilde{g}(\sqrt{2}) = \lim_{n\to\infty} \tilde{g}(x_n) = \lim_{n\to\infty} g(x_n) = \lim_{n\to\infty} x_n^2 = 2,$$

and

$$\tilde{g}(\sqrt{2}) = \lim_{n\to\infty} \tilde{g}(y_n) = \lim_{n\to\infty} g(y_n) = \lim_{n\to\infty} 0 = 0.$$

This is clearly a contradiction, so no such $\tilde{g}$ can exist.

What is the difference between these two cases? Why can f be extended but g cannot be? The difference has to do with *uniform continuity*. Notice that, on any bounded set $(-a, a) \cap \mathbb{Q}$, f is uniformly continuous since, given any $\varepsilon > 0$, the choice

$$\delta = \max\left\{1, \frac{\varepsilon}{100a}\right\}$$

which does not depend on x_0, suffices to show that $|f(x) - f(x_0)| < \varepsilon$ (this can be shown similar to the computations in step (ii), above). On the other hand, g is *not* uniformly continuous on $(-a, a)$ whenever $a > \sqrt{2}$. We can, roughly[12], see this since the choice of δ in the steps (i) and (ii) above depends on the distance between x_0^2 and 2.

3.1.2. **Topology in metric spaces.** Let (X, d) be a metric space throughout this section. We introduce a few fundamental concepts.

Definition 3.1.2. *The* metric open ball $B_r(x)$ *of radius r around x is the set of all points less than distance r from x:*

$$B_r(x) := \{y \in X : d(x, y) < r\}.$$

A ball is the quintessential open set and the basis for all other open sets.

Definition 3.1.3. *A set $\Omega \subset X$ is open if, for every point $\omega \in \Omega$, there is a radius $r_\omega > 0$ such that $B_r(\omega)$. We denote the set of open subsets of X by $\mathcal{T}$.*

Example 3.1.4. *(i) $(X, d) = (\mathbb{R}, |\cdot|)$. Then the open balls are $B_r(x) = (x - r, x + r)$. Some open sets are those of the form*

$$(a, b), \qquad (-\infty, a), \qquad \bigcup_{i\in\mathbb{Z}} (i, i + \tfrac{1}{1+|i|}) \qquad \mathbb{R}, \quad \textit{and} \qquad \emptyset.$$

[12]This is not a proof! Perhaps we just "did a bad job" of choosing δ and there is a better δ that we could have chosen.

Exercise 3.1.2. *Prove that g is not uniformly continuous on $(-2, 2)$.*

The last example vacuously *satisfies Definition 3.1.3 – there is* no $x \in \emptyset$ *to test the condition* $B_r(x) \subset \emptyset$ *so it is true by default. Let us check the first example. Fix any* $x \in (a,b)$*. Then* $a < x < b$*, so we can set*

$$r = \min\{x-a, b-x\} > 0.$$

Then $B_r(x) = (x-r, x+r) \subset (a,b)$*. Why? If* $y \in (x-r, x+r)$*, we have*

$$y > x - r \geq x - (x-a) = a \quad \text{and} \quad y < x + r \leq x + (b-x) = b.$$

Some non-open sets are those of the form

$$[a,b], \qquad (a,b], \qquad [a,b), \qquad \mathbb{Q}, \qquad \mathbb{Z}, \quad \text{and} \quad \{a\}.$$

Let us think about the last example. Why is this not open? Take any $r > 0$ *and look at* $B_r(a)$*. Is* $B_r(a) \subset \{a\}$*? No! For example* $a + r/2 \in B_r(a)$ *but* $a + r/2 \notin \{a\}$*.*

Exercise 3.1.3. *Prove the openness or non-openness of all the examples above.*

(ii) (X, d_{disc})*. Every set is open. Indeed, fix any set* $S \subset X$ *and any* $s \in S$*. Let* $r = 1/2$*. Then* $B_r(s) = \{s\} \subset S$*. Hence,* S *is open.*

(iii) $(\ell^p, \|\cdot\|_{\ell^p})$ *with* $p \in [1, \infty]$*. Let*

$$S = \{\bar{x} : |x_1| < 1\}.$$

This is open. Indeed, fix any $\bar{x} \in S$ *and let* $r = 1 - |x_1| > 0$*. Then, if* $\|\bar{y} - \bar{x}\|_p < r$*, we have*

$$|y_1 - x_1| \leq \|\bar{y} - \bar{x}\|_{\ell^p} < r.$$

Hence,

$$|y_1| \leq |y_1 - x_1| + |x_1| < r + |x_1| = (1 - |x_1|) + |x_1| = 1.$$

In other words, $\bar{y} \in S$*. As we just showed that* $B_r(\bar{x}) \subset S$*, we conclude that* S *is open.*

What about the set $c_0 \subset \ell^\infty$*? Is this open? No! Take any element* $\bar{x}$ *and any* $r > 0$*. Since* $\bar{x} \in c_0$*, there is* N *so that, for all* $n \geq N$*,* $|x_n| < r/3$*. Define a new point* $\bar{y}$ *by*

$$y_n = \begin{cases} x_n & \text{if } n < N, \\ \frac{(-1)^n r}{3} & \text{if } n \geq N. \end{cases}$$

Then, we have that $x_n - y_n = 0$ *if* $n < N$*, and* $|x_n - y_n| < |x_n| + |y_n| < r/3 + r/3 < r$ *for all* $n \geq N$*. Hence*

$$\|\bar{x} - \bar{y}\|_{\ell^\infty} < r.$$

On the other hand, clearly $\bar{y} \notin c_0$*. Hence,* $\bar{y} \in B_r(\bar{x})$ *but* $\bar{y} \notin c_0$*. We conclude that* c_0 *is not open.*

Exercise 3.1.4. *What about* c*? Is* c *an open subset of* ℓ^∞*?*

Exercise 3.1.5. *Show that if* $U_1, U_2, \ldots$ *are open subsets of* X *then so is*

$$\bigcup_{i=1}^{\infty} U_i = \{u \in X : \text{ there exists } i \in \mathbb{N} \text{ such that } u \in U_i\}.$$

If you are feeling up to it, let $\mathcal{I}$ *be any indexing set such that, for every* $\iota \in \mathcal{I}$*, there is an associated open set* U_ι*. Then show that*

$$\bigcup_{\iota \in \mathcal{I}} U_\iota = \{u \in X : \text{ there exists } \iota \in \mathcal{I} \text{ such that } u \in U_\iota\}.$$

Hint: this is essentially the same as the previous case, where $\mathcal{I} = \mathbb{N}$*.*

Exercise 3.1.6. *Show that, for any metric space (X, d), any point $x \in X$, and any $r > 0$, $B_r(x)$ is open.*

Every set, open or not, has a "largest" open subset that we call the interior.

Definition 3.1.5. *For any $S \subset X$, the interior of S, denoted*[13] $\operatorname{Int}(S)$ *or S°, is the set of all points s such that there is $r_s > 0$ such that $B_{r_s}(s) \subset S$. In other words*

$$\operatorname{Int} S = \{s \in S : \exists r_s > 0 \text{ such that } B_{r_s}(s) \subset S\}.$$

Exercise 3.1.7. *For any $S \subset X$, show that $\operatorname{Int} S$ is the largest open subset; that is, for any open set $U \subset X$, then $U \subset \operatorname{Int} S$. Deduce that*

$$\operatorname{Int} S = \bigcup_{U \in \mathcal{T}, U \subset S} U.$$

Example 3.1.6. *Consider the set $[3, 7] \subset \mathbb{R}$ with the Euclidean metric. We claim that*

$$\operatorname{Int}[3, 7] = (3, 7). \tag{3.1.2}$$

First, by definition, we have that

$$\operatorname{Int}[3, 7] \subset [3, 7]. \tag{3.1.3}$$

We first claim that

$$(3, 7) \subset \operatorname{Int}[3, 7].$$

Indeed, take $x \in (3, 7)$ and, since $(3, 7)$ is open, there is $r > 0$ such that

$$B_r(x) \subset (3, 7) \subset [3, 7].$$

Hence $x \in \operatorname{Int}[3, 7]$. By the arbitrariness of x, we have

$$(3, 7) \subset \operatorname{Int}[3, 7]. \tag{3.1.4}$$

Now, by (3.1.3) and (3.1.4), we need only show that $3, 7 \notin \operatorname{Int}[3, 7]$ to conclude (3.1.2). We show that $3 \notin \operatorname{Int}[3, 7]$, but a similar proof establishes the analogous claim for 7. Take any $r > 0$, and notice that

$$3 - \frac{r}{2} \in B_r(3) \quad \text{and} \quad 3 - \frac{r}{2} \notin [3, 7].$$

Hence $B_r(3) \not\subset [3, 7]$. By the arbitrariness of r, it follows that no such ball exists that is a subset of $[3, 7]$. Hence, $3 \notin \operatorname{Int}[3, 7]$.

We can also define density in a slightly different way:

Definition 3.1.7. *A set $D \subset X$ is dense if, for every $x \in X$ and $r > 0$,*

$$D \cap B_r(x) \neq \emptyset.$$

Note that X is trivially always dense in X.

[13] I will always denote this as $\operatorname{Int} S$ because I prefer notation that is self-explanatory (both for my own sake and because I think this makes it easier for other readers).

Exercise 3.1.8. *Fix any open set $U \subset X$ such that $U \neq \emptyset$ and $U \neq X$. Can U be dense in X? Can the complement of U be dense in X? The complement is*

$$X \setminus U := U^c := \{x \in X : x \notin U\}.$$

Let us go over these concepts with some examples.

Example 3.1.8. *(i) $\mathbb{Q}$ is dense in $\mathbb{R}$: Fix any $x \in \mathbb{R}$ and $r > 0$, and we must show that there is $q \in \mathbb{Q}$ such that $|x - q| < r$. Let*

$$N > \frac{1}{r}. \tag{3.1.5}$$

Let

$$k = \max\{\ell \in \mathbb{Z} : \ell/N < x\}. \tag{3.1.6}$$

This is well defined because $\ell/N \to \infty$ as $\ell \to \infty$. Moreover, by the choice of N (3.1.6), we have that $k \geq 1$. Finally, by (3.1.6), we have

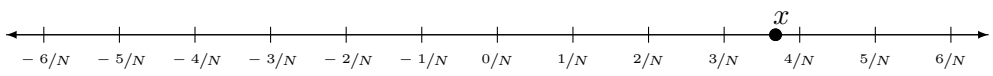
$$\frac{k}{N} < x \leq \frac{k+1}{N}. \tag{3.1.7}$$

Thus,

$$|x - {}^{k}/_{N}| = x - \frac{k}{N} \leq \frac{k+1}{N} - \frac{k}{N} = \frac{1}{N} < r,$$

where the first equality and the first inequality follows from (3.1.7) and the last inequality uses the choice of N (3.1.5). This completes the proof of the density of $\mathbb{Q}$ with the choice $q = {}^{k}/_{N}$.

What is the intuition here? Basically, we created a 'grid' of all the rational elements of the form ℓ/N. This grid ranges from $-\infty$ to ∞ and every point of it is $1/N$ apart. Hence, x has to be between two gridpoints, which implies that it is within $1/N$ of the two rational numbers defining these gridpoints.



(ii) $\mathbb{Q}^n$ is dense in $\mathbb{R}^n$ (with the metric d_p for any $p \in [1, \infty]$): Let us show this for $p = 1$, that is, the usual Euclidean metric. Fix any $\bar{x} \in \mathbb{R}^n$ and any $r > 0$. For each i, use the previous example to find $q_i \in \mathbb{Q}$ such that

$$|x_i - q_i| < \frac{r}{n}.$$

Let $\bar{q} = (q_1, q_2, \ldots, q_n) \in \mathbb{Q}^n$. Then

$$d(\bar{x}, \bar{q}) = \sum_{i=1}^{n} |x_i - q_i| < \sum_{i=1}^{n} \frac{r}{n} = r.$$

Hence $\bar{q} \in B_r(\bar{x})$, as desired. This completes the proof that $\mathbb{Q}^n$ is dense in $\mathbb{R}^n$.

Exercise 3.1.9. *Show that each metric d_p is equivalent to d_1, and use this to show that $\mathbb{Q}^n$ is dense in $\mathbb{R}^n$ for each $p \in [1, \infty]$.*

(iii) *The only dense subset of (X, d_{disc}) is X. Clearly X is dense in X. Hence, we need only show that any other dense set is actually X. Let $D \subset X$ be dense. Pick any $x \in X$. Since D is dense in X, there must be $\theta \in D$ such that $\theta \in D \cap B_{1/2}(x)$. Hence, $d_{\text{disk}}(x, \theta) < 1/2$. By definition of the discrete metric, it follows that $x = d$. Hence, $x \in D$.*

We started with an arbitrary point in X and showed that it was a point in D. It follows that $X \subset D$. On the other hand, we have, by assumption, that $D \subset X$. Hence, $X = D$.

(iv) *For $p \in [1, \infty)$. Take the set*

$$X = \{\bar{x} \in \ell^p : x_i \in \mathbb{Q} \text{ for all } i \text{ and } x_n = 0 \text{ for all but finitely many } i \in \mathbb{N}\}.$$

Roughly, $\bar{x}$ are all elements of ℓ^p such that all entries are rational and there is N such that $x_n = 0$ for all $n \geq N$.

We claim that X is a dense subset of ℓ^p. Fix any point $\bar{z} \in \ell^p$ and any $r > 0$. The proof is complete if we can find $\bar{x}$ such that

$$\bar{x} \in X \cap B_r(\bar{z}). \tag{3.1.8}$$

We define $\bar{x} = (x_1, x_2, \dots)$ as follows. Fix $N > 0$ so that

$$\sum_{i=N+1}^{\infty} |z_i|^p \leq \Big(\frac{r}{100}\Big)^p.$$

For $i > N$, let $x_i = 0$. By the density of $\mathbb{Q}$ in $\mathbb{R}$, we can choose, for all $i \in \{1, 2, 3, \dots, N\}$, $x_i \in \mathbb{Q}$ such that

$$|z_i - x_i| \leq \frac{r}{100(1+N)^{1/p}}.$$

Hence, we have defined $\bar{x} = (x_1, x_2, \dots)$. Then

$$\begin{aligned} \|\bar{z} - \bar{x}\|_{\ell^p}^p &= \sum_{i=1}^{N} |z_i - x_i|^p + \sum_{i=N+1}^{\infty} |z_i|^p < \sum_{i=1}^{N} \frac{r^p}{100^p(1+N)} + \Big(\frac{r}{100}\Big)^p \\ &= \frac{r^p N}{100^p(1+N)} + \Big(\frac{r}{100}\Big)^p < \frac{2r^p}{100^p} < r^p. \end{aligned}$$

Hence, (3.1.8) *holds.*

Exercise 3.1.10. *When $p = \infty$, is Example 3.1.8.*((ii)) *still true? That is, is X a dense subset of ℓ^∞?*

Exercise 3.1.11. *Show that $\mathbb{R} \setminus \mathbb{Q}$ is dense in $\mathbb{R}$. You may use that $\pi \in \mathbb{R} \setminus \mathbb{Q}$, and you may find it helpful to show that $\pi - q \in \mathbb{R} \setminus \mathbb{Q}$ for every $q \in \mathbb{Q}$.*

Before we conclude with this quick foray into metric spaces, let us mention two last characterizations of dense sets.

Proposition 3.1.9. *Let $Y \subset X$, where (X, d) is a metric space. The following are equivalent:*

(i) *Y is dense;*

(ii) *for every $x \in X$, there is a sequence $y_1, y_2, y_3, \dots \in Y$ such that*

$$\lim_{n \to \infty} y_n = x;$$

(iii) *for every open set U, we have $U \cap Y \neq \emptyset$.*

3.2. Countability and uncountability.

3.2.1. **How many rational numbers are there anyways!?.** Of course, the answer is there are an infinite number. But, how does the 'size' of $\mathbb{Q}$ relate to other infinite sets we know, like

$$\mathbb{N}, \quad \mathbb{Z}, \quad \mathbb{R}, \quad \mathbb{R}^n, \quad \text{and} \quad \ell^\infty?$$

Well, it depends on how you quantify the size of a set. Here, we will look at cardinality. For finite sets, this is easy to define: it is the number of elements in the set. For example,

$$\begin{aligned} &\operatorname{card}(\{1, 2, 277, 469, \pi\}) = 5, \qquad \operatorname{card}(\{\Delta, \square, \circ\}) = 3, \\ &\text{and} \quad \operatorname{card}(\{x \in \mathbb{R} : x^3 - x = 0\}) = 3 \end{aligned} \tag{3.2.1}$$

Of course, when we can count, it is pretty easy to say which sets are the same size. How do we formalize the process of counting? When we count a bunch of objects, we point our finger at each, in turn, and say 1 then 2 and so on. In essence, what we are doing here is associating, to every element of the set, a unique natural number. Hence, we say that a set A has n elements if we can *uniquely* associate to *every* element a of A a natural number $i_a \in \{1, 2, \dots, n\}$. The above procedure sounds an awful lot like defining functions!

This leads us to following preliminary definition.

Definition 3.2.1. *We say that a set A has n elements if there is a bijection $\varphi : \{1, \dots, n\} \to A$.*

We say that a set A is finite if either $A = \emptyset$ (in which case we say that it has 0 elements) or A has n elements for some $n \in \mathbb{N}$.

OK, we have used bijections before, but for completeness let us write down the definition.

Definition 3.2.2. *A function $f : A \to B$ is a bijection if it is an injection (one-to-one) and a surjection (onto).*

- *f is an injection if, for every $x, y \in A$, $f(x) = f(y)$ if and only if $x = y$.*
- *f is a surjection if, for every $b \in B$, there exists $a_b \in A$ such that $f(a_b) = b$.*

In the example above, we are essentially saying that A has n elements because A and $\{1, \dots, n\}$ have the same number of elements (cardinality), and we know this because the bijection between them associates to every element of one set a unique element of the other set. This way of thinking that two sets have the same cardinality does not really require finiteness! We formalize it in the following:

Definition 3.2.3. *We say that* $\operatorname{card}(A) = \operatorname{card}(B)$ *if there exists a bijection $\varphi : A \to B$.*

Of course, using equality here only makes sense if it is symmetric. This is guaranteed in the following result.

Proposition 3.2.4. *If $\varphi : A \to B$ is a bijection, then there is a bijection $\psi : B \to A$. As a consequence, if* $\operatorname{card}(A) = \operatorname{card}(B)$*, then* $\operatorname{card}(B) = \operatorname{card}(A)$.

Proof. It is clear that the second statement follows from the first, so we only prove the first. We define $\psi : B \to A$ as follows. Fix any $b \in B$ and, by the surjectivity of φ, there is a_b such that $\varphi(a_b) = b$. Let $\psi(b) = a_b$.

Let us check that ψ is well-defined (i.e., that our description above determines a ψ that is a function). For this, we need only show that, for any b, if $\varphi(a_b) = b$ and $\varphi(\tilde{a}_b) = b$ (making both a_b and $\tilde{a}_b$ potential "outputs" for $\psi(b)$) then $a_b = \tilde{a}_b$. Since φ is injective, $\varphi(a_b) = b = \varphi(\tilde{a}_b)$ implies that $a_b = \tilde{a}_b$, as desired. Thus, ψ is well-defined.

Let us point out that our construction of ψ yields

$$\varphi(\psi(b)) = \varphi(a_b) = b \quad \text{and} \quad \psi(\varphi(a)) = a. \tag{3.2.2}$$

The last equality holds because, by construction $a = a_{\varphi(a)}$. Note that this means that ψ is the inverse function of φ.

We check that ψ is injective. Suppose that $\psi(b) = \psi(\tilde{b})$. Using the first identity in (3.2.2), we see that

$$b = \varphi(\psi(b)) = \varphi(\psi(\tilde{b})) = \tilde{b},$$

Thus, ψ is injective.

We check that ψ is surjective. Fix any $a \in A$. Let $b = \varphi(a)$. Then, by the second identity in (3.2.2), we have

$$\psi(b) = \psi(\varphi(a)) = a.$$

Thus, ψ is surjective.

We conclude that $\psi : B \to A$ is a bijection, which concludes the proof. □

Exercise 3.2.1. *If $f : A \to B$ is a bijection, then there is a bijection $g : B \to A$ such that $f \circ g(b) = b$ for all $b \in B$ and $g \circ f(a) = a$ for all $a \in A$.*

Exercise 3.2.2. *If $f : A \to B$ and $g : B \to C$ are...*

(i) ...surjections, then so is $g \circ f : A \to C$.

(ii) ...injections, then so is $g \circ f : A \to C$.

(iii) ...bijections, then so is $g \circ f : A \to C$.

Notice that there is a bijection between the two three element sets in (3.2.1): let $f : \{\Delta, \square, \circ\} \to \{x \in \mathbb{R} : x^5 + 8x^4 - 44x^3 - 240x^2 = 0\}$ be defined by

$$f(\Delta) = 0, \quad f(\square) = -1, \quad \text{and} \quad f(\circ) = 1.$$

This is not, however, any bijection $f : \{1, 2, 277, 469, \pi\} \to \{\Delta, \square, \circ\}$, due to the Pigeonhole principle (By injectivity, $f(1)$, $f(2)$, and $f(277)$ are all distinct and, hence, must be equal to, in some order, Δ, $\square$, and $\circ$. Since $f(469)$ will be equal to one of Δ, $\square$, or $\circ$, injectivity is violated).

How can we use this approach to define an infinite set?

Definition 3.2.5. *A set A is infinite if, for every n, there is an injection $\varphi_n : \{1, \ldots, n\} \to A$.*

Exercise 3.2.3. *Let A be any set. The following are equivalent:*

(i) A *is infinite;*

(ii) there is $B \subsetneq A$ *and a bijection* $\varphi_B : B \to A$*;*

(iii) A *is not finite.*

It is pretty clear that any finite set does not satisfy this condition since bijections preserve the number of elements in a set. But let us see this in action in some infinite sets that we are used to...

Example 3.2.6. *(i)* $\mathbb{Z}$ *– Notice that* $f : \mathbb{Z} \to 2\mathbb{Z}$*, defined by* $f(z) = 2z$ *for all* $z \in \mathbb{Z}$*, is a bijection (here* $2\mathbb{Z}$ *is the set of even integers).*

(ii) $\mathbb{Z}$ *– Notice that* $f : \mathbb{Z} \to \mathbb{N}$ *defined by*

$$f(z) = \begin{cases} 2z - 1 & \text{if } z > 0, \\ -2z & \text{if } z \leq 0, \end{cases}$$

is a bijection.

(iii) $\mathbb{R}$ *– Notice that* $\varphi : \mathbb{R} \to (0, 1)$ *defined by* $\varphi(x) = e^x/(1 + e^x)$ *for all* $x \in \mathbb{R}$*, is a bijection. The injectivity is because* $\varphi' > 0$ *(it is strictly increasing so that* $x \neq y$ *implies* $\varphi(x) \neq \varphi(y)$*), and the surjectivity is due to the intermediate value theorem, the continuity of* φ*, and the fact that*

$$\lim_{x \to -\infty} \varphi(x) = 0 \quad \text{and} \quad \lim_{x \to \infty} \varphi(x) = 1.$$

(iv) ℓ^∞ *– Define the shift operator* $S : \ell^\infty \to Z$*, where*

$$Z = \{\bar{x} \in \ell^\infty : x_1 = 0\},$$

that is defined by

$$S\bar{x} = (0, x_1, x_2, x_3, \dots).$$

This is a bijection, but $Z \subsetneq \ell^\infty$*.*

With all of these tools in hand, one can ask all sorts of questions[14], for our purposes, though, we will focus only on the distinction between countable and uncountable sets.

Definition 3.2.7. *A set* A *is countable if* $\text{card}(\mathbb{N}) = \text{card}(A)$ *and it is uncountable if* A *is infinite it is not countable.*

Sometimes this is easy to do: above we showed that $\text{card}(\mathbb{N}) = \text{card}(\mathbb{Z}) = \text{card}(2\mathbb{Z})$ so all three sets are countable. Sometimes it is difficult... An easier criteria is the following:

Theorem 3.2.8 (Cantor-Schröder-Bernstein)**.** *Two sets* A *and* B *have the same cardinality if and only if there are injections* $f : A \to B$ *and* $g : B \to A$*.*

[14]There has been a whole world of research on these questions – e.g., are cardinalities discrete? In particular, we'll show that, in some sense, $\text{card}(\mathbb{N}) < \text{card}(\mathbb{R})$... but is there any set A such that $\text{card}(\mathbb{N}) < \text{card}(A) < \text{card}(\mathbb{R})$? It turns out this 'continuum hypothesis' is independent of mathematics under the standard axioms we use (ZFC); that is, it can be neither proven nor disproven.

This proof is extremely long, and the main ideas, while elegant, are not useful to the remainder of this course. Hence, we relegate its proof to Appendix A, where only the very interested and intrepid reader will find it.

The intuition here is that, roughly, $\text{card}(A) \leq \text{card}(B)$ when there is an injection $f : A \to B$ (otherwise A would not 'fit' into B). Hence, if there are injections in both directions, we should think of $\text{card}\, A \leq \text{card}\, B$ and $\text{card}\, B \leq \text{card}\, A$. Then Theorem 3.2.8 tells us that this $\leq$ intuition is reasonable. Actually, the same can be said with surjections: if $f : A \to B$ is a surjection then $\text{card}\, A \geq \text{card}\, A$.

Exercise 3.2.4. *Show that, in this case, there must be an injection $g : B \to A$.*

Example 3.2.9. *$\mathbb{Q}$ is countable – We will show this in a few steps.*

Step one: $\mathbb{N} \times \mathbb{N}$ is countable. *It is easy to see that $f : \mathbb{N} \to \mathbb{N} \times \mathbb{N}$, defined by $f(n) = (1, n)$ is an injection. Hence, we need only find an injection $g : \mathbb{N} \times \mathbb{N} \to \mathbb{N}$. One way to define this is the following:*

$$g(n, m) = 2^n \cdot 3^m.$$

This is an injection because prime factorizations are unique – if $2^n \cdot 3^m = 2^{\tilde{n}} \cdot 3^{\tilde{m}}$ then $n = \tilde{n}$ and $m = \tilde{m}$. Thus, by Theorem 3.2.8, $\mathbb{N} \times \mathbb{N}$ is countable.

Step two: $\mathbb{Z} \times \mathbb{N}$ is countable. *We already showed that $\mathbb{Z}$ is countable. Hence, there is $\psi : \mathbb{Z} \to \mathbb{N}$ that is a bijection. It follows that*

$$\Psi : \mathbb{Z} \times \mathbb{N} \to \mathbb{N} \times \mathbb{N},$$

defined by $\Psi(z, n) = (\psi(z), n)$, is a bijection. Hence, $\text{card}(\mathbb{Z} \times \mathbb{N}) = \text{card}(\mathbb{N} \times \mathbb{N}) = \text{card}(\mathbb{N})$.

Step three: $\mathbb{Q}$ is countable. *First note that*

$$\iota : \mathbb{N} \to \mathbb{Q}$$

defined by $\iota(n) = n$ is clearly an injection. Hence, we need only find an injection from $\mathbb{Q}$ into $\mathbb{N}$. Let $\rho : \mathbb{Z} \times \mathbb{N} \to \mathbb{N}$ be a bijection (this exists due to the work in Step two). Then we define

$$f : \mathbb{Q} \to \mathbb{Z} \times \mathbb{N}$$

as follows: for $q \in \mathbb{Q}$, there is a unique way to write $q = {}^{a_q}/_{b_q}$, where $a_q \in \mathbb{Z}$ and $b_q \in \mathbb{N}$, in lowest terms[15]*. Using this representation, we let*

$$f(q) = f({}^{a_q}/_{b_q}) = (a_q, b_q).$$

By the uniqueness of the representation $q = a_q/b_q$, this map is injective.

Exercise 3.2.5. *Is it surjective?*

Then we note that $\rho \circ f : \mathbb{Q} \to \mathbb{N}$ is an injection.

Exercise 3.2.6. *Check this!*

Applying Theorem 3.2.8, we see that $\mathbb{Q}$ is countable.

Exercise 3.2.7. *Show the following sets are countable:*

[15]'Lowest terms' means that there is no element $d \in \mathbb{N} \setminus \{1\}$ such that $b_q/d \in \mathbb{N}$ and $a_q/d \in \mathbb{Z}$.

(i) $\mathbb{N}^n$ *for any* n*;*

(ii) $\mathbb{Z}^n$ *for any* n*;*

(iii) $\mathbb{Q}^n$ *for any* n*;*

(iv) X *defined in Example 3.1.8.*((ii))*;*

(v) the set of all polynomials with coefficients in $\mathbb{Q}$*;*

(vi) the set of finite subsets of $\mathbb{N}$*.*

Next, we show understand how countability interacts with unions and complements.

Lemma 3.2.10. *Suppose that* B *is a finite or countable set. Fix any other set* A*.*

(i) Suppose that A *is countable. Then so is* $A \cup B$*.*

(ii) Suppose that A *is uncountable. Then* $A \setminus B$ *is uncountable.*

(iii) Suppose that A *is uncountable. Then* $\operatorname{card}(A) = \operatorname{card}(A \setminus B)$*.*

Proof of (i). Let us take B to be countable; however, it is easy to modify the proof if B is finite. By definition, we have bijections

$$\varphi : A \to \mathbb{N} \quad \text{and} \quad \psi : B \to \mathbb{N}.$$

Let us apply Theorem 3.2.8, as usual. We need only demonstrate the existence of injections from $A \cup B$ to $\mathbb{N}$ and from $\mathbb{N}$ to $A \cup B$.

We begin with the easy direction. Since φ is a bijection, so is $\varphi^{-1} : \mathbb{N} \to A$. Define the usual inclusion map

$$\iota : A \to A \cup B,$$

by $\iota(a) = a$. This is injective. Hence, by Exercise 3.2.2, we have that

$$\iota \circ \varphi^{-1} : \mathbb{N} \to A \cup B$$

is an injection.

We now proceed with the more difficult direction. Fix any bijection

$$\eta : \mathbb{N} \times \mathbb{N} \to \mathbb{N},$$

which exists by Example 3.2.9. Then, define

$$f : A \cup B \to \mathbb{N} \times \mathbb{N},$$

by

$$f(x) = \begin{cases} (\varphi(x), 1) & \text{if } x \in A, \\ (1, 1 + \psi(x)) & \text{if } x \in B \setminus A. \end{cases}$$

Exercise 3.2.8. *Show that* f *is injective.*

Hence, we have, again by Exercise 3.2.2, that

$$\eta \circ f : A \cup B \to \mathbb{N}$$

is injective. This concludes the proof. □

Proof of (ii). Let us show the uncountability of $A \setminus B$ first. We show this by establishing the contrapositive. Suppose that it is countable and we show that A is then countable. Then, by Lemma 3.2.10.((ii)), we have

$$A = (A \setminus B) \cup B$$

is countable. □

Proof of (iii). Let us again assume that B is countable, as this is the trickier claim to prove. If B is finite, the modifications are straightforward. Again, let

$$\psi : B \to \mathbb{N}$$

be a bijection.

Since $A \setminus B$ is uncountable, it is nonempty. Hence, there is

$$c_1 \in A \setminus B.$$

Applying Lemma 3.2.10.((ii)), we have that $A \setminus (B \cup \{c_1\})$ is uncountable and, thus, nonempty. Choose

$$c_2 \in A \setminus (B \cup \{c_1\}).$$

Continuing in this way, by selecting, for each n,

$$c_n \in A \setminus (B \cup \{c_1, \ldots, c_{n-1}\})$$

we find points

$$c_1, c_2, c_3, \cdots \in A \setminus B \quad \text{such that } c_i \neq c_j$$

whenever $i \neq j$. Define

$$\varphi : A \to A \setminus B$$

as follows. Fix any $x \in A$. If

$$x \in A \setminus (B \cup \{c_1, c_2, \ldots\}),$$

then let

$$\varphi(x) = x.$$

If $x = c_i$ for some i, let

$$\varphi(x) = c_{2i-1}.$$

If $x \in B$, let $i = \psi^{-1}(x)$ and define

$$\varphi(x) = c_{2i}.$$

Exercise 3.2.9. *The function φ is bijective.*

□

3.2.2. **How many real numbers are there?.** What about uncountable sets? Let us give Georg Cantor's "diagonalization" argument to show that $\mathbb{R}$ is uncountable. First, we consider a case that, while essentially the same, allows us to see the argument a little bit more cleanly. Let us stress that this is the heart of the argument that $\mathbb{R}$ is uncountable – nearly everything else is "technical wrangling."

Proposition 3.2.11. *The set $\{0,1\}^{\mathbb{N}}$ is uncountable.*

Proof. First, we observe that $\{0,1\}^{\mathbb{N}}$ is clearly infinite. Hence, we may proceed by contradiction, assuming that $\{0,1\}^{\mathbb{N}}$ is countable and that there is, thus, a bijection

$$\varphi : \mathbb{N} \to \{0,1\}^{\mathbb{N}}.$$

For each n, let $f_n = \varphi(n)$. It follows that

$$\{0,1\}^{\mathbb{N}} = \{f_1, f_2, f_3, \dots\}; \tag{3.2.3}$$

that is, if $f \in \{0,1\}^{\mathbb{N}}$ then $f = f_i$ for some $i \in \mathbb{N}$.

We now construct an $f \in \{0,1\}^{\mathbb{N}}$ such that this is not the case. To guarantee that $f \neq f_n$, it is natural to make them differ in how they evaluate n; that is, $f(n) \neq f_n(n)$. Thus, we define f as follows: for $n \in \mathbb{N}$, let

$$f(n) = \begin{cases} 0 & \text{if } f_n(n) = 1, \\ 1 & \text{if } f_n(n) = 0. \end{cases}$$

Notice that $f(n) \neq f_n(n)$, which implies that $f \neq f_n$. We deduce that $f \in \{0,1\}^{\mathbb{N}}$, but $f \neq f_i$ for any i, which contradicts (3.2.3). This concludes the proof. □

Let us now begin the slog of upgrading the uncountability of $\{0,1\}^{\mathbb{N}}$ to that of $\mathbb{R}$. First, we show that $\{0,1\}^{\mathbb{N}}$ is "mostly" made up of sequences that do not terminate in repeating decimals. For shorthand, let us temporarily use

$$S = \{0,1\}^{\mathbb{N}}.$$

Recall that we showed that S is uncountable.

Lemma 3.2.12. *Let S_{repeat} be the set of all sequences in S that end in repeated 1's:*

$$S_{\text{repeat}} = \{\bar{b} \in S : \exists N \text{ such that } b_n = 1 \text{ for all } n \geq N\}.$$

Then S_{repeat} is countable.

Proof. As usual, we use Theorem 3.2.8. Notice that $f : \mathbb{N} \to S_{\text{repeat}}$ is easy to construct:

$$f(n) = (\underbrace{0, \dots, 0}_{n \text{ places}}, 1, 1, \dots).$$

This is clearly an injection.

Now we establish an injection $g : S_{\text{repeat}} \to \mathbb{N}$. Define g by, for any $\bar{x}$,

$$g(\bar{x}) = \sum_{i=1}^{\infty} |x_i - 1| 2^i.$$

Notice that $|x_i - 1|$ is one if $x_i = 0$ and is 0 if $x_i = 1$. This is a compact (if opaque) way of writing the procedure of swapping all ones and zeros. Since $\bar{x}$ terminates in all ones, $g(\bar{x})$ is a finite sum and, thus, well-defined.

Let us show by contradiction that g is an injection. Let us argue by contradiction assuming that there are $\bar{x}, \bar{y} \in S_{\text{repeat}}$ such that $\bar{x} \neq \bar{y}$ and

$$g(\bar{x}) = g(\bar{y}).$$

Let

$$\ell = \max\{i : x_i \neq y_i\}.$$

Since $\bar{x}$ and $\bar{y}$ are not equal but terminate in all 1's, ℓ is well-defined. Then

$$\begin{aligned} 0 = g(\bar{x}) - g(\bar{y}) &= \sum_{i=1}^{\infty} |x_i - 1|2^i - \sum_{i=1}^{\infty} |y_i - 1|2^i \\ &= \sum_{i=1}^{\ell} |x_i - 1|2^i - \sum_{i=1}^{\ell} |y_i - 1|2^i. \end{aligned} \tag{3.2.4}$$

Without loss of generality, assume that $x_i = 0$ and $y_i = 1$. Also, notice that, for $i < \ell$,

$$||x_i - 1| - |y_i - 1|| \leq 1.$$

Then, from (3.2.4)

$$\begin{aligned} 2^\ell = |x_\ell - 1| - |y_\ell - 1| &= -\sum_{i=1}^{\ell-1} |x_i - 1|2^i + \sum_{i=1}^{\ell-1} |y_i - 1|2^i \\ &= \sum_{i=1}^{\ell-1} (|y_i - 1| - |x_i - 1|)2^i \leq \sum_{i=1}^{\ell-1} 2^i = 2^\ell - 2. \end{aligned}$$

This is clearly a contradiction, so g must be injective. This completes the proof. □

We now reduce the proof down to showing that intervals are uncountable.

Lemma 3.2.13. *Let $a < b$ and suppose that I is an interval of the form (a, b), $[a, b)$, $(a, b]$, or $[a, b]$. Then* $\text{card}(I) = \text{card}(\mathbb{R})$.

Proof. This holds due to Theorem 3.2.8. Indeed, $\iota : I \to \mathbb{R}$ defined by $\iota(x) = x$ is injective, and $\varphi : \mathbb{R} \to I$ defined by $\varphi(x) = a + (b - a)(e^x/(1 + e^x))$ is injective. □

Note that the injectivity of φ, above, is a consequence of the following exercise.

Exercise 3.2.10. *Fix any sets $A, B \subset \mathbb{R}$. Let $\varphi : A \to B$ be a strictly increasing function; that is, $\varphi(x_1) < \varphi(x_2)$ if $x_1 < x_2$. Show that any strictly increasing function in injective.*

We finally show that intervals are uncountable by showing that they have the same cardinality as $S = \{0, 1\}^{\mathbb{N}}$.

Lemma 3.2.14. *We have* $\text{card}(S) = \text{card}\{[0, 1)\}$. *Thus, $[0, 1)$ is uncountable.*

Proof. In view of Lemma 3.2.10 and Lemma 3.2.12, it is enough to show that

$$\operatorname{card}(S \setminus S_{\text{repeat}}) = \operatorname{card}([0,1)).$$

Any element $x \in [0,1)$ can be given by a unique binary expansion

$$x = \sum_{i=k}^{\infty} \frac{b_k}{2^k},$$

where $b_k \in \{0,1\}$ and where there is no N such that $b_k = 1$ for all $k \geq N$ (e.g., we take the binary expansion of $1/2$ to be $b_1 = 1$ and $b_k = 0$ for all $k > 1$ instead of the expansion $b_1 = 0$ and $b_k = 1$ for all $k > 1$). Thus, let

$$f : S \setminus S_{\text{repeat}} \to [0,1),$$

be defined by

$$f(\bar{x}) = \sum_{i=1}^{\infty} \frac{x_i}{2^i}.$$

It follows that f is a bijection.[16] This completes the proof. □

Theorem 3.2.15. *The real numbers $\mathbb{R}$ are not countable. Further, if $a < b$ then any interval I of the form (a,b), $[a,b)$, $(a,b]$, or $[a,b]$ is not countable.*

Proof. It follows from Proposition 3.2.11 that $[0,1)$ is not countable. By Lemma 3.2.13, $\mathbb{R}$, I, and $[0,1)$ all have the same cardinality. This completes the proof. □

Exercise 3.2.11. *Show that there is a bijection between $[0,1)$ and $\mathcal{P}^{\mathbb{N}}$. Recall that $\mathcal{P}^{\mathbb{N}} = \{A : A \subset \mathbb{N}\}$.*

As we show on the homework, $\operatorname{card} A \neq \operatorname{card} \mathcal{P}^A$, for any set A. Of course, it is easy to make an injection $A \to \mathcal{P}^A$ (indeed, take $a \mapsto \{a\}$), so we conclude that $\operatorname{card} A < \operatorname{card} \mathcal{P}^A$. This tells us that, not only are there different levels of infinity, there are an infinite number of different levels of infinity:

$$\operatorname{card} \mathbb{N}, \operatorname{card} \mathcal{P}^{\mathbb{N}}, \operatorname{card} \mathcal{P}^{\mathcal{P}^{\mathbb{N}}}, \operatorname{card} \mathcal{P}^{\mathcal{P}^{\mathcal{P}^N}}, \ldots$$

Actually, there is an interesting question called the **Continuum Hypothesis.**. The Continuum Hypothesis asserts that there is no cardinality between that of $\mathbb{N}$ and that of $\mathbb{R}$ (or, equivalently, that of $\mathcal{P}^{\mathbb{N}}$). Gödel proved that the Continuum Hypothesis cannot be disproved using the standard mathematical axioms (Zermelo-Frankael set theory with the Axiom of Choice) in 1940, while Cohen proved that the Continuum Hypothesis cannot be proved using standard mathematical axioms in 1963. Interestingly, both Gödel and Cohen did not believe that the Continuum Hypothesis was true.

Separability: Countable sets are useful because we can enumerate them! Usually, however, we do not have the luxury of working with countable sets (indeed, we are not number theorists and, thus, have to work *at least* $\mathbb{R}$). On the other hand, many useful spaces have dense countable sets. We give a name for these:

[16] Injectivity can be proved directly using the ideas in Lemma 3.2.12. Surjectivity is a bit harder; however, the main idea is, for a fixed $r \in [0,1)$, inductively define x_i as follows. If you have already picked $x_1, \ldots, x_{i-1}$, then let $x_i = 0$ if $r - (x_1/2 + \cdots + x_{i-1}/2^{i-1} < 2^{-i}$. Otherwise, let $x_i = 1$. One can then check that $f(\bar{x}) = r$.

Definition 3.2.16. *A metric space (X, d) is* separable *if there is a countable dense set $A \subset X$.*

Example 3.2.17. *(i) $\mathbb{R}$ is separable because $\mathbb{Q}$ is a countable dense subset of $\mathbb{R}$.*

(ii) $\mathbb{R}^n$ is separable because $\mathbb{Q}^n$ is a countable, dense subset of $\mathbb{R}^n$. The density of $\mathbb{Q}^n$ was established in Example 3.1.8.((ii))*. Here is a proof of countability.*

We show countability by induction with the case $n = 1$ already complete (see above). Hence, we assume the claim is true for $\mathbb{Q}^{n-1}$ and we use that to show it is true for $\mathbb{Q}^n$. Recall that we showed that $\mathbb{Q}$ and $\mathbb{N} \times \mathbb{N}$ are countable. Hence, there are bijections $f : \mathbb{Q}^{n-1} \to \mathbb{N}$, $g : \mathbb{N} \times \mathbb{N} \to \mathbb{N}$, and $h : \mathbb{Q} \to \mathbb{N}$. We define

$$\psi : \mathbb{Q}^n \to \mathbb{N} \qquad \text{by } \psi(\bar{q}, q_n) = g(f(\bar{q}), h(q_n)) \quad \text{for } \bar{q} \in \mathbb{Q}^{n-1}, q_n \in \mathbb{Q}.$$

Is this injective? If $\psi(\bar{q}, q_n) = \psi(\bar{p}, p_n)$, then, by the injectivity of g, we have $f(\bar{q}) = f(\bar{p})$ and $h(q_n) = h(p_n)$. By the injectivity of f and h, we conclude that $\bar{q} = \bar{p}$ and $q_n = p_n$; that is, $(\bar{q}, q_n) = (\bar{p}, p_n)$. We conclude that ψ is injective.

Is this surjective? Fix any $n \in \mathbb{N}$. Let $(n_1, n_2) \in \mathbb{N} \times \mathbb{N}$ be such that $g(n_1, n_2) = n$. This exists by the surjectivity of g. Then, by the surjectivity of f and h, we find $\bar{q} \in \mathbb{Q}^{n-1}$ and $q_n \in \mathbb{Q}$ such that $f(\bar{q}) = n_1$ and $h(q_n) = n_2$. It follows that $\psi(\bar{q}, q_n) = n$. We conclude that ψ is surjective.

Hence, $\text{card}(\mathbb{Q}^n) = \text{card}(\mathbb{N})$ *and the proof is complete.*

(iii) The space X from Example 3.1.8.((ii))* is a countable, dense subset of ℓ^p for $p \in [1, \infty)$. We showed the density above, while you showed the countability on your homework.*

(iv) By the Stone-Weierstrass theorem, the set of all polynomials with coefficients in $\mathbb{Q}$ is dense in $C([a, b])$ (for any $-\infty < a < b < \infty$ and with the $\|\cdot\|_\infty$-norm). By a homework problem, this set is countable.

(v) (Important example that we justify in the future) The set of all functions of the form:

$$\sum_{n=1}^{N} (a_n \sin(2\pi n x) + b_n \cos(2\pi n x)),$$

where $N \in \mathbb{N}$, $a_n, b_n \in \mathbb{Q}$, are dense in $L^2([0, 1])$.

3.3. **Complete Metric Spaces.** Let us return to our motivating questions: given a continuous function between metric spaces $f : X \to Y$, when can f extend to larger metric spaces $\bar{X}$ and $\bar{Y}$? Abstractly, we would like to "extend by density" in the following way:

1. given $\bar{x} \in \bar{X}$, pick a sequence $x_n \to \bar{x}$;
2. find the limit $\bar{y} = \lim_{n\to\infty} f(x_n)$;
3. let $\bar{f}(\bar{x}) = \bar{y}$.

The first step is the only one that is not fraught (assuming that X is dense in $\bar{X}$).

In the second step, why should $\bar{y}$ exist? Even if we expect that the x_n's converging means that the $f(x_n)$'s should be "close together," what if there is a hole that they are centered around? For

example, if we try to extend $f : \mathbb{Q} \to \mathbb{Q}$, defined by $f(x) = x^2$, to $\bar{f} : \mathbb{R} \to \mathbb{Q}$, we will have problems with $f(2^{1/4})$ since this "should" be $\sqrt{2}$. Hence need the codomain to be *complete*.

In the last step, our definition of $\bar{f}(\bar{x})$ depends on the choice of x_n... what if we chose a different sequence $\tilde{x}_n$? Hence, our $\bar{f}$ might not be well-defined. We need to know that, if the x_n's and $\tilde{x}_n$'s converge to the same point $\bar{x}$, then $f(x_n)$ and $f(\tilde{x}_n)$ converge to the same point. Since $\bar{x} \notin X$, continuity is not enough... we will need uniform continuity.

The first step to understanding this is the introduce two new concepts:

Definition 3.3.1. *A sequence $x_1, x_2, \dots$ is a* Cauchy sequence *if, for all $\varepsilon > 0$, there is N_ε so that*

$$d(x_n, x_m) < \varepsilon \qquad \text{whenever } n, m \geq N.$$

Exercise 3.3.1.

- *Show that any convergent sequence is a Cauchy sequence.*
- *Show that if $x_1, x_2, \dots$ is a Cauchy sequence that has a convergence subsequence $x_{n_1}, x_{n_2}, \dots$, then $x_1, x_2, \dots$ is convergent and*

$$\lim_{n\to\infty} x_n = \lim_{k\to\infty} x_{n_k}.$$

Definition 3.3.2. *A metric space (X, d) is* complete *if, for any $x_1, x_2, \dots$ that is Cauchy, there is x_∞ such that*

$$x_\infty = \lim_{n\to\infty} x_n.$$

Example 3.3.3.

- *$\mathbb{Q}$ is not complete: Indeed, we showed above that $\mathbb{Q}$ is dense in $\mathbb{R}$. Hence, for each n, we can find $x_n \in \mathbb{Q}$ such that $|x_n - \sqrt{2}| < 1/n$. This is a Cauchy sequence: for $\varepsilon > 0$, let N be such that $N \geq 2\varepsilon$. Then, if $n, m \geq N$,*

$$|x_n - x_m| \leq |x_n - \sqrt{2}| + |\sqrt{2} - x_m| \leq \frac{1}{n} + \frac{1}{m} \leq \frac{2}{N} < \varepsilon.$$

  But x_n cannot converge to a rational number (since it converges to $\sqrt{2}$!).

- *$\mathbb{R}$ is complete: We essentially take this as an axiom. (It is impossible for us to prove since we never 'constructed' the real numbers...)*
- *$\mathbb{N}$ with the metric $d(n, m) = |n - m|$ is complete. Indeed, Cauchy sequences are those that are eventually constant (there is N such that $x_n = x_m$ for all $n, m \geq N$). These clearly converge.*
- *$\mathbb{N}$ with the metric $d(n, m) = |1/n - 1/m|$.*

  Exercise 3.3.2. *Check that this is a metric*

  What do Cauchy sequences $(x_n)_n$ look like?

  If the sequence is bounded (by, say, M), let $\varepsilon = 1/(M+1)^2$ and find N such that $d(x_n, x_m) < \varepsilon$ for $n, m \geq N$. Then

$$\frac{1}{(M+1)^2} = \varepsilon > d(x_n, x_m) = \left|\frac{1}{x_n} - \frac{1}{x_m}\right| \geq \begin{cases} 0 & \text{if } x_n = x_m, \\ \frac{1}{M^2} & \text{if } x_n \neq x_m. \end{cases}$$

  Clearly this implies that $x_n = x_m$ for all $n \geq N$. Hence, $(x_n)_n$ is eventually constant. (These clearly converge to a point $\bar{x} \in \mathbb{N}$).

If the sequence is not bounded, then we claim that $x_n \to \infty$ as $n \to \infty$. Fix any $M > 0$. Let $\varepsilon = 1/2M$ and let N be such that $n, m \geq N$, then $d(x_n, x_m) < \varepsilon$. Since the sequence is not bounded, there is $n_0 > N$ such that $x_{n_0} > 2M$. Then, for any $n \geq N$, we have

$$\frac{1}{x_n} - \frac{1}{2M} < \frac{1}{x_n} - \frac{1}{x_{n_0}} \leq \left| \frac{1}{x_n} - \frac{1}{x_{n_0}} \right| = d(x_n, x_{n_0}) < \varepsilon = \frac{1}{2M}.$$

Hence, $1/x_n < 1/M$; that is $x_n > M$. By the arbitrariness of M, this is precisely the definition of convergence to ∞.

In summary, any Cauchy sequence must either be eventually constant or tend to infinity. Notice that $x_n = n$ is, thus, a Cauchy sequence, but does not converge to anything in $\mathbb{N}$! Hence $\mathbb{N}$, with this metric, is not complete.

Is there a complete metric space containing $\mathbb{N}$ (similar to how $\mathbb{R}$ is a complete metric space containing $\mathbb{Q}$)? Let $\bar{\mathbb{N}} = \mathbb{N} \cup \{\infty\}$, where ∞ is an added distinguished point. We take the metric $\bar{d} : \bar{\mathbb{N}} \times \bar{\mathbb{N}} \to [0, \infty)$ such that $\bar{d}(n, m) = |1/n - 1/m|$ if $n, m \in \mathbb{N}$, $\bar{d}(\infty, n) = \bar{d}(n, \infty) = 1/n$ if $n \in \mathbb{N}$, and $\bar{d}(\infty, \infty) = 0$.

One can check that $\bar{d}$ is a metric, and it agrees with d on $\mathbb{N} \times \mathbb{N}$. On the other hand, $\bar{\mathbb{N}}$ is now a complete metric space.

Exercise 3.3.3. *Show this!*

Definition 3.3.4. *We say that a sequence $x_1, x_2, \ldots$ in a metric space (X, d) is* **bounded** *if there exists $R > 0$ such that $d(x_1, x_n) < R$ for every n.*

Exercise 3.3.4. *Suppose that $x_1, x_2, \ldots$ is a Cauchy sequence in (X, d), a metric space. Show that it bounded.*

3.3.1. **Completing a metric space.** Notice that, for $\mathbb{Q}$ and $(\mathbb{N}, d)$, we could 'complete' the metric space to find a (unique-ish!) metric space containing them. We now show how to do this in general.

Definition 3.3.5 (Equivalence relation)**.** *Given a set A, an equivalence relation $\sim$ satisfies:*

(i) (Reflexivity) $a \sim a$;

(ii) (Transitivity) $a \sim b$ and $b \sim c$ implies that $a \sim c$;

(iii) (Symmetry) $a \sim b$ implies that $b \sim a$.

[Note: one can write this very formally as $\sim \subset A \times A$ such that $(a, a) \in \sim$ for all $a \in A$, etc... I do not personally find this useful as it is overly formal for my liking, but I am happy to develop it with you in office hours if you are curious / think it will be helpful!]

Example 3.3.6. *(i) $=$ is an equivalence relation on $\mathbb{R}$;*

(ii) For $a, b \in \mathbb{N}$, $a \sim b$ if $a \mod 2 = b \mod 2$ (that is, if both a and b are even or if both a and b are odd) is an equivalence relation.

(iii) Consider $C(\mathbb{R})$. An equivalence relation is: $f \sim g$ if $f(0) = g(0)$.

(iv) For a 'nice' set of functions, an equivalence relation is: $f \sim g$ if f and g are equal except at a countable number of points.

Definition 3.3.7 (Equivalence classes)**.** *Given an equivalence relation $\sim$ on a set A, we can form the (disjoint) equivalence classes: for each $a \in A$, its equivalence class is*

$$[a] := \{a' \in A : a \sim a'\}.$$

Note that, if $a \sim a'$, then $[a] = [a']$.

We note that, if set up well, the equivalence classes have the same features as their elements. For example, in Example 3.3.6.((ii)), we can add elements: indeed, for any sets A and B that are a part of a larger linear space V:

$$A + B = \{z \in V : z = a + b \text{ for some } a \in A, b \in B\}.$$

Applying this, we see that $[0]+[0] = [0]$, $[0]+[1] = [1]$, and $[1]+[1] = [0]$... this make sense because they encode "even plus even is odd," "even plus odd is odd," and "odd plus odd is even," respectively, and each of these is true irrespective of which element of the equivalence class you take.

Idea: If we wish to "complete" a metric space, let us use equivalence classes of Cauchy sequences to fill in the "holes" in the metric space. Let (X, d) be a metric space and let

$$\bar{X} = \{[(x_n)_n] : x_1, x_2, \dots \text{ is a Cauchy sequence in } X\},$$

where we use the equivalence relation

$$(x_n)_n \sim (y_n)_n \quad \text{ if } \lim_{n\to\infty} d(x_n, y_n) = 0.$$

Exercise 3.3.5. *Check that this is an equivalence relation!*

This inherits a distance function $\bar{d}$ from X:

$$\bar{d}([(x_n)_n], [(y_n)_n]) = \lim_{n\to\infty} d(x_n, y_n). \tag{3.3.1}$$

Exercise 3.3.6. *Show this is well-defined; that is, show that the limit on the right exists and is independent of the choice of representatives $(x_n)_n$ and $(y_n)_n$ of the equivalence class.*

The last part of this exercise is important – we want our operations to work equally well on *any* element of the equivalence class.

Note that "a copy" of X "lives inside" the larger space $\bar{X}$:

$$X_{\text{copy}} = \{[(x_n)_n] \in \bar{X} : x_1 = x_2 = \cdots = x_n = \cdots\}. \tag{3.3.2}$$

There is a clear bijection $f : X \to X_{\text{copy}}$ such that

$$d(x, y) = \bar{d}(f(x), f(y)) \qquad \text{for all } x, y \in X. \tag{3.3.3}$$

Importantly, one can show that the new space is itself complete

Exercise 3.3.7. *Show that $(\bar{X}, \bar{d})$ is complete. This is a little tricky, but if you want to shore up your technical skills, it is worthwhile!*

Moreover, if $(\bar{\bar{X}}, \bar{\bar{d}})$ is another complete metric space in which X is dense, then one can find $f : \bar{X} \to \bar{\bar{X}}$ that is a bijection and preserves distance (as in (3.3.3)).

Exercise 3.3.8. *Show this!*

Let us sum all of the above up into one result:

Theorem 3.3.8 (Cauchy completion theorem)**.** *Suppose that (X, d) is a metric space. Let $\sim$ be an equivalence relation on the set $\tilde{X} \subset X^{\mathbb{N}}$ of all Cauchy sequences in X defined by*

$$(x_n)_n \sim (y_n)_n \quad \text{if and only if } \lim_{n\to\infty} d(x_n, y_n) = 0.$$

Define $(\bar{X}, \bar{d})$ to be defined by

- $\bar{X} = \{[(x_n)_n] : (x_n)_n \in \tilde{X}\}$;
- $\bar{d}([(x_n)_n)], [(y_n)_n]) = \lim_{n\to\infty} d(x_n, y_n)$.

This is a complete metric space. Additionally, there is a copy of X living in $\bar{X}$; that is, there is a map

$$\iota : X \to \bar{X} \qquad \text{defined by } \iota(x) = (x, x, \dots)$$

that satisfies $d(x, y) = \bar{d}(\iota(x), \iota(y))$ for any $x, y \in X$.

Example 3.3.9.

- *$\mathbb{N}$ with the metric $d(n, m) = |1/n - 1/m|$. We showed in Example 3.3.3 that $\mathbb{N}$ is completed by adding in a point ∞. Let us see this at the level of Cauchy sequences. Let $(x_n)_n$ be a Cauchy sequence. As we showed in Example 3.3.3, either x_n is eventually constant (that is, there is N such that $x_n = \bar{x}$ for some $\bar{x}$ and all $n \geq N$) or $x_n \to \infty$ as $n \to \infty$ (that is, for any $M > 0$, there is N such that $x_n \geq M$ for all $n \geq N$). In the former case, we identify $\bar{x}$ with $[(x_n)_n]$. In the latter case, we identify ∞ with ∞ with $[(x_n)_n]$. Hence, the completion procedure outlined above yields $\mathbb{N} \cup \{\infty\}$ as in Example 3.3.3.*
- *One way to define $\mathbb{R}$ is as the completion of $\mathbb{Q}$.*

Exercise 3.3.9. *Find the completion of $(\mathbb{N}, \tilde{d})$ where $\tilde{d}$ is defined as:*

$$\tilde{d}(n, m) = |((-1)^n + 1/n) - ((-1)^m + 1/m)|.$$

Back to extending functions: So if we have a uniformly[17] continuous function $f : X \to Y$ and the completions $\bar{X}$ and $\bar{Y}$ of f, can we define $\bar{f} : \bar{X} \to \bar{Y}$ that agrees[18] with f on X and Y and is itself continuous?

A useful fact to know is the following:

Lemma 3.3.10. *Suppose that $f : X \to Y$ is uniformly continuous. Then $(f(x_n))_n$ is Cauchy whenever $(x_n)_n$ is Cauchy.*

Proof. Take $x_1, x_2, \dots$ that is Cauchy in X and fix $\varepsilon > 0$. We show that $f(x_1), f(x_2), \dots$ is Cauchy in Y. Since f is uniformly continuous, there is $\delta > 0$ such that $d(x, y) < \delta$ implies that $d(f(x), f(y)) < \varepsilon$. By the Cauchy-ness of x_n, take N such that $n, m \geq N$ implies that $d(x_n, x_m) < \delta$. Hence, if $n, m \geq N$, we have

$$d_Y(f(x_n), f(x_m)) < \varepsilon.$$

This shows that $f(x_n)$ is Cauchy. Hence, the proof is complete. □

[17] Actually, our motivating example was locally uniformly continuous – it was uniformly continuous on any open ball $B_r(x_0)$. Actually, one can show that any Cauchy sequence is bounded, so that the assuming uniformly continuous instead of locally uniformly continuous is really just a technical assumption and can be relaxed at the expense of more work... so let us just do less work!

[18] OK, truly we should use the "copy" of X and Y living in $\bar{X}$ and $\bar{Y}$ (see (3.3.2)), but let us not be so pedantic...

Now we will extend f to get $\bar{f}$:

- Cauchy sequences in X get taken to Cauchy sequences in Y: This follows from Lemma 3.3.10.
- Defining $\bar{f} : \bar{X} \to \bar{Y}$: Fix an element $\bar{x} = [(x_n)_n] \in \bar{X}$ and let

  $$\bar{f}(\bar{x}) = [(f(x_n))_n].$$

  Well-defined? By the previous set $(f(x_n)$ is a Cauchy sequence). But we need to be careful that this value does not depend on the choice of representative. In other words, take $(x_n)_n$ and $(y_n)_n$ such that $(x_n)_n \sim (y_n)_n$ and we need to show that $(f(x_n))_n \sim (f(y_n)_n$. Fix $\varepsilon > 0$ and take δ as in the definition of uniform continuity for f. Next, find N such that, if $n \geq N_x$, we have

  $$d_X(x_n, y_n) < \delta.$$

  Hence, $d_Y(f(x_n), f(y_n)) < \delta$. It follows that $(f(x_n))_n \sim (f(y_n))_n$ and $\bar{f}$ is well-defined.
- Uniform continuity: Fix any $\varepsilon > 0$. Using the uniform continuity of f, take $\delta > 0$ such that

  $$d_Y(f(x), f(z)) < \frac{\varepsilon}{2} \quad \text{whenever } d_X(x, z) < \delta.$$

  Suppose that $\bar{x}$ and $\bar{z} \in \bar{X}$ satisfy

  $$\bar{d}_{\bar{X}}(\bar{x}, \bar{z}) < \frac{\delta}{2}.$$

  By (3.3.1), there is N such that, if $n \geq N$, we have

  $$d_X(x_n, z_n) < \delta.$$

  Hence, whenever $n \geq N$,

  $$d_Y(f(x_n), f(z_n)) < \frac{\varepsilon}{2}.$$

  It follows that

  $$d_{\bar{Y}}(\bar{f}(\bar{x}), \bar{f}(\bar{z})) = \lim_{n\to\infty} d_Y(f(x_n), f(z_n)) \leq \frac{\varepsilon}{2} < \varepsilon.$$

  This is exactly the proof that $\bar{f}$ is uniformly continuous.
- Uniqueness: Suppose that $\bar{\bar{f}} : \bar{X} \to \bar{Y}$ is continous and $\bar{\bar{f}}|_X = f$.

  Exercise 3.3.10. *Show that $\bar{\bar{f}} = \bar{f}$!*

Theorem 3.3.11. *Suppose $f : X \to Y$ is a uniformly continuous function and $\bar{X}$ and $\bar{Y}$ are complete metric spaces such that X is a dense subset of $\bar{X}$ and Y is a dense subset of $\bar{Y}$. Then there is a unique continuous function $\bar{f} : \bar{X} \to \bar{Y}$ such that $\bar{f}|_X = f$.*

3.3.2. How to tell a space is complete? And closed sets.

Definition 3.3.12 (Closed sets)**.** *A set $C \subset X$ of a metric space (X, d) is closed if $X \setminus C = \{x \in X : x \notin C\}$ is open.*

WARNING: open and closed are not opposites!

- (Sets can be open and closed): $\emptyset$ and X are both open and closed.

- (Sets can be neither open nor closed): $[0,1)$ is neither open nor closed in $\mathbb{R}$ (with the usual metric).

Example 3.3.13. *(i) $\mathbb{R}$ – the following sets are closed:*

$$\{a\}, \quad [a,b], \quad [a,\infty), \quad (-\infty,a].$$

(ii) (X, d_{disc}) – any subset $S \subset X$ is closed (and open!).

(iii) The subspaces of ℓ^∞, c_0 and c, are closed (see, e.g., HW1 Problem 8).

Proposition 3.3.14 (Alternate definition of closed)**.** *A subset S of a metric space (X,d) is closed if and only if, for every convergent sequence s_n of points in S, $\lim_{n\to\infty} s_n \in S$.*

Proof of $\Longrightarrow$. Let s_n be an arbitrary convergent sequence of points in S and denote its limit $\bar{x} \in X$. Suppose, by way of contradiction that $\bar{x} \in X \setminus S$. Since S is closed, $X \setminus S$ is open. Hence, there is $r > 0$ such that

$$B_r(\bar{x}) \subset X \setminus S. \tag{3.3.4}$$

On the other hand, since $s_n \to \bar{x}$, there is N such that, if $n \geq N$, we have

$$d(s_n, \bar{x}) < r; \qquad \text{that is, } s_n \in B_r(\bar{x}).$$

This contradicts (3.3.4), which concludes the proof. □

Proof of $\Longleftarrow$. We prove this by showing the contrapositive. Assume that S is not closed and, hence, that $X \setminus S$ is not open. This implies that there exists $\bar{x} \in X \setminus S$ such that, for every $r > 0$,

$$B_r(\bar{x}) \not\subset X \setminus S.$$

Thus, for each $n \in \mathbb{N}$, we may find

$$s_n \in B_{1/n}(\bar{x}) \cap S.$$

In other words, $s_n \in S$ and

$$d(s_n, \bar{x}) < \frac{1}{n}.$$

It follows that $s_n \to \bar{x}$ as $n \to \infty$, $s_n \in S$ for each n, and $\bar{x} \notin S$. This completes the proof. □

Why might closed sets be useful for us? One reason is that it gives an easy way to tell if a metric space is complete:

Proposition 3.3.15. *Suppose that (X,d) and (Y,ρ) are metric spaces with $X \subset Y$ and $\rho|_{X\times X} = d$. If Y is complete and X is closed (as a subset of Y), then X is a complete metric space.*

Proof. Suppose that x_n is a Cauchy sequence in X. Since Y is complete, there is $\bar{y} \in Y$ such that $x_n \to \bar{y}$ as $n \to \infty$. As X is closed, $\bar{y} \in X$ by Proposition 3.3.14. Hence, X is complete. □

Some other important properties of closed sets.

Proposition 3.3.16. *If $f : X \to Y$ is a continuous function with $A \subset Y$ closed, then $f^{-1}(Y)$ is closed.*

Exercise 3.3.11. *Prove this!*

This is particularly useful when thinking about the kernel of an operator: if $\lambda : X \to \mathbb{R}$ is a bounded linear functional, then $\lambda^{-1}(\{0\})$ is closed.

Proposition 3.3.17. *Suppose (X, d) is a metric space. The following are true:*

(i) If $\mathcal{I}$ is an indexing set and A_ι is a collection of closed subsets, then

$$\bigcap_{\iota \in \mathcal{I}} A_\iota \quad \text{is closed.}$$

(ii) The closure of A, denoted $\bar{A}$, and given by:

$$\bar{A} = \bigcap_{B \supset A, B \text{ is closed}} B$$

is a closed set. It is the smallest closed set containing A: if $B \supset A$ and B is closed, then $\bar{A} \subset B$.

(iii) The closure is the set of points in A and its limit points:

$$\bar{A} = A \cup A_{lp}$$

where $A_{lp} = \{x \in X : \text{there is a sequence } a_n \in A \text{ such that } a_n \to x \text{ as } n \to \infty\}$.

Exercise 3.3.12. *Prove this! (Hint: these are simply analogous to the same statements for open sets, unions, and interiors)*

3.4. Compactness.

A thought experiment. Suppose we have some sufficiently nice function $F : X \to Y$, where (X, d_X) and (Y, d_Y) are metric spaces. In order to solve a problem

$$F(\bar{x}) = \bar{y}, \tag{3.4.1}$$

where $\bar{y} \in Y$ is given but $\bar{x}$ is unknown. The "nicest" case is when X is finite. Then we can simply take every element $x_1, x_2, \ldots, x_N$ in X and try them out; that is, we just check $F(x_1)$ then $F(x_2)$ and so on until we find the correct x_i such that $F(x_i) = \bar{y}$.

If X is infinite, we can no long do that. However, the whole world of numerical analysis has developed algorithms for many problems like this. Hence, we should be able to "approximately" solve it. Indeed, perhaps there is an algorithm to get "close" to the solution: we can find $x_1 \in X$ and $y_1 \in Y$ such that

$$F(x_1) = y_1 \quad \text{with } y_1 \approx y,$$

and then iterate to get "closer":

$$F(x_n) = y_n \quad \text{with } d(y_n, y) < d(y_{n-1}, y).$$

If we are lucky, x_n is convergent to some x and F is continuous so that $F(x) = y$.

There is a tension here: we want convergence to be "easy" so that the sequence x_n converges, but if it is too "easy," then this makes it "harder" for F to be continuous (more convergent sequences to test in the definition of sequential continuity). For example:

Exercise 3.4.1. *Suppose that (X, d) is a nonempty metric space in which all sequences are convergent. Then X has exactly one point.*

A more refined idea might be to work in a space where all sequences x_n have subsequences x_{n_k} that converge. Notice that, were this the case, our strategy to solve (3.4.1) would still be successful.

Example 3.4.1. *(i) Assume that X is a finite set: there is N such that*

$$X = \{x_1, x_2, \dots, x_N\}.$$

Suppose that s_n is a sequence of elements of X. We identify s_n with the function $f : \mathbb{N} \to X$ defined by $f(n) = s_n$. Since

$$\mathbb{N} = \bigcup_{k=1}^{N} f^{-1}(\{x_k\}),$$

there must be some $\ell \in \{1, \dots, N\}$ such that $f^{-1}(x_\ell)$ is infinite. This set is, of course, countable, so we can enumerate it $n_1, n_2, n_3, \dots$. It follows that

$$(s_{n_1}, s_{n_2}, s_{n_3}, \dots) = (x_\ell, x_\ell, x_\ell, \dots).$$

Hence, s_n has a convergent subsequence.

Actually, there are a number of other advantages to working in this sort of finite setting, the biggest one being the existence of maxima and minima of any functions. But notice that this will be true of a space in which all sequences have convergent subsequences (we will show this in a bit). This shows us that this sort of "generalized finite" space is very useful. Let us now formally define it.

Definition 3.4.2 (Sequentially compact)**.** *A metric space (X, d) is sequentially compact if, for every sequence x_n of points in X, there is $\bar{x}$ and a subsequence x_{n_k} such that*

$$\lim_{k\to\infty} x_{n_k} = \bar{x}.$$

Note: There is another notion that is equivalent to sequential compactness in the setting of metric spaces. A space X is called "compact" if, for every collection $\mathcal{O} = \{U_\iota : \iota \in \mathcal{I}\}$ of open sets U_ι such that

$$X \subset \bigcup_{\iota\in\mathcal{I}} U_\iota,$$

there is a finite subcollection $U_{\iota_1}, \dots, U_{\iota_n} \in \mathcal{O}$ such that

$$X \subset \bigcup_{k=1}^{n} U_{\iota_k}.$$

We will not use this notion, but you should be aware that it exists and is sometimes more useful.

Example 3.4.3. *(i) $\mathbb{R}$ is* not *sequentially compact – Take the sequence $x_n = n$. No subsequence converges since, for any subsequence x_{n_k}, we have $x_{n_k} = n_k \to \infty$ as $k \to \infty$.*

(ii) $(0, 1)$ is not *sequentially compact – Take $x_n = 1 - 1/n$. Clearly, $x_n \to 1$ as $n \to \infty$, but $1 \notin (0, 1)$! Similarly, if $S \subsetneq \mathbb{R}$ is a nonempty, open set, then S is not compact.*

Exercise 3.4.2. *Prove this!*

(iii) $[0,1]$ *is sequentially compact – We will show this soon...*

(iv) The unit ball $B_1(0)$ *of* ℓ^p *(or even the* closure *of the unit ball!) is* not *sequentially compact – Take the sequence* $\bar{x}_n = (1/2)\bar{e}_n$*, which is all zeros except a* $1/2$ *in the* n*th place. Since*

$$\|s_n - s_m\|_{\ell^p} \geq \frac{1}{2} \quad \text{if } n \neq m,$$

then s_n *is not Cauchy (and, thus, not convergent).*

(v) $(\mathbb{N} \cup \{\infty\}, d)$ *with* $d(n,m) = |1/n - 1/m|$ *as in Example 3.3.3.*

Exercise 3.4.3. *Show this is sequentially compact. Hint: we already characterized all convergent sequences.*

Proposition 3.4.4. *Any closed, finite interval* $[a,b]$ *is sequentially compact.*

Intuitively, we want to use the pigeonhole principle to make this argument, similar to how we approached Example 3.4.1.((i)). Let us "cut" the space in half – one of those halves has to have infinitely many points of the sequence, so we will take our first element of our subsequence from that half. Then, we cut this half in half again, taking our next point from there. And so on and so forth.

Proof. Fix a sequence $x_n \in [a,b]$, and we show that it has a convergent subsequence.

Step one: Let $f : \mathbb{N} \to [0,1]$ be the function such that $f(n) = x_n$. We choose a sequence of intervals $[a_n, b_n]$ inductively so that, for all n,

$$f^{-1}([a_n, b_n]) \text{ is infinite}, \quad [a_{n+1}, b_{n+1}] \subset [a_n, b_n], \quad \text{and} \quad b_n - a_n = (b-a)2^{-n}. \tag{3.4.2}$$

First, let $[a_0, b_0] = [a,b]$. Now we show how to find $[a_{n+1}, b_{n+1}]$ whenever we have been given $[a_n, b_n]$. Let $\delta = b_n - a_n > 0$. As in Example 3.4.1.((i)), either

$$f^{-1}([a_n, a_n + \delta/2]) \quad \text{or} \quad f^{-1}([a_n + \delta/2, b_n])$$

must be infinite. If the former (or if both are infinite), let $[a_{n+1}, b_{n+1}] = [a_n, a_n + \delta/2]$, and if the latter, let $[a_{n+1}, b_{n+1}] = [a_n + \delta/2, b_n]$.

Step two: We now choose a subsequence inductively as follows. Let $X_n = f^{-1}([a_n, b_n])$. Let $n_1 = \min X_1$. Now, suppose that we have chosen the indices $n_1, n_2, \ldots, n_k$. Choose n_{k+1} to be

$$n_{k+1} = \min X_{k+1} \setminus \{1, 2, 3, \ldots, n_k\}.$$

We point out two things: (1) X_{k+1} is infinite, by construction, so that the set on the right is nonempty, and (2) the set on the right is a set of natural numbers so that a minimum does exist (as opposed to an infimum). Hence, n_{k+1} is well-defined and, by construction

$$n_1 < n_2 < n_3 < \cdots$$

so that x_{n_k} is an admissible subsequence of x_n. We note that

$$n_k = (n_k - n_{k-1}) + n_{k+1} \geq 1 + n_{k+1}$$

Iterating this, we have

$$n_k \geq k. \tag{3.4.3}$$

Step three: We show that x_{n_k} is a Cauchy sequence, which will finish the proof (by the completeness of $[a, b]$, x_{n_k} is convergent). Fix $\varepsilon > 0$ and choose N such that

$$(b-a)2^{-N} < \varepsilon.$$

Notice that, by (3.4.2), we have, if $k, \ell \geq N$,

$$x_{n_k}, x_{n_\ell} \in [a_N, b_N],$$

where we recall that, by (3.4.3), $n_k, n_\ell \geq N$. Hence,

$$|x_{n_k} - x_{n_\ell}| \leq |b_N - a_N| = (b-a)2^{-N} < \varepsilon.$$

By the arbitrariness of ε, we deduce that x_{n_k} is a Cauchy sequence. Since $[a, b]$ is complete, it follows that x_{n_k} is convergent. This concludes the proof. □

We can easily upgrade this to $\mathbb{R}^d$.

Proposition 3.4.5. *Let $a_i < b_i$ for $i = 1, 2, \ldots, d$. Then*

$$[a_1, b_1] \times [a_2, b_2] \times \cdots \times [a_d, b_d]$$

is a sequentially compact subset of $\mathbb{R}^d$.

Proof. Fix any sequence $\bar{x}_1, \bar{x}_2, \ldots$ in $[a_1, b_1] \times \cdots \times [a_d, b_d]$. Notationally let,

$$\bar{x}_k = (x_k^{(1)}, x_k^{(2)}, \ldots, x_k^{(d)});$$

that is, the superscript indicates the coordinate and the subscript denotes where in the sequence $\bar{x}_k$ falls.

Notationally, this is an intricate proof. But the idea is that, fixing. a coordinate i, the sequence $x^{(i)}$ has to converge (up to taking a subsequence). So we can take a subsequence for the first coordinate that converges, then we take a subsequence of that to get convergence of these second coordinate, and so on until the dth coordinate. Since there are only finitely many coordinates, the last subsequence (which is, in a sense, a sub-sub-$\cdots$-subsequence) is still a subsequence of the original one.

Using Proposition 3.4.4, we can find a subsequence $n_1^{(1)}, n_2^{(1)}, \ldots$ and a point $x^{(1)}$ such that

$$x^{(1)}_{n_k^{(1)}} \to x^{(1)} \quad \text{as } k \to \infty.$$

We can repeat this process: for each $j \in \{2, \ldots, d\}$, we can apply Proposition 3.4.4 to find a subsequence $n_1^{(j)}, n_2^{(j)}, \ldots$ of $n_1^{(j-1)}, n_2^{(j-1)}, \ldots$ and a point $x^{(j)}$ such that

$$x^{(j)}_{n_k^{(j)}} \to x^{(j)} \quad \text{as } k \to \infty.$$

Since $n_k^{(d)}$ is a subsequence of $n_k^{(j-1)}$, we have $x^{(d-1)}_{n_k^{(d)}} \to x^{(d-1)}$ as $k \to \infty$. Similarly, $x^{(i)}_{n_k^{(d)}} \to x^{(i)}$ as $k \to \infty$ for every $i \in \{1, \ldots, d\}$. In other words

$$\bar{x}_{n_k^{(d)}} \to (x^{(1)}, x^{(2)}, \ldots, x^{(d)}) \quad \text{as } k \to \infty.$$

□

Exercise 3.4.4. *Suppose that $(X_1, d_1), \ldots, (X_N, d_N)$ are sequentially compact metric spaces. Then $X_1 \times \cdot \times X_N$ is a sequentially compact metric space.*

While this gives us the quintessential example of a sequentially compact set, it also is a key piece in characterizing all sequentially compact subsets of $\mathbb{R}^n$. We do this now.

3.4.1. **The Bolzano-Weierstrass Theorem.** Our goal is to prove the characterization of sequentially compact sets in $\mathbb{R}^n$. Along the way, we also prove several general results that will be useful to us.

Theorem 3.4.6 (Bolzano-Weierstrass Theorem)**.** *A set $K \subset \mathbb{R}^n$ is sequentially compact if and only if K is closed and bounded.*

Let me make a large disclaimer here: this theorem is *only* true in $\mathbb{R}^n$ and other special cases. See, for example, Example 3.4.3.((iv)). We should also clarify the definition of bounded:

Definition 3.4.7. *A set C is bounded in a metric space (X, d) if there is some $x_0 \in X$ and $r > 0$ such that $C \subset B_r(x_0)$.*

The first step is the following.

Proposition 3.4.8. *If (X, d) is a sequentially compact metric space and $K \subset X$ is closed set, then (K, d) is a sequentially compact metric space.*

Proof. If k_n is a sequence of points in K, then it is a sequence of points in X. By the sequential compactness of X, there is a subsequence k_{n_i} and a point $x \in X$ so that $k_{n_i} \to x$ as $i \to \infty$. Since K is closed, $x \in K$ by Proposition 3.3.14. By the arbitrariness of the original sequence, K is sequentially compact. □

The next ingredient is that sequentially compact sets are closed and bounded.

Proposition 3.4.9. *Suppose that (X, d) is a metric space and $K \subset X$ such that (K, d) is sequentially compact. Then K is closed as a subset of X and bounded.*

Proof. This is easiest to prove by contradiction. Let us do bounded first.

Boundedness: Suppose that C is not bounded. Fix any $x_0 \in X$. Then, for each n, there is $x_n \in C \cap (B_r(x_0)^c)$. Suppose that $x_n \to \bar{x}$ for some $\bar{x} \in C$. Notice that

$$d(x_n, \bar{x}) \geq d(x_n, x_0) - d(x_0, \bar{x}) \to \infty \quad \text{as } n \to \infty.$$

This contradicts. that $x_n \to \bar{x}$.

Closedness: Suppose that K is not closed. Then, by Proposition 3.3.14, there is $\bar{x} \in X \setminus K$ and $k_n \in K$ such that $k_n \to \bar{x}$ as $n \to \infty$. Since K is sequentially compact, there is a $\bar{k} \in K$ and a subsequence $k_{n_1}, k_{n_2}, \ldots$ such that

$$k_{n_i} \to \bar{k} \qquad \text{as } i \to \infty.$$

Of course, k_{n_i} must converge to $\bar{x}$ as well (it is a subsequence of a sequence converging to $\bar{x}$). It follows that $\bar{x} = \bar{k}$. This is a contradiction because $\bar{x} \in X \setminus K$ and $\bar{k} \in K$. □

We can now prove the Bolzano-Weierstrass theorem.

Proof of Theorem 3.4.6. $\Longrightarrow$: This follows directly from Proposition 3.4.9.

$\Longleftarrow$: Since K is bounded, there is $r > 0$ and $\bar{x} \in \mathbb{R}^n$ such that $K \subset B_r(\bar{x})$. It is easy to check that

$$K \subset B_r(\bar{x}) \subset [-R, R]^n,$$

where $R = |\bar{x}| + r$. By Proposition 3.4.5. □

3.4.2. **Sequentially compact sets and continuous functions.**

Proposition 3.4.10 (Extreme Value theorem)**.** *Suppose that $f : X \to \mathbb{R}$ is a continuous function and (X, d) is a sequentially compact metric space. Then there is $x_{\max} \in X$ such that*

$$f(x_{\max}) = \sup_{x \in X} f(x).$$

Proof. There exists a sequence $x_n \in X$ such that

$$f(x_n) \to \sup_{x \in X} f(x).$$

By sequential compactness, there is a subsequence x_{n_k} such that $x_{n_k} \to \bar{x}$ as $k \to \infty$ for some $\bar{x} \in X$. By the continuity of f, we have

$$f(\bar{x}) = \lim_{k \to \infty} f(x_{n_k}) = \lim_{n \to \infty} f(x_n) = \sup_{x \in X} f(x).$$

□

You are probably already aware that this is very useful. You are probably also aware that the hypotheses may not be relaxed:

Example 3.4.11. *(i) Define $f : (0, 1) \to \mathbb{R}$ by $f(x) = x$. Then $\sup f = 1$, but $f(x) < 1$ for all $x \in (0, 1)$. Hence, "closedness" cannot be relaxed.*

(ii) Define $g : (0, 1) \to \mathbb{R}$ by $g(x) = 1/x$. Then $\sup g = +\infty$, and, thus, there is no $x_{\max}$ such that $g(x_{\max}) = \sup g$. Again we see that "closedness" cannot be relaxed.

(iii) Define $h : [0, \infty) \to \mathbb{R}$ by $h(x) = 1 - e^{-x}$. Then $\sup h = 1$, but $h(x) < 1$ for all $x \in [0, \infty)$. Hence, "boundedness" cannot be relaxed.

It turns out that Proposition 3.4.10 is extremely useful! Let us prove a few results using it in order to illustrate its power.

First, you now have it in your power to actually prove the following (this was a homework problem where we used the Extreme Value theorem out of the box...)

Proposition 3.4.12. *Let $\|\cdot\|$ be any norm on $\mathbb{R}^d$. Then $\|\cdot\|$ is equivalent to $\|\cdot\|_2$.*

Second, although not an application of the Extreme Value theorem, is similar in spirit to it.

Proposition 3.4.13. *Suppose that K and C are, respectively, sequentially compact and closed subsets of a metric space (X, d) with $K \cap C = \emptyset$. Then there exists $\delta > 0$ such*

$$\delta = \operatorname{dist}(K, C) := \inf\{d(k, c) : k \in K, c \in C\}.$$

Exercise 3.4.5. *Suppose, in Proposition 3.4.13, that C is sequentially compact as well. Then there exists $k \in K$ and $c \in C$ such that*

$$0 < d(k_m, c_m) = \operatorname{dist}(K, C).$$

The easiest proof of this uses Proposition 3.4.10 (the Extreme Value theorem).

Proof. We prove this by contradiction: assume that $\delta = 0$. Take a sequence of points $k_n \in K$ and $c_n \in C$ such that

$$d(k_n, c_n) \to \delta \qquad \text{as } n \to \infty.$$

Since K is sequentially compact, there is a subsequence k_{n_ℓ} and a point $\bar{k}$ such that $k_{n_\ell} \to \bar{k}$ as $\ell \to \infty$. We note that $\bar{k} \in K$ so, by the disjointness of K and C, $\bar{k} \notin C$.

We have

$$0 = \delta = \lim_{\ell \to \infty} d(k_{n_\ell}, c_{n_\ell}) \geq \limsup_{\ell \to \infty} d(c_{n_\ell}, \bar{k}) - \lim_{\ell \to \infty} d(k_{n_\ell}, \bar{k}) = \limsup_{\ell \to \infty} d(c_{n_\ell}, \bar{k}).$$

Hence, $c_{n_\ell} \to \bar{k}$ as $\ell \to \infty$. Since C is closed, we deduce that $\bar{k} \in C$, which is a contradiction. We conclude that $\delta > 0$,m as claimed. □

Example 3.4.14. *The conditions that K is sequentially compact and C is closed are very important. For example, let us consider the following simple examples of disjoint sets:*

(i) Let

$$C_1 = \mathbb{N} \qquad \text{and} \qquad C_2 = \{n + 1/n : n \in \mathbb{N}, n \geq 2\}.$$

Clearly $C_1 \cap C_2 = \emptyset$ and each set C_i is closed. On the other hand, the result of Proposition 3.4.13 is false:

$$|n - (n + 1/n)| = \frac{1}{n} \to 0 \qquad \text{as } n \to \infty,$$

so $\operatorname{dist}(C_1, C_2) = 0$.

(ii) Let $K = \{0\}$, which is sequentially compact, and $V = (0, 1)$. While K is sequentially compact, V is not closed. The result of Proposition 3.4.13 is not true since

$$|0 - 1/n| \to 0 \qquad \text{as } n \to \infty,$$

so $\operatorname{dist}(K, V) = 0$.

Proposition 3.4.15. *Let (C, d) be a compact metric space, and let (Y, ρ) be any metric space. If $f : C \to Y$ is continuous, then f is uniformly continuous.*

Proof. We argue by contradiction. If f is not uniformly continuous, then there is $\varepsilon > 0$ such that, for every n, there is $x_n, y_n \in C$ with

$$\rho(f(x_n), f(y_n)) \geq \varepsilon_0 \quad \text{ and } d(x_n, y_n) < \frac{1}{n}.$$

By sequential compactness, we have convergent sequences x_{n_k} and y_{n_k} so that, letting

$$x_\infty = \lim_{k\to\infty} x_{n_k} \quad \text{and} \quad y_\infty = \lim_{k\to\infty} y_{n_k}.$$

Notice that

$$d(x_\infty, y_\infty) \le d(x_\infty, x_{n_k}) + d(x_{n_k}, y_{n_k}) + d(y_{n_k}, y_\infty) < d(x_\infty, x_{n_k}) + \frac{1}{n_k} + d(y_{n_k}, y_\infty).$$

Taking the limit $k \to \infty$, we find

$$d(x_\infty, y_\infty) = 0, \quad \text{that is, } x_\infty = y_\infty.$$

On the other hand,

$$\varepsilon_0 \le \rho(f(x_{n_k}), f(y_{n_k})) \le \rho(f(x_{n_k}), f(x_\infty)) + \rho(f(x_\infty), f(y_\infty)) + \rho(f(y_\infty), f(y_{n_k})),$$

which implies, after taking the limit $k \to \infty$ and using the continuity of f,

$$0 < \varepsilon_0 \le 0 + \rho(f(x_\infty), f(y_\infty)) + 0 = \rho(f(x_\infty), f(y_\infty)) = 0.$$

The last equality follows from the fact that $x_\infty = y_\infty$. This is clearly a contradiction, which completes the proof. □

One final point to make about sequentially compact spaces is that they must be complete:

Exercise 3.4.6. *Show that if (X, d) is sequentially compact, then it is also complete.*

3.5. **Contraction mappings.** Returning to the idea of solving an equation, one nice way of finding solutions is to frame a problem

$$F(x) = y$$

as a "fixed point" problem:

$$G(x) = x.$$

In this case, we can do so by writing $G(x) = F(x) - y + x$. There is a whole world full of "fixed point theorems." We will give the basic one here.

Theorem 3.5.1. *Suppose that (X, d) is a complete metric space and $T : X \to X$ is a map such that there is $\theta \in [0, 1)$ such that*

$$d(T(x), T(y)) \le \theta d(x, y) \qquad \text{for every } x, y \in X. \tag{3.5.1}$$

(We say that T is a contraction mapping*). Then there is a unique point x_f such that $T(x_f) = x_f$.*

Remark 3.5.2.

- *The restriction $\theta < 1$ is extremely important (as we shall see in the proof). Indeed, consider the case $T : \mathbb{R}^n \to \mathbb{R}^n$ such that $T(x) = e_1 + x$. This satisfies $d(T(x), T(y)) = d(x, y)$ (that is, (3.5.1) with $\theta = 1$), but it clearly does not have a fixed point.*

- *Let us give another example about the restriction $\theta < 1$. It is not enough to have $d(T(x), T(y)) < \theta d(x, y)$ for all $x \neq y$. Indeed, consider $f : [1, \infty) \to [1, \infty)$ defined by $f(x) = x + e^{-x}$.*

  Consider any $x < y$. Notice that, by the Mean Value Theorem, there exists $\xi \in [x, y]$ such that

  $$|f(x) - f(y)| = |f'(\xi)||x - y| = (1 - e^{-\xi})|x - y| < |x - y|.$$

  On the other hand, we cannot have a fixed point x_0 since this would imply that

  $$x_0 = f(x_0) = x_0 + e^{-x_0} > x_0.$$

  This is clearly a contradiction.

- *One way that this is often applied (especially in numerics) is, if we are trying to solve an equation $S(x) = 0$ on a normed vector space M, we look for a number μ such that $T(x) = x - \mu S(x)$ is a contraction mapping. In this case, the fixed point x_f must satisfy $S(x_f) = 0$!*

  This method of solving $S(x) = 0$, that follows the proof below, is often called "relaxation".

- *Any contraction mapping is uniformly continuous. In fact, the condition* (3.5.1) *(with or without the assumption that $\theta < 1$) is a particular type of continuous called "Lipschitz continuous." Why is it continuous? Given $\varepsilon > 0$, let $\delta = \varepsilon/(\theta + 1)$, and, if $d(x, y) < \delta$,*

  $$d(T(x), T(y)) \leq \theta d(x, y) \leq \theta\delta = \theta \frac{\delta}{\theta + 1} < \varepsilon.$$

Proof. There are two parts to this proof: existence of a fixed point and uniqueness of the fixed point. We show these separately.

Existence: The intuition here is that if we keep applying T to some point, it always brings the successive points closer together. In the limit, the points will have to be infinitely close together; that is, the limit and T applied to it, will have to be the same!

We inductively define a sequence x_n as follows. Let $x_1 \in X$ be any point. Then, for any n, let $x_{n+1} = T(x_n)$. We note that

$$\begin{aligned} d(x_{n+1}, x_n) &= d(T(x_n), T(x_{n-1})) \leq \theta d(x_n, x_{n-1}) = \theta d(T(x_{n-1}), T(x_{n-2})) \\ &\leq \cdots \leq \theta^{n-1} d(x_2, x_1). \end{aligned} \tag{3.5.2}$$

We claim that x_n is a Cauchy sequence. Fix $\varepsilon > 0$ and N_ε to be chosen[19] (depending only on x_1, x_2, and ε). Fix any $n, m \geq N_\varepsilon$ and assume that $n \geq m$ without loss of generality. Applying the triangle inequality and then (3.5.2), we find

$$\begin{aligned} d(x_m, x_n) &\leq \sum_{k=0}^{n-m} d(x_{m+k}, x_{m+k+1}) \leq \sum_{k=0}^{n-m} \theta^{m+k-1} d(x_1, x_2) \leq d(x_1, x_2) \sum_{k=0}^{\infty} \theta^{m+k-1} \\ &= d(x_1, x_2) \frac{\theta^{m-1}}{1 - \theta} \leq d(x_1, x_2) \frac{\theta^{N_\varepsilon - 1}}{1 - \theta}. \end{aligned}$$

Choosing

$$N_\varepsilon > 1 + \frac{\log\left(\frac{(1-\theta)\varepsilon}{d(x_1, x_2)}\right)}{\log \theta},$$

[19] The non-lazy way to choose N_ε is to explicit write it here. But, in practice, mathematicians often write that it is "to be chosen" and then determine it after the computations. One has to be very careful here to make sure to not end up with a circular proof – this is why I have specified at the outset that it will only depend on x_1, x_2, and ε.

we conclude that

$$d(x_n, x_m) < \varepsilon,$$

and, hence, that x_n is Cauchy.

It follows that x_n is convergent to some point x_f. By the continuity of T, we have

$$\lim_{n\to\infty} T(x_n) = T(\lim_{n\to\infty} x_n) = T(x_f).$$

On the other hand $T(x_n) = x_{n+1} \to x_f$. It follows that

$$x_f = T(x_f).$$

Hence, a fixed point exists.

Uniqueness: We now check that this is the only fixed point. The intuition is the following: two fixed points do not get moved by T, but the condition (3.5.1) implies that T moves points closer together. These two behaviors are not compatible!

Let us prove it now. Fix any point $\tilde{x}_f$ such that $T(\tilde{x}_f) = \tilde{x}_f$. Then, by (3.5.1), we find

$$d(x_f, \tilde{x}_f) = d(T(x_f), T(\tilde{x}_f)) \leq \theta d(x_f, \tilde{x}_f).$$

Since $\theta < 1$, this is not possible unless $d(x_f, \tilde{x}_f) = 0$. Thus, we have proven that $x_f = \tilde{x}_f$; that is, the fixed point is unique. □

Example 3.5.3. • **(Newton's method)** *Fix a C^1 function $f : \mathbb{R} \to \mathbb{R}$ and suppose that we know that*

- *f has a zero somewhere in $[x_0 - \varepsilon, x_0 + \varepsilon]$ (perhaps we know that $f(x_0 - \varepsilon) < 0$ and $f(x_0 + \varepsilon) > 0$ by some other means);*
- *$|f| < \delta$ on $[x_0 - \varepsilon, x_0 + \varepsilon]$ for some small δ (this is not crazy since we are "near" a zero so, by decreasing ε, we should be able to make δ arbitrarily small);*
- *satisfies $|f'| > \sqrt{\delta}$ on $[x_0 - \varepsilon, x_0 + \varepsilon]$ (this is the strongest assumption since it is completely reasonable that f and f' can vanish as the same place; e.g., $f(x) = x^2$).*

Let us find this zero. Consider the map:

$$T(x) = x - \frac{f(x)}{f'(x)}.$$

Any fixed point $x_* = T(x_*)$ is also a zero of f: $f(x_*) = 0$. Using Taylor's theorem on f/f', we find

$$\begin{aligned} |T(x) - T(y)| &= \left| x - \frac{f(x)}{f'(x)} - y + \frac{f(y)}{f'(y)} \right| \\ &= \left| (x-y) - (x-y)\frac{f'(\xi_{x,y})f'(\xi_{x,y}) - f(\xi_{x,y})f''(\xi_{x,y})}{f'(\xi_{x,y})^2} \right| \\ &= \left| (x-y)\frac{f(\xi_{x,y})f''(\xi_{x,y})}{f'(\xi_{x,y})^2} \right| \leq |x-y|\sqrt{\delta}\|f''\|_{L^\infty([x_0-\varepsilon, x_0+\varepsilon])}, \end{aligned}$$

where $\xi_{x,y}$ is some point between x and y. If $\delta < \|f''\|^2_{L^\infty([x_0-\varepsilon,x_0+\varepsilon])}$, this is a contraction map. Hence, there is a fixed point of T, which means that f has zero!

Importantly, the proof of Theorem 3.5.1 gives a procedure to find the zero that can easily be implemented numerically.

- **(Iterative solutions of matrix equations)** *Let $T : V \to V$ be a mapping from a complete normed vector space to itself satisfying $\|T\| < 1$. Suppose that we seek a fixed point of M defined by*

$$M(v) = Tv + b.$$

for some vector $b \in V$. Notice that

$$\|Mv - Mw\| = \|T(v - w)\| \leq \|T\|\|v - w\|.$$

Since $\|T\| < 1$, this is a contraction mapping ($\|T\|$ plays the role of θ here). Hence, M has a fixed point v_*. Examining the proof of Theorem 3.5.1, we see that

$$v_* = b + Tb + T^2b + T^3b + \cdots$$

Why? Let $v_0 = b$. Then $v_1 = Mb = Tb + b$, $v_2 = Mv_1 = M(Tb + b) = T^2b + Tb + b$, etc.

Does this make any sense!? Let

$$T_n = I + T + T^2 + T^3 + \cdots + T^n.$$

Notice that, for any $n \geq m$, we have

$$\|T_n - T_m\| \leq \|T^n + T^{n-1} + \cdots + T^{m+1}\| \leq \sum_{k=m+1}^{n} \|T^k\|$$

$$\leq \sum_{k=m+1}^{n} \|T\|^k \leq \frac{\|T\|^m}{1 - \|T\|}.$$

Hence, $\|T_nb - T_mb\| \leq \|T\|^m(1 - \|T\|)^{-1}\|b\|$, which implies that T_nb is a Cauchy sequence. The assumed completeness of V then implies that v_* is well-defined and satisfies

$$Tv_* + b = Mv_* = v_* \qquad \text{or, equivalently, } (I - T)v_* = b.$$

That is $(I - T)^{-1} = I + T + T^2 + \cdots$. That is, the standard geometric series identity holds for linear operators as well!

The operator $I + T + T^2 + \cdots$ is called a **Neumann** *series.*

How can we use this insight in more generality? Let us say we want to solve $Ax = b$. Then we try to split $A = D + E$ where D is an "easily invertible" matrix (say a diagonal or lower triangular matrix). Then we rewrite the problem $Ax = b$ as

$$x + D^{-1}Ex = D^{-1}b.$$

If $\|D^{-1}E\| < 1$, then we can apply the above method to solve this. (Note: we get to choose the norms here, so it might be that the condition $\|D^{-1}E\| < 1$ can be obtained by choosing "good" norms). In this case, the matrix D is called a **preconditioner**.

To be able to apply some of these ideas in more sophisticated ways, we need to show that a few spaces that we know and love are complete. Let us begin with $C([0, 1])$.

Proposition 3.5.4. *For any $a < b$, the space $C([a, b])$ is complete (with the L^∞-norm).*

Before we prove this, let us show an application of it.

Example 3.5.5 (Solving ODE by contraction mappings)**.** *Suppose we look for a solution of the ordinary differential equation*

$$x' = f(t,x) \qquad x(0) = x_0 \tag{3.5.3}$$

where f is Lipschitz continuous in x; that is, there is C_f such that $|f(t,x) - f(t,y)| \le C_f|x-y|$ for all $x, y \in \mathbb{R}$ and all $t \in \mathbb{R}$. By integrating this, it can be recast as solving the integral equation

$$x(t) = x_0 + \int_0^t f(s, x(s))ds. \tag{3.5.4}$$

Let $\varepsilon = 1/(2C_f)$. Define $T : C([0,\varepsilon]) \to C([0,\varepsilon])$ by

$$(Tx)(t) = x_0 + \int_0^t f(s, x(s))ds.$$

Notice that

$$\begin{aligned}\|Tx - Ty\|_\infty &= \Big\| \int_0^t f(s,x(s))ds - \int_0^t f(s,y(s))ds \Big\|_\infty \le \int_0^t \|f(s,x(s)) - f(s,y(s))\|_\infty ds \\ &\le \int_0^t C_f \|x-y\|_\infty ds \le \varepsilon C_f \|x-y\|_\infty < \frac{1}{2}\|x-y\|_\infty.\end{aligned}$$

Hence, T is a contraction mapping. It has a fixed point $x_* \in C([0,\varepsilon])$, which, by (3.5.4), is a solution of the ODE (3.5.3) on the time interval[20] *$[0,\varepsilon]$.*

Proof. Fix any Cauchy sequence $f_n \in C([a,b])$.

Existence of a limiting function f: Since f_n is Cauchy, we can choose a subsequence n_1 such that

$$\|f_{n_k} - f_{n_m}\|_\infty \le 2^{-\min\{k,m\}} \tag{3.5.5}$$

for every k, m.

Exercise 3.5.1. *Prove this!*

Then we define

$$f(x) = f_{n_1}(x) + \sum_{k=1}^{\infty} (f_{n_{k+1}}(x) - f_{n_k}(x)) = f_{n_1}(x) + \lim_{N\to\infty} \sum_{k=1}^{N} (f_{n_{k+1}}(x) - f_{n_k}(x)). \tag{3.5.6}$$

It is easy to see that the limit on the right hand side exists because the series on the right hand side is absolutely summable[21]:

$$\sum_{k=1}^{N} |f_{n_{k+1}}(x) - f_{n_k}(x)| \le \sum_{k=1}^{N} \|f_{n_{k+1}} - f_{n_k}\|_\infty \le \sum_{k=1}^{N} 2^{-k} < 1.$$

We have now identified f, but it remains to show that $f_n \to f$ in the L^∞-norm and that $f \in C([a,b])$.

[20] This can actually be upgrade to a solution on $\mathbb{R}$ by iterating the result on $[\varepsilon, 2\varepsilon]$ then $[2\varepsilon, 3\varepsilon]$, etc.

[21] You may worry that this "calculus fact" is not rigorously justified. If you are one of these people, like me, then it is a good exercise to show that the sequence of partial sums is Cauchy and, hence, convergent.

Convergence of f_n to f: We first show that $f_{n_k} \to f$ in the L^∞-norm. This follows easily from (3.5.6). Fix $\varepsilon > 0$ and let N be such that $2^{-N+1} < \varepsilon$. Then notice that, for any k,

$$f = f_{n_k} + \sum_{\ell=k}^{\infty} (f_{n_{\ell+1}} - f_{n_\ell})$$

so that, for $k > N$,

$$\begin{aligned} \|f - f_{n_k}\|_\infty &= \Big\| \sum_{\ell=k}^{\infty} (f_{n_{\ell+1}} - f_{n_\ell}) \Big\|_\infty \leq \sum_{\ell=k}^{\infty} \|(f_{n_{\ell+1}} - f_{n_\ell})\|_\infty \leq \sum_{\ell=k}^{\infty} 2^{-\ell} \\ &= \frac{2^{-k}}{1-(1/2)} = 2^{-k+1} \leq 2^{-N+1} < \varepsilon. \end{aligned} \tag{3.5.7}$$

Thus, we have shown that $f_{n_k} \to f$.

Now we show that $f_n \to f$. Fix $\varepsilon > 0$ and let N_1 be such that, if $n, m \geq N_1$,

$$\|f_n - f_m\|_\infty < \frac{\varepsilon}{2},$$

and N_2 be such that $2^{-N_2+1} < \frac{\varepsilon}{2}$. Let $N = \max\{N_1, N_2\}$. If $k \geq N$, we have

$$\|f - f_k\|_\infty \leq \|f - f_{n_k}\|_\infty + \|f_{n_k} - f_k\|_\infty \leq 2^{-k+1} + \frac{\varepsilon}{2} < \varepsilon.$$

The first inequality is due to the triangle inequality, the second is due to the work in (3.5.7) and the fact that $n_k \geq k \geq N$, and the third is due to the fact that $2^{-k+1} \leq 2^{-N+1} < \varepsilon/2$.

Finally, we prove that f is continuous. Fix $\varepsilon > 0$ and $x \in [a, b]$. Let N be such that $\|f - f_N\|_\infty < \varepsilon/100$. Since f_N is continuous, there is $\delta > 0$ such that if $|x - y| < \delta$ then $|f_N(x) - f_N(y)| < \varepsilon/100$. Hence, if $|x - y| < \delta$, we have

$$\begin{aligned} |f(x) - f(y)| &\leq |f(x) - f_N(x)| + |f_N(x) - f_N(y)| + |f_N(y) - f(y)| \\ &\leq \|f - f_N\|_\infty + |f_N(x) - f_N(y)| + \|f - f_N\|_\infty < \frac{3\varepsilon}{100} < \varepsilon. \end{aligned}$$

In other words, f is continuous. This completes the proof. □

Next, we show that the ℓ^p spaces are complete as well. We state first a useful lemma.

Lemma 3.5.6. *If x_n is a Cauchy sequence in a metric space (X, d) and there is a convergent subsequence x_{n_k} to a point $\bar{x}$, then $x_n \to \bar{x}$.*

Exercise 3.5.2. *Prove this!*

Proposition 3.5.7. *For $p \in [1, \infty]$, ℓ^p is complete.*

Proof. Fix any Cauchy sequence $\bar{x}_n \in \ell^p$. Notationally, we write $\bar{x}_n$ as the sequence

$$\bar{x}_n = (x_{n,1}, x_{n,2}, x_{n,3}, \dots).$$

We use the same trick as we did in showing the completeness of $C([a, b])$. Arguing exactly as in (3.5.5), we take n_k such that

$$\|\bar{x}_{n_k} - \bar{x}_{n_m}\|_{\ell^p} < 2^{-\min\{k,m\}} \tag{3.5.8}$$

and then, for each k, we define

$$x_{\infty,k} = x_{n_1} + \sum_{k=1}^{\infty} (x_{n_{k+1}} - x_{n_k}).$$

This is a finite real number for the same reasons as in the proof of Proposition 3.5.4.

Fix any k, m and let us consider the case $p < \infty$. Then, using the identity

$$x_{\infty,k} = x_{n_m,k} + \sum_{i=m}^{\infty} (x_{n_{i+1},k} - x_{n_i,k})$$

we find

$$\|\bar{x}_\infty - \bar{x}_{n_m}\|_{\ell^p} = \left\| \sum_{k=m}^{\infty} (\bar{x}_{n_{k+1}} - \bar{x}_{n_k}) \right\|_{\ell^p} \leq \sum_{k=m}^{\infty} 2^{-k} = 2^{-m+1}. \tag{3.5.9}$$

Hence $\bar{x}_{n_m} \to \bar{x}_\infty$. The proof is then concluded by Lemma 3.5.6.

Actually, there is a subtle lie above! It may not be kosher to take the triangle inequality an infinite number of times. So how do we repair this? Fix any N. Define the vectors $v_1, \ldots, v_L$, by

$$v_i = (x_{n_{i+1},1} - x_{n_i,1}, x_{n_{i+1},2} - x_{n_i,2}, \ldots, x_{n_{i+1},N} - x_{n_i,N}) \in \mathbb{R}^N.$$

We also momentarily define the notation $\|\cdot\|_{\ell^p,\mathbb{R}^N}$ denote the ℓ^p norm on $\mathbb{R}^N$. Then

$$\begin{aligned}
\left(\sum_{k=1}^{N} |x_{\infty,k} - x_{n_m,k}|^p \right)^{1/p} &= \lim_{L\to\infty} \left(\sum_{k=1}^{N} \Big| \sum_{i=m}^{L} (x_{n_{i+1},k} - x_{n_i,k}) \Big|^p \right)^{1/p} = \lim_{L\to\infty} \left\| \sum_{i=m}^{L} v_i \right\|_{\ell^p,\mathbb{R}^N} \\
&\leq \limsup_{L\to\infty} \sum_{i=M}^{L} \|v_i\|_{\ell^p,\mathbb{R}^N} = \limsup_{L\to\infty} \sum_{i=m}^{L} \left(\sum_{k=1}^{N} |x_{n_{i+1},k} - x_{n_i,k}|^p \right)^{1/p} \\
&\leq \limsup_{L\to\infty} \sum_{i=m}^{L} 2^{-i} = 2^{-m+1},
\end{aligned}$$

where the first inequality is due to Minkowski's inequality and the last inequality is due to (3.5.8). Then (3.5.9) follows by taking $N \to \infty$ in the above.

Exercise 3.5.3. *Do the case $p = \infty$!*

□

4. Measure theory

4.1. **Generalities.** At the highest level, measure theory is about providing a "better" integral (the Lebesgue integral) than the Riemann integral. It will be more powerful in that it will allow us to integrate more functions than the Riemann integral. Additionally, and perhaps most importantly, using it will allow us to have complete spaces of integrable functions: L^p will be a complete normed linear space (a Banach space) for every $p \in [1, \infty]$. This is not possible with the Riemann integral. Beyond these benefits, measure theory is the basis for probability theory, allowing us to make sense of randomness, which is essential in modern applied mathematics.

In a sense, the very basic goal of measure theory is simply to determine the "length" of sets. When the set is an interval, it is easy to say that its measure should be the difference between the endpoints:

$$m([a,b]) = m((a,b)) = m([a,b)) = m((a,b]) = b - a. \tag{4.1.1}$$

Some other easy cases are the emptyset and any infinite or half-infinite interval:

$$m(\emptyset) = 0 \quad \text{and} \quad m(\mathbb{R}) = m(a,\infty) = m(-\infty, b) = \infty.$$

Of course, any "reasonable" notion of length should involve *finite additivity*: whenever A, B are measurable[22] and disjoint then

$$m(A \cup B) = m(A) + m(B). \tag{4.1.2}$$

Additionally, we would expect "bigger" sets to have greater measure: $A \subset B$ implies that

$$m(A) \leq m(B). \tag{4.1.3}$$

These are not exactly the conditions that we put on generic measures, but at the moment we are only speaking roughly in order to build up our intuition.

It turns out that, if (4.1.2)-(4.1.3) are your only requirements, there are many measures!

Example 4.1.1. *(i) (Trivial measure) Define $m : \mathcal{P}^{\mathbb{R}} \to \mathbb{R}_+$ by:*

$$m(A) = \begin{cases} 0 & \text{if } A = \emptyset, \\ \infty & \text{otherwise.} \end{cases}$$

(ii) (Counting measure) Given any set A, defined

$$\#(A) = \begin{cases} \text{card}(A) & \text{if } A \text{ is finite}, \\ \infty & \text{if } A \text{ is infinite.} \end{cases}$$

(iii) (Delta measure) Fix a point x_0 in a space X. Then, we define

$$\delta_{x_0}(A) = \begin{cases} 1 & \text{if } x_0 \in A, \\ 0 & \text{if } x_0 \notin A. \end{cases}$$

Although this is a somewhat trivial example,

(iv) (Coin Flip) Roughly, we think of a coin flip as taking the value heads, which we will represent with 1, or tails, which we will represent as 0, each with probability $1/2$. This gives rise to the probability measure:

$$\mathbb{P} : \mathcal{P}^{\{0,1\}} \to \mathbb{R}_+$$

defined by

$$\mathbb{P}(\emptyset) = 0, \quad \mathbb{P}(\{0\}) = 1/2 \quad \mathbb{P}(\{1\}) = 1/2, \quad \text{and} \quad \mathbb{P}(\{0,1\}) = 1.$$

The first and last statement mean, respectively, that the probability of getting neither heads nor tails is zero and that the probability of getting at least one of heads or tails is 1.

There is no reason that the probabilities have to be balanced: one could look at a Bernoulli*(p) random variable for any p:*

$$\mathbb{P}(\emptyset) = 0, \quad \mathbb{P}(\{0\}) = 1 - p \quad \mathbb{P}(\{1\}) = p, \quad \text{and} \quad \mathbb{P}(\{0,1\}) = 1. \tag{4.1.4}$$

Notice that, actually, only the second and third equalities are require, while the first and third arise via (4.1.2)-(4.1.3).

[22] We will make this precise later, but for now we just mean that $m(A)$ and $m(B)$ "make sense."

Exercise 4.1.1. *Show this!*

(v) Sequence of n independent Bernoulli(p) random variables: Let X_i be the ith coin flip. It follows (4.1.4)*. On the other hand, we can define the probability measure on the sequence $(f_1, \dots, f_n)$ by:*

$$\mathbb{P}((f_1, \dots, f_n) = (a_1, \dots, a_n)) = p^{n_0}(1-p)^{n-n_0} \tag{4.1.5}$$

where $(a_1, \dots, a_n) \in \{0,1\}^n$ with

$$n_0 = |\{i \in \{1, \dots, n\} : a_i = 1\}|.$$

Accommodating (4.1.1) is much harder. Indeed, how do we measure the "length" of something that is not made up of intervals? For example, what is $m(\mathbb{Q})$? Recall that $\mathbb{Q}$ and $\mathbb{R} \setminus \mathbb{Q}$ are dense in $\mathbb{R}$, so that both have "holes everywhere." It is clear that we cannot write these as unions of disjoint intervals.

4.2. **Defining the Lebesgue measure.** The measure satisfying the natural property (4.1.1) is called the Lebesgue measure. It was (amazingly!) introduced by Henri Lebesgue in his PhD thesis at the beginning of the 20th century. For the moment, we will mainly be concerned with this measure. Later, we will widen our gaze to include more measures, such as those in Example 4.1.1.

4.2.1. **Some basic facts about sets.** Before we get started on the careful construction of the Lebesgue measure, let us recall some

Remark 4.2.1 (Set Operations)**.** *Some basic properties of $\cup$, $\cap$, and $\cdot^c$:*

- $A \cup B = B \cup A$;
- $A \cap B = B \cap A$;
- $(A \cup B) \cup C = A \cup (B \cup C)$;
- $(A \cap B) \cap C = A \cap (B \cap C)$;
- $(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$;
- $(A \cap B) \cup C = (A \cup C) \cap (B \cup C)$;
- $(A^c)^c = A$;
- $(A \cup B)^c = A^c \cap B^c$;
- $(A \cap B)^c = A^c \cup B^c$.

Let us begin by first working with "nice" sets, that is, those that can be written as a disjoint collection of intervals.

Proposition 4.2.2. *A set $W \subset \mathbb{R}$ is open and nonempty if and only if it is a* disjoint *collection of intervals: there is $N \in \mathbb{N}$, possibly infinite and $a_i < b_i$ such that*

$$W = \bigcup_{i=1}^{N} (a_i, b_i),$$

where $(a_i, b_i) \cap (a_j, b_j) = \emptyset$ when $i \neq j$. We allow for the possibilities that $a_i = -\infty$ and $b_i = \infty$.

Proof. $\Longleftarrow$: This direction is obvious thanks to our earlier work in Section 3.

$\Longrightarrow$: Fix any arbitrary open set W. Let $q_1, q_2, \dots$ be an enumeration of the elements of $W \cap \mathbb{Q}$ that exists due to the countability of the rational numbers[23]. We collect intervals (a_i, b_i) inductively as follows.

For $i = 1$, we let

$$a_1 = \inf\{t < q_1 : (t, q_1) \subset W\} \quad \text{and} \quad b_1 = \sup\{t > q_1 : (q_1, t) \subset W\}.$$

The openness of W ensures that a_1 and b_1 exist and satisfy $a_1 < b_1$. Also, it follows from their definition that $(a_1, b_1) \subset W$. We let $n_i^{(1)}$ denote an enumeration of the subsequence of $q_1, q_2, \dots$ of elements in $(W \cap \mathbb{Q}) \setminus (a_1, b_1)$:

$$\{q_{n_1^{(2)}}, q_{n_2^{(2)}}, q_{n_3^{(2)}}, q_{n_4^{(2)}}, \dots\} = (W \cap \mathbb{Q}) \setminus (a_1, b_1).$$

We use the convention that $n_k^{(1)} = n_k$. We point out that $q_1, \dots, q_{n_1^{(1)}-1} \in (a_1, b_1)$ and $a_1, b_1 \notin W$. (Indeed, if, for example, $a_1 \in W$, then, by the definition of openness, there is $r > 0$ such that $(a_1 - r, a_1 + r) \subset W$, which implies that $a_1 - r$ is a "smaller infimum" than a_1, which is a contradiction.)

For general i, we assume that we have already defined disjoint $(a_1, b_1), \dots, (a_i, b_i)$ as above. If

$$W = \bigcup_{k=1}^{i} (a_k, b_k),$$

we are finished. Otherwise, we define:

$$a_{i+1} = \inf\{t < q_{n_1^{(i)}} : (t, q_{n_1^{(i)}}) \subset W\} \quad \text{and} \quad b_{i+1} = \sup\{t > q_{n_1^{(i)}} : (q_{n_1^{(i)}}, t) \subset W\}.$$

Similarly as above, the openness of W ensures that a_{i+1} and b_{i+1} exist and satisfy $a_{i+1} < b_{i+1}$. Also, it follows from their definition that $(a_{i+1}, b_{i+1}) \subset W$. We let $n_1^{(i+2)}, n_2^{(i+2)}, \dots$ denote an enumeration of the subsequence of $q_{n_1^{(i+1)}}, q_{n_2^{(i+1)}}, \dots$ of elements in

$$(W \cap \mathbb{Q}) \setminus \left(\bigcup_{\ell=1}^{i+1} (a_\ell, b_\ell) \right);$$

that is

$$\{q_{n_1^{(i+2)}}, q_{n_2^{(i+2)}}, q_{n_3^{(i+2)}}, q_{n_4^{(i+2)}}, \dots\} = (W \cap \mathbb{Q}) \setminus \left(\bigcup_{\ell=1}^{i+1} (a_\ell, b_\ell) \right).$$

We point out that $a_{i+1}, b_{i+1} \notin W$ and

$$q_1, \dots, q_{n_1^{(i+1)}-1} \in \bigcup_{\ell=1}^{i+1} (a_\ell, b_\ell). \tag{4.2.1}$$

We claim that (a_{i+1}, b_{i+1}) is disjoint from (a_ℓ, b_ℓ) for all $\ell < i + 1$. We argue by contradiction, assuming that $(a_{i+1}, b_{i+1}) \cap (a_\ell, b_\ell) \neq \emptyset$. Without loss of generality, assume that $q_{n_1^{(i+1)}} < q_{n_1^{(\ell)}}$. Then

$$W \not\ni a_\ell \in (q_{n_1^{(i+1)}}, q_{n_1^{(\ell)}}) \subset (\min\{a_{i+1}, a_\ell\}, \max\{b_{i+1}, b_\ell\}) = (a_{i+1}, b_{i+1}) \cup (a_\ell, b_\ell) \subset W.$$

[23] One might worry that there is not an infinite number of elements in $W \cap \mathbb{Q}$. This can be seen as follows. Since W is open, there is $r > 0$ and $x_0 \in W$ such that $B_r(x_0) \subset W$. Since $\mathbb{Q}$ is dense in $\mathbb{R}$, there is $q \in B_r(x_0)$. Thus, $q < x_0 + r$, and, as a result, there is N sufficiently large that, $q + 1/N < x_0 + r$. It follows that $x_0 - r < q < q + 1/n < x_0 + r$ for every $n \geq N$, which gives an infinite set of points in $B_r(x_0) \subset W$.

This is a contradiction.

If N is finite, the proof is finished, as noted at the beginning of the inductive step. Otherwise, we must show that

$$W = \bigcup_{i=1}^{\infty} (a_i, b_i).$$

Clearly "$\supset$" holds. If "$\subset$" does not hold, then there is $w \in W$ such that $w \notin (a_i, b_i)$ for every i. Since W is open, there is $r > 0$ such that $(w - r, w + r) \subset W$. Since $a_i, b_i \notin W$, it follows that $a_i, b_i \notin (w - r, w + r)$. It follows that

$$(w - r, w + r) \subset W \setminus \left(\bigcup_{i=1}^{\infty} (a_i, b_i) \right). \tag{4.2.2}$$

Then there is ℓ_0 such that $q_{\ell_0} \in (w - r, w + r)$ and then we must have that, by (4.2.1),

$$q_{\ell_0} \in \bigcup_{\ell=1}^{\ell_0+1} (a_\ell, b_\ell)$$

which contradicts (4.2.2). □

Thus, for any open set U we can declare that

$$m(U) = \sum_{i=1}^{N} |b_i - a_i|,$$

where $a_1, b_1, a_2, b_2, \ldots$ is from the decomposition in Proposition 4.2.2... But what about the rest of the sets $A \subset \mathbb{R}$?

Philosophy: Two "close" sets should have measures that are "close." In other words, we expect our measure to be continuous. This is apparent when we look at the following: let $\varepsilon > 0$ and $a < b$, then

$$|m((a, b + \varepsilon)) - m((a, b))| = |(b + \varepsilon - a) - (b - a)| = \varepsilon.$$

Our goal is to use this to define the measure of "nice" sets:

$$m(A) = \inf_{U \supset A, \text{open}} m(U).$$

While the comment about "niceness" might seem unimportant, let me stress that not all sets will be "nice," and one must take care with this issue!

4.2.2. **Outer measure.** We begin with a preliminary form of measure:

Definition 4.2.3 (Outer measure)**.** *The* outer measure *is* $m^* : \mathcal{P}^{\mathbb{R}} \to \mathbb{R}_+$ *defined by*

$$m^*(A) = \inf \left\{ \sum_{i=1}^{\infty} |b_i - a_i| : A \subset \bigcup_{i=1}^{\infty} (a_i, b_i) \right\}.$$

We notice that this works nicely for open sets. Thanks to Proposition 4.2.2, we immediately see that, if U is open:

$$m^*(U) = m(U).$$

Actually, there is a tiny bit to prove here, but let us ignore that for now. Recall that m is, at this point, only defined on open sets. Further, it is not hard to prove that

$$m^*(A) = \inf_{U \text{ open}, U \supset A} m(U). \tag{4.2.3}$$

Exercise 4.2.1. *Do this!*

Rephrased, we have that, if $U \supset A$ is an open set then $m^*(A) \leq m(U)$ and, for all $\varepsilon > 0$, there is an open set $U_\varepsilon \supset A$ such that

$$m^*(A) \geq m(U_\varepsilon) - \varepsilon.$$

The advantage to this is that it is defined on all sets and will allow us to identify "nice" sets. The disadvantage is that it will lead to some nasty contradictions before we reduce to "nice" sets.

We point out that an immediate consequence of (4.2.3) is that, if $A \subset B$, then

$$m^*(A) \leq m^*(B). \tag{4.2.4}$$

A nice property of outer measures (that the Lebesgue measure will inherit!) is *countable subadditivity*.

Proposition 4.2.4. *Let $A_1, A_2, \ldots$ be a collection of sets. Then, letting*

$$\mathcal{A} = \bigcup_{i=1}^{\infty} A_i,$$

we have

$$m^*(\mathcal{A}) \leq \sum_{i=1}^{\infty} m^*(A_i).$$

Proof. Before beginning, we note that there is nothing to prove if

$$\sum_{i=1}^{\infty} m^*(A_i) = \infty,$$

hence, we assume that this sum is finite.

Fix $\varepsilon > 0$. For each i, there is collection of intervals $(a_{i,k}, b_{i,k})$ with $k = 1, 2, \ldots$ such that

$$m^*(A_i) \geq \sum_{k=1}^{\infty} |b_{i,k} - a_{i,k}| - \frac{\varepsilon}{2^i} \qquad \text{and } A_i \subset \bigcup_{k=1}^{\infty} (a_{i,k}, b_{i,k}).$$

Notice that $(a_{i,k}, b_{i,k})$ is a countable collection of intervals whose union contains $\mathcal{A}$. Hence,

$$m^*(\mathcal{A}) \leq \sum_{i=1}^{\infty} \sum_{k=1}^{\infty} |b_{i,k} - a_{i,k}| \leq \sum_{i=1}^{\infty} \left(m^*(A_i) + \frac{\varepsilon}{2^i} \right) = \sum_{i=1}^{\infty} m^*(A_i) + \varepsilon.$$

Since this is true for all ε, the claim follows. □

Notice that, for any open set U, $m^*(U) = m(U)$. There are other sets that we can find the outer measure of in a relatively straightforward way.

Proposition 4.2.5. *The set of rational numbers has outer measure* 0*:* $m^*(\mathbb{Q}) = 0$*. Further, any countable set has outer measure zero.*

Warning: not every measure zero set is countable!

Proof. Fix any $\varepsilon > 0$ and let $\mathbb{Q} = \{q_1, q_2, \dots\}$. For each i, let

$$a_i = q_i - \frac{\varepsilon}{2^{i+1}} \qquad \text{and} \qquad b_i = q_i + \frac{\varepsilon}{2^{i+1}}.$$

It is clear that

$$\mathbb{Q} \subset \bigcup_{i=1}^{\infty} (a_i, b_i),$$

so that

$$0 \leq m^*(\mathbb{Q}) \leq \sum_{i=1}^{\infty} |b_i - a_i| = \sum_{i=1}^{\infty} \frac{\varepsilon}{2^i} = \varepsilon.$$

Since this is true for all ε, it follows that $m^*(\mathbb{Q}) = 0$. □

This sort of proof is very common when showing something has measure zero. Try it yourself!

Exercise 4.2.2. *Suppose that* $A_1, A_2, \dots$ *have outer measure zero. Show that* $\mathcal{A} = A_1 \cup A_2 \cup \cdots$ *has outer measure zero. Use this to show that any countable set has outer measure zero.*

Exercise 4.2.3. *A metric on sets. For any* $A, B \subset \mathbb{R}$*, let*

$$\tilde{d}(A, B) = m^*(A \Delta B) = m^*((A \setminus B) \cup (B \setminus A)).$$

Show that $\tilde{d}$ *may be infinite and that there are* $A \neq B$ *such that* $\tilde{d}(A, B) = 0$*. Instead of considering* $\mathcal{P}^{\mathbb{R}}$*, let* $\mathcal{P}^{\mathbb{R}}/\sim$ *where* $A \sim B$ *if* $m^*(A \Delta B) = 0$*. We say that then* A *and* B *are the same up to a set of measure zero. Let*

$$d([A], [B]) = \frac{\tilde{d}(A, B)}{1 + \tilde{d}(A, B)}.$$

Show that d *is a metric on* $\mathcal{P}^{\mathbb{R}}/\sim$*.*

4.2.3. **Lebesgue measurability.** Here we answer the question: what are the "nice" sets on which we can define the measure? First, notice that Proposition 4.2.4 implies that

$$m^*(E) \leq m^*(A \cap E) + m^*(A^c \cap E). \tag{4.2.5}$$

We would expect that equality holds (since $A \cap E$ and $A^c \cap E$ are disjoint), but we have not proven that yet. In fact, it is not always true! If A and A^c are "very intertwined" then a cover of A by integrals might include "a lot" of points of A^c as well. It takes some effort to construct such messy sets, and essentially every set you know does not have this bad property. So let us just restrict our universe to just the nice sets where equality holds in (4.2.5).

Definition 4.2.6 (Measurability)**.** *A set* $A \subset \mathbb{R}$ *is* (Lebesgue) measurable *if, for every* $E \subset \mathbb{R}$*,*

$$m^*(E) = m^*(A \cap E) + m^*(A^c \cap E).$$

We denote the set of measurable sets by $\mathcal{M}$*. If* $A \in \mathcal{M}$*, then we denote its measure*

$$m(A) = m^*(A).$$

We note that, due to (4.2.5), a set A is measurable if and only if

$$m^*(E) \geq m^*(A \cap E) + m^*(A^c \cap E). \tag{4.2.6}$$

Example 4.2.7. *(i)* **Measure zero sets are measurable.** *Note that this includes the empty set, finite sets, and countable sets.*

Assume that $m^*(A) = 0$. Fix any set $E \subset \mathbb{R}$. Then

$$m^*(A \cap E) + m^*(A^c \cap E) = 0 + m^*(A^c \cap E) \leq m^*(E).$$

Here we used (4.2.4) twice. The above is precisely (4.2.6), which yields the measurability of A.

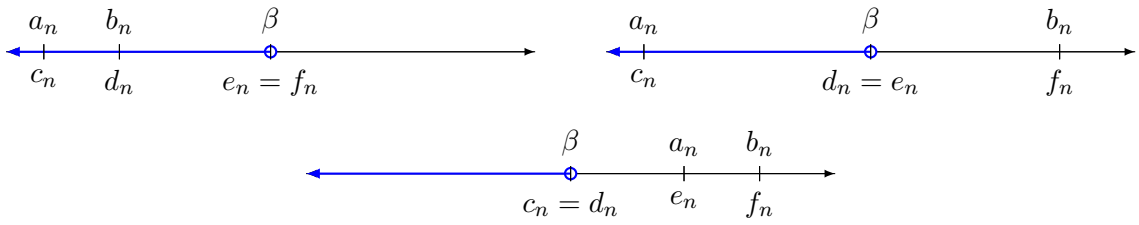
(ii) **The half-open interval $(-\infty, \beta)$ is measurable.** *Fix any set $E \subset \mathbb{R}$. Take any $\varepsilon > 0$ and $(a_1, b_2), (a_2, b_2), \ldots$ to be a collection of open sets such that*

$$\sum_{n=1}^{\infty} |b_n - a_n| \leq m^*(E) + \varepsilon \qquad \text{and} \qquad E \subset \bigcup_{n=1}^{\infty} (a_n, b_n).$$

We now define two new collections of intervals as follows. For each n, if $b_n \leq \beta$, we let

$$c_n = \min\{a_n, \beta\}, \quad d_n = \min\{b_n, \beta\}, \quad e_n = \max\{a_n, \beta\}, \quad \text{and} \quad f_n = \max\{b_n, \beta\}.$$

Here are the three different situations that the above choices represent:



Notice that

$$f_n - e_n + d_n - c_n = b_n - a_n.$$

Additionally, notice that

$$(-\infty, \beta) \cap E \subset \bigcup_{n=1}^{\infty} (c_n, d_n) \quad \text{and} \quad (\beta, \infty) \cap E \subset \bigcup_{n=1}^{\infty} (e_n, f_n).$$

Hence,

$$\begin{aligned} m^*((-\infty, \beta) \cap E) + m^*((-\infty, \beta)^c \cap E) &= m^*((-\infty, \beta) \cap E) + m^*([\beta, \infty) \cap E) \\ &\leq m^*((-\infty, \beta) \cap E) + m^*(\{\beta\}) + m^*((\beta, \infty) \cap E) \\ &\leq \sum_{n=1}^{\infty} (d_n - c_n) + 0 + \sum_{n=1}^{\infty} (f_n - e_n) = \sum_{n=1}^{\infty} (b_n - a_n) \leq m^*(E) + \varepsilon. \end{aligned}$$

Taking $\varepsilon \to 0$, we arrive at (4.2.6), which proves that $(-\infty, \beta)$ is measurable.

We can get many more measurable sets by observing measurability is preserved under the operations of set complementation, countable unions, and countable intersections. This, along with Example 4.2.7, shows that all open and closed sets are measurable, as well as, essentially, any set we have ever come across in the wild before.

We prove this in several results below. We first consider set complements.

Proposition 4.2.8. *If $A \subset \mathbb{R}$ is measurable, then so is A^c.*

Proof. This is essentially obvious after noting that $(A^c)^c = A$. Indeed, for any E, we have, by the measurability of A,

$$m^*(E) = m^*(A \cap E) + m^*(A^c \cap E) = m^*((A^c)^c \cap E) + m^*(A^c \cap E).$$

This is precisely the definition of measurability of A^c. □

In order to establish the result about countable unions and intersections, we first prove a preliminary version of it for finite unions and intersections.

Lemma 4.2.9. *Let $N \in \mathbb{N}$ and $A_1, \ldots, A_N$ be measurable sets.*

(i) The following sets are measurable:

$$\mathcal{U}_N = \bigcup_{i=1}^{N} A_i \quad \text{and} \quad \mathcal{I}_N = \bigcap_{i=1}^{N} A_i.$$

(ii) If we further assume that the sets are pairwise disjoint (i.e., $A_i \cap A_j = \emptyset$ for $i \neq j$), then

$$m^*(E \cap \mathcal{U}_N) = \sum_{i=1}^{N} m^*(E \cap A_i) \tag{4.2.7}$$

for any set $E \subset \mathbb{R}$. In particular,

$$m(\mathcal{U}_N) = \sum_{i=1}^{N} m(A_i).$$

Proof of Lemma 4.2.9.(i). We need only show the measurability of $\mathcal{U}_N$; that is, that any finite union of measurable sets is itself measurable. Indeed, let us show how the measurability of $\mathcal{I}_N$ follows from this fact. Observe that

$$\mathcal{I}_N^c = \bigcup_{i=1}^{N} A_i^c$$

and that

$$\mathcal{I}_N = (\mathcal{I}_N^c)^c. \tag{4.2.8}$$

Observe that, for all i, A_i^c is measurable, by Proposition 4.2.8. Using this and the measurability of finite unions of measurable sets, we see that $\mathcal{I}_N^c$ is measurable. Finally, by (4.2.8) and Proposition 4.2.8, we deduce that $\mathcal{I}_N$ is measurable.

Next, we prove the measurability of $\mathcal{U}_N$ by induction. The base case $N = 1$ is immediate since $\mathcal{U}_1 = A_1$, which is measurable by assumption. Then we consider the inductive step. Suppose that $N > 1$ and that $\mathcal{U}_{N-1}$ and A_N are measurable. We show that $\mathcal{U}_N$ is measurable.

Fix any set $E \subset \mathbb{R}$. Observe some basic identities:

$$\begin{aligned} &\mathcal{U}_N = \mathcal{U}_{N-1} \cup A_N, \\ &(\mathcal{U}_{N-1} \cup A_N) \cap E = (\mathcal{U}_{N-1} \cap E) \cup (\mathcal{U}_{N-1}^c \cap A_N \cap E), \qquad \text{and} \\ &(\mathcal{U}_{N-1} \cup A_N)^c \cap E = \mathcal{U}_{N-1}^c \cap A_N^c \cap E. \end{aligned} \tag{4.2.9}$$

Fix any set E. Then

$$\begin{aligned} &m^*((\mathcal{U}_{N-1} \cup A_N) \cap E) + m^*((\mathcal{U}_{N-1} \cup A_N)^c \cap E) \\ &\qquad = m^*\left((\mathcal{U}_{N-1} \cap E) \cup (\mathcal{U}_{N-1}^c \cap A_N \cap E)\right) + m^*\left(\mathcal{U}_{N-1}^c \cap A_N^c \cap E\right) \\ &\qquad \leq m^*(\mathcal{U}_{N-1} \cap E) + m^*(\mathcal{U}_{N-1}^c \cap A_N \cap E) + m^*(\mathcal{U}_{N-1}^c \cap A_N^c \cap E) \\ &\qquad \leq m^*(\mathcal{U}_{N-1} \cap E) + m^*\left(A_N \cap (\mathcal{U}_{N-1}^c \cap E)\right) + m^*\left(A_N^c \cap (\mathcal{U}_{N-1}^c \cap E)\right). \end{aligned} \tag{4.2.10}$$

In the first equality, we simply re-wrote the unions and intersections using (4.2.9), and in the first inequality, we used (4.2.5). The last inequality is simply a reordering of the intersections for expository reasons. Next, the measurability of A_N yields

$$m^*\left(A_N \cap (\mathcal{U}_{N-1}^c \cap E)\right) + m^*\left(A_N^c \cap (\mathcal{U}_{N-1}^c \cap E)\right) = m^*(\mathcal{U}_{N-1} \cap E).$$

Hence, (4.2.10) becomes

$$\begin{aligned} m^*((\mathcal{U}_{N-1} \cup A_N) \cap E) + m^*((\mathcal{U}_{N-1} \cup A_N)^c \cap E) \\ \leq m^*(\mathcal{U}_{N-1} \cap E) + m^*(\mathcal{U}_{N-1} \cap E). \end{aligned} \tag{4.2.11}$$

We then similarly using the measurability of $\mathcal{U}_{N-1}$ to deduce that the right hand side of (4.2.11) is $m^*(E)$. We, thus, deduce that

$$\begin{aligned} &m^*((\mathcal{U}_{N-1} \cup A_N) \cap E) + m^*((\mathcal{U}_{N-1} \cup A_N)^c \cap E) \\ &\qquad \leq m^*(\mathcal{U}_{N-1} \cap E) + m^*(\mathcal{U}_{N-1}^c \cap A_N \cap E) + m^*(\mathcal{U}_{N-1}^c \cap A_N^c \cap E) \\ &\qquad = m^*(\mathcal{U}_{N-1} \cap E) + m^*(\mathcal{U}_{N-1}^c \cap E) = m^*(E). \end{aligned}$$

This is precisely (4.2.6), which completes the proof of the measurability of $\mathcal{U}_N$. □

Proof of Lemma 4.2.9.(ii). Again, we proceed by induction. The case $N = 1$ is immediate since $\mathcal{U}_1 = A_1$. Next, we consider the inductive step. Fix any $E \subset \mathbb{R}$. Suppose that $N > 1$ and that (4.2.7) holds for $N - 1$; that is,

$$m^*(E \cap \mathcal{U}_{N-1}) = \sum_{i=1}^{N-1} m^*(E \cap A_i). \tag{4.2.12}$$

We notice that

$$\begin{aligned} &(E \cap (\mathcal{U}_{N-1} \cup A_N)) \cap \mathcal{U}_{N-1} = E \cap \mathcal{U}_{N-1} \qquad \text{and} \\ &(E \cap (\mathcal{U}_{N-1} \cup A_N)) \cap \mathcal{U}_{N-1}^c = E \cap A_N. \end{aligned} \tag{4.2.13}$$

In the second equality, we used that $\mathcal{U}_{N-1}$ and A_N are disjoint.

Using the measurability of $\mathcal{U}_{N-1}$ (due to Lemma 4.2.9.(i)) and then (4.2.13), we have

$$\begin{aligned} m^*(E \cap (\mathcal{U}_{N-1} \cup A_N)) &= m^*((E \cap (\mathcal{U}_{N-1} \cup A_N)) \cap \mathcal{U}_{N-1}) + m^*((E \cap (\mathcal{U}_{N-1} \cup A_N)) \cap \mathcal{U}_{N-1}^c) \\ &= m^*(E \cap \mathcal{U}_{N-1}) + m^*(E \cap A_N). \end{aligned}$$

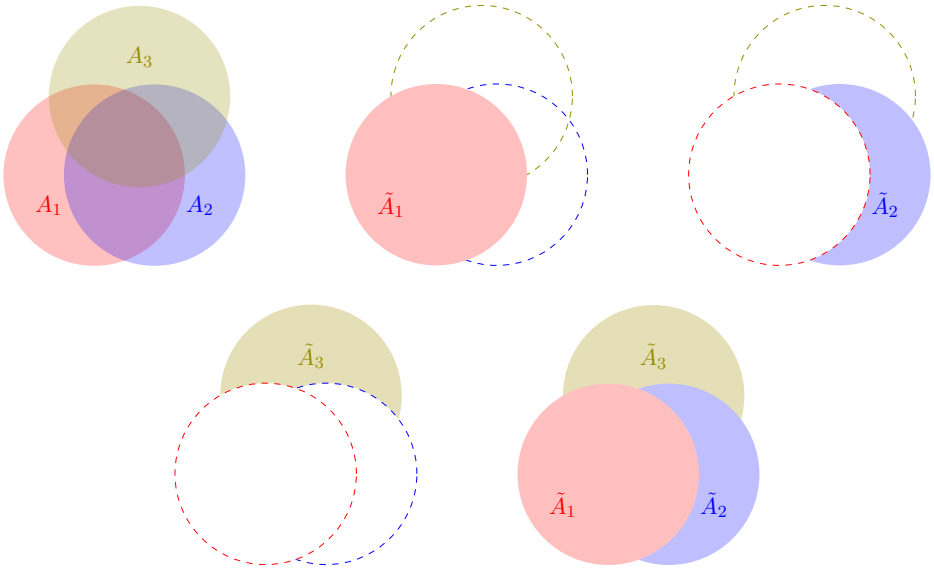


Figure 5: A cartoon of the definition of $\tilde{A}_1$, $\tilde{A}_2$, $\tilde{A}_3$.

We then apply (4.2.12) to deduce that

$$\begin{aligned} m^*(E \cap (\mathcal{U}_{N-1} \cup A_N)) &= m^*(E \cap \mathcal{U}_{N-1}) + m^*(E \cap A_N) \\ &= \left(\sum_{i=1}^{N-1} m^*(E \cap A_i) \right) + m^*(E \cap A_N) = \sum_{i=1}^{N} m^*(E \cap A_i). \end{aligned}$$

The proof is complete. □

We now bootstrap this to the countably infinite case.

Proposition 4.2.10. *Suppose that $A_1, A_2, \ldots$ are measurable sets.*

(i) The following sets are measurable:

$$\mathcal{U} = \bigcup_{i=1}^{\infty} A_i \quad \text{and} \quad \mathcal{I} = \bigcap_{i=1}^{\infty} A_i.$$

(ii) If we further assume that the sets are pairwise disjoint, then

$$m(\mathcal{U}) = \sum_{i=1}^{\infty} m(A_i).$$

Proof of Proposition 4.2.10.(i). As in the proof Proposition 4.2.10.(i), we need only show the measurability of $\mathcal{U}$.

We define a new collection of sets

$$\tilde{A}_n = A_n \setminus (A_1 \cup \cdots \cup A_{n-1}) = A_n \cap (A_1^c \cap \cdots \cap A_{n-1}^c).$$

The advantage to this collection is encapsulated in the following exercise; that is, this collection is pairwise disjoint and its union is still $\mathcal{U}$.

Exercise 4.2.4. *Show that the sets $\tilde{A}_1, \tilde{A}_2, \ldots$ have the property that*

$$\tilde{A}_n \cap \tilde{A}_m = \emptyset \quad \text{if } n \neq m \qquad \text{and} \qquad \mathcal{U} = \bigcup_{i=1}^{\infty} \tilde{A}_i.$$

By parts (i) and (ii), each $\tilde{A}_i$ is measurable. Using the measurability of

$$\mathcal{U}_N = \bigcup_{i=1}^{N} \tilde{A}_i = \bigcup_{i=1}^{N} A_i,$$

we find

$$m^*(E) = m^*(\mathcal{U}_N \cap E) + m^*(\mathcal{U}_N^c \cap E).$$

Since $\mathcal{U}_N^c \subset \mathcal{U}^c$, we have

$$m^*(E) \geq m^*(\mathcal{U}_N \cap E) + m^*(\mathcal{U}^c \cap E).$$

Using the fact that each $\tilde{A}_i$ is disjoint as well as part (ii), we have

$$m^*(E) \geq \sum_{i=1}^{N} m^*(\tilde{A}_i \cap E) + m^*(\mathcal{U}^c \cap E).$$

Since this holds for every N, we may take N to infinity to find

$$m^*(E) \geq \sum_{i=1}^{\infty} m^*(\tilde{A}_i \cap E) + m^*(\mathcal{U}^c \cap E) \geq m^*(\mathcal{U} \cap E) + m^*(\mathcal{U}^c \cap E).$$

In the second inequality, we used Proposition 4.2.4. This completes the proof. □

Proof of Proposition 4.2.10.(ii). From Proposition 4.2.4 and the measurability of $\mathcal{U}$, we immediately get that

$$m(\mathcal{U}) \leq \sum_{i=1}^{\infty} m(A_i). \tag{4.2.14}$$

On the other hand, for any N, we have $\mathcal{U}_N \subset \mathcal{U}$, which, along with Lemma 4.2.9.(ii) implies that

$$\sum_{i=1}^{N} m(A_i) = m(\mathcal{U}_N) \leq m(\mathcal{U}).$$

Since this holds for all N, we may take the limit as $N \to \infty$ to deduce that

$$\sum_{i=1}^{\infty} m(A_i) \leq m(\mathcal{U}). \tag{4.2.15}$$

The combination of (4.2.14) and (4.2.15) completes the proof. □

Before giving the proof of this, we see what it tells us about measurable sets. First, we note that it is not easy to understand what the space of Lebesgue measurable sets looks like. It is quite large, and contains some very messy sets! Second, let us look at some new examples of measurable sets.

Example 4.2.11. *(i)* **Open sets** –

Exercise 4.2.5. *Homework problem!*

(ii) **Closed sets** – *If K is a closed set, then K^c is open and, hence, measurable. It follows from Proposition 4.2.8 that $K = (K^c)^c$ is open.*

(iii) **The irrational numbers** – $\mathbb{R} \setminus \mathbb{Q} = \mathbb{Q}^c$.

(iv) **Countable unions and intersections of open sets and closed sets** – *This is an important "σ-algebra" (we define this below, see Definition 4.3.4) called the Borel σ-algebra. It yields all sorts of sets, such as*

$$[0,1] \cap (\mathbb{R} \setminus \mathbb{Q}), \quad \{z : \sin(z) \in \mathbb{Q}\}, \qquad \text{and} \qquad \bigcup_{k=1}^{\infty} \bigcup_{n=1}^{\infty} \left[k + \frac{1}{10n}, k + \frac{1}{5n} \right].$$

Exercise 4.2.6. *You may find it useful to think hard about why each of the above arises in the way claimed.*

Exercise 4.2.7 (The Cantor set)**.** *We form a subset $C \subset [0,1]$ in the following way. Let $C_0 = [0,1]$. At each step i, $C_i \subset C_{i-1}$ obtained by removing the middle third of each interval making up C_{i-1}:*

$$\begin{aligned}
C_1 &= [0, 1/3] \cup [2/3, 1], \\
C_2 &= [0, 1/9] \cup [2/9, 3/9] \cup [6/9, 7/9] \cup [8/9, 1], \\
&\vdots \\
C_\ell &= \bigcap_{n=1}^{\ell} \bigcup_{i=0}^{3^{n-1}-1} \left(\left[\frac{3i}{3^n}, \frac{3i+1}{3^n} \right] \bigcup \left[\frac{3i+2}{3^n}, \frac{3i+3}{3^n} \right] \right), \\
&\vdots
\end{aligned}$$

Another way to characterize C_k is that it is all real numbers between 0 and 1 (inclusive) with a ternary *expansion in which the first k terms in the expansion have a numerator of 0 or 2:*

$$C_k = \left\{ x \in [0,1] : x = \sum_{i=1}^{\infty} \frac{\alpha_i}{3^i}, \text{ where } \alpha_i \in \{0,2\} \text{ if } i \leq k \text{ and } \alpha_i \in \{0,1,2\} \text{ for all } i \right\}.$$

Let

$$C = \bigcap_{k=1}^{\infty} C_k.$$

Show that C is uncountable and has measure zero.

4.2.4. **An example of a non-measurable set.** First, we must define modular arithmetic. This is an intuitively simple concept that, in our case of "mod 1 arithmetic," essentially boils down to "forgetting" the integer part of a number; for example, $.7 + .8 \mod 1 = .5$ since it "usually" is 1.5. More precisely, we say that, for any $x, y \in \mathbb{R}$,

$$x + y \mod 1 = x + y - n$$

where $n = \max\{n' \in \mathbb{Z} : n' \leq x + y\}$. One can check that all of the usual manipulations of addition hold in this setting.

Next, define an equivalence relation $\sim$ by

$$x \sim y \iff \text{ there is } q \in \mathbb{Q} \text{ such that } x = y + q.$$

We can partition $[0,1)$ with this equivalence relation into disjoint equivalence classes $[x] \subset [0,1)$ for each $x \in [0,1)$.

Using the Axiom of Choice, we obtain a set B that contains exactly one element from each equivalence class $[x]$ defined using $\sim$. In other words, for every $p, \tilde{p} \in P$, either $p = \tilde{p}$ or $p \not\sim \tilde{p}$.

Let $q_1, q_2, \dots$ be an enumeration of $\mathbb{Q} \cap [0,1)$. Define

$$P_i = P + q_i \bmod 1 = \{z \in [0,1) : z = p + q_i \mod 1 \text{ for some } p \in P\}.$$

The goal is to show that P is not measurable.

We proceed by contradiction, assuming that P is measurable. Then $m(P) = m(P_i)$ for all i (see Lemma 4.2.12). On the other hand, (see Lemma 4.2.13) the P_i are disjoint and their union is $[0,1)$, so that, by Proposition 4.2.10.(ii),

$$1 = m([0,1)) = m\Big(\bigcup_{i=1}^{\infty} P_i\Big) = \sum_{i=1}^{\infty} m(P_i) = \sum_{i=1}^{\infty} m(P).$$

If $m(P) = 0$, the right hand side is zero, which is a contradiction. If $m(P) > 0$, the right hand side is infinite, which is also a contradiction.

Let us now prove the two component pieces of this argument. To show that $P_i = P + q_i \mod 1$ is measurable (if we, by contradiction, assume that P is measurable), we establish a more general claim.

Lemma 4.2.12. *Suppose that A is measurable and $b \in \mathbb{R}$.*

(i) The set $A + b$ is measurable and $m(A) = m(A + b)$.

(ii) If $A \subset [0,1)$ and $b \in [0,1)$. then so is $A_b := A + b \mod 1$ is as well and $m(A_b) = m(A)$.

Proof of (i). We first show the following claim:

$$\text{for any } S \subset \mathbb{R} \text{ and } b \in \mathbb{R}, \quad m^*(S) = m^*(S + b). \tag{4.2.16}$$

To prove this, we simply show that

$$m^*(S + b) \subset m^*(S).$$

Indeed, since $S = (S - b) + b$ then the opposite inequality follows. Fix any $\varepsilon > 0$ and any $(a_1, b_1), (a_2, b_2,) \dots$ such that

$$S \subset \bigcup_{i=1}^{\infty} (a_i, b_i) \qquad \text{and} \qquad \sum_{i=1}^{\infty} |\beta_i - \alpha_i| + \sum_{i=1}^{\infty} |b_i - a_i| \leq m^*(S) + \varepsilon.$$

collection. Then

$$S + b \subset \bigcup_{i=1}^{\infty} (a_i + b, b_i + b).$$

Hence,

$$m^*(S+b) \le \sum_{i=1}^{\infty} |\beta_i - \alpha_i| + \sum_{i=1}^{\infty} |(b_i+b)-(a_i+b)| = \sum_{i=1}^{\infty} |\beta_i - \alpha_i| + \sum_{i=1}^{\infty} |b_i - a_i| \le m^*(S) + \varepsilon.$$

Then (4.2.16) follows by taking $\varepsilon \to 0$.

Fix any set $E \subset \mathbb{R}$. Notice that

$$\begin{aligned}(A+b)\cap E &= (A\cap(E-b))+b \qquad \text{and}\\ (A+b)^c\cap E &= (A^c+b)\cap E = (A^c\cap(E-b))+b.\end{aligned}$$

Using this, followed by (4.2.16), the measurability of A, and (4.2.16) again, we find

$$\begin{aligned}m^*((A+b)\cap E)+m^*((A+b)^c\cap E) &= m^*((A\cap(E-b))+b)+m^*((A^c\cap(E-b))+b)\\ &= m^*((A\cap(E-b)))+m^*((A^c\cap(E-b)))\\ &= m^*(E-b) = m^*(E).\end{aligned}$$

Thus, $A+b$ is measurable. That $m(A+b) = m(A)$ follows from (4.2.16). This completes the proof. □

Proof of (ii). Let $A_1 = A\cap[0,1-b)$ and $A_2 = A\cap[1-b,1)$. Notice that A_1 and A_2 are measurable, disjoint, and their union is A. Hence,

$$m(A) = m(A_1) + m(A_2).$$

Moreover, notice that

$$A+b \mod 1 = (A_1+b)\cup(A_2+b-1),$$

and the two sets on the right hand side are disjoint. This, along with Lemma 4.2.12.(i) yields

$$m(A+b) = m(A_1+b)+m(A_2+b-1) = m(A_1)+m(A_2) = m(A).$$

The proof is complete. □

Next we show that the P_i are pairwise disjoint and their union is $[0,1)$.

Lemma 4.2.13. *If $i \ne j$, then $P_i \cap P_j = \emptyset$. Further,*

$$\bigcup_{i=1}^{\infty} P_i = [0,1). \tag{4.2.17}$$

Proof. We begin by showing the disjointness first. We argue by contradiction. Fix $i \ne j$ and suppose there is

$$z \in P_1 \cap P_j.$$

Then there exist $p_i \in P_i$ and $p_j \in P_j$ such that

$$z = p_i + q_i \mod 1 = p_j + q_j \mod 1.$$

Thus, we have

$$p_i - p_j = q_j - q_i \mod 1.$$

Recall that $q_i - q_j \mod 1 \neq 0$. It follows that $p_i - p_j$ differ by a non-zero rational number, which implies that $p_j \sim p_i$. This contradicts the construction of P. Thus, $P_i \cap P_j = \emptyset$.

Now we show that (4.2.17) holds. It is clear that "$\subset$" holds, by construction. We show that "$\supset$" holds as well. Take any $x \in [0, 1)$. We need only find i such that $x \in P_i$. By construction of P, there exists $p \in P$ such that $p \sim x$. By definition of $\sim$, there is i such that $x = p + q_i \mod 1$. Thus, $x \in P_i$. This completes the proof. □

4.3. **General theory of measures.** Let us move beyond the notion of the Lebesgue measure. As we see below, there are many measures out there that are important in the world. Perhaps the largest source of these is that of probability measures (which formalize random events). Of course, *the reader is encouraged, on the first pass through this section, to special all theorems to the Lebesgue measure* for the sake of intuition.

In order to carefully define the general notion of a measure on an arbitrary set X, we need to first carefully define which sets we can measure! As we saw in the previous section (Section 4.2.4), it may not be possible to assign measure *every* set. Which sets should we be able to measure? Certainly, we should have a notion of the size of the whole space; i.e., X should be measurable. Of course, if we can measure a set A and we can measure X, we should be able to measure its complement $A^c = X \setminus A$. Finally, if we can measure a (countable) collection of sets, we should be able to measure their union just by adding.

This leads us to the following:

Definition 4.3.1. *A σ-algebra on a set X is a set $\mathcal{F} \subset \mathcal{P}^X$ such that*

(i) $X \in \mathcal{F}$;

(ii) if $A \in \mathcal{F}$, *then* $A^c \in \mathcal{F}$;

(iii) if $A_1, A_2, \cdots \in \mathcal{F}$, *then* $A_1 \cup A_2 \cup \cdots \in \mathcal{F}$.

A σ-algebra is the collection of sets to which we will assign measure when we construct a measure space. Note that we showed that the set of Lebesgue measurable sets satisfies these properties. In other words, we arrive at the following:

Proposition 4.3.2. *Let $\mathcal{L} \subset \mathcal{P}^{\mathbb{R}}$ denote the set of Lebesgue measurable sets. Then $\mathcal{L}$ is a σ-algebra.*

Exercise 4.3.1. *Show that $\emptyset \in \mathcal{F}$.*

Example 4.3.3. *Here are two further (simple) σ-algebras. Fix any set X.*

(i) Let $\mathcal{F} = \{\emptyset, X\}$. Then $\mathcal{F}$ is a σ-algebra.

(ii) The power set $\mathcal{P}^X$ is a σ-algebra.

Given a collection of sets, one can generate the "smallest" σ-algebra containing it. This is similar to the process we used to define the interior, the closure, and the convex/affine/conic hulls of sets. Let $P \subset \mathcal{P}^X$, then

$$\sigma(P) := \bigcap_{\Sigma \supset P, \Sigma \text{ is a } \sigma\text{-algebra}} \Sigma. \tag{4.3.1}$$

One should check that $\sigma(P)$ is, in fact, a σ-algebra:

Exercise 4.3.2. *Given any set X and a collection $P \subset \mathcal{P}^X$, the set $\sigma(P)$ is a σ-algebra. More generally, any arbitrary intersection of σ-algebras is a σ-algebra.*

In general, taking the σ-algebra $\sigma(P)$ generated by a collection of sets P is a pretty messy process. Indeed, the first example of this that we have seen is the Borel σ-algebra (see Example 4.2.11.(iv)):

Definition 4.3.4. *The Borel σ-algebra is the one generated by open subsets of $\mathbb{R}$:*

$$\mathcal{B} = \sigma\left(\{U \subset \mathbb{R} : U \text{ is open}\}\right).$$

We say a set S is Borel measurable *if $S \in \mathcal{B}$. Often we simply say that S is a Borel set.*

Let us note that the definition of the Borel σ-algebra makes sense in *any* metric space. This is one reason that is quite useful in applications.

We do not have an explicit characterization of Borel σ-algebra – given a set, one must "check by hand" if it is Borel. Unfortunately, this is typical of generated σ-algebras; however it is usually not an issue. In practice, it is reassuring that such a set exists (as we shall see, this is useful in probability theory), and often one does not really need to know explicitly what $\sigma(P)$ looks like. One thing that we can say is that Borel sets are Lebesgue measurable:

Lemma 4.3.5. *Let $\mathcal{B}$ denote the Borel σ-algebra and $\mathcal{L}$ denote the σ-algebra of Lebesgue measurable sets. Then $\mathcal{B} \subset \mathcal{L}$.*

Proof. Let $\mathcal{T} = \{U \subset \mathbb{R} : U \text{ is open}\}$. Then $\mathcal{T} \subset \mathcal{L}$ by Example 4.2.11. It follows that $\mathcal{B} \subset \mathcal{L}$ because $\mathcal{L}$ is one of the σ-algebras in the intersection (4.3.1) defining $\mathcal{B}$. □

We present one "nice" case here where we can characterize $\sigma(P)$.

Example 4.3.6. *Suppose $P = \{A_1, \ldots, A_n\}$ is a* partition *of X (that is, $A_i \cap A_j = \emptyset$ if $i \neq j$ and $X = A_1 \cup \cdots \cup A_n$). Then $\sigma(P)$ must contain*

$$\emptyset, X, A_1, A_2, \ldots, A_n. \tag{4.3.2}$$

It must also contain any finite union of the A_i; that is, any set of the form

$$\bigcup_{i \in S} A_i, \tag{4.3.3}$$

where $S \in \mathcal{P}^{\{1,\ldots,n\}}$. Actually, correctly interpreted, the unions in (4.3.3) *represents all the elements in* (4.3.2) *as well.*

One can check that this is the entirety of $\sigma(P)$. Indeed, the disjointness of the A_i implies that taking complements does not add any new elements because, for any $S \in \mathcal{P}^{\{1,\ldots,n\}}$,

$$\left(\bigcup_{i \in S} A_i\right)^c = \bigcup_{i \in S^c} A_i. \tag{4.3.4}$$

Hence,

$$\sigma(P) = \left\{\bigcup_{i \in S} A_i : S \in \mathcal{P}^{\{1,\ldots,n\}}\right\}.$$

Exercise 4.3.3. *Prove* (4.3.4) *and deduce that there is a bijection between* $\sigma(P)$ *and* $\mathcal{P}^{\{1,\dots,n\}}$.

We can then define measurable spaces and measure spaces:

Definition 4.3.7. *A* measurable space *is a pair* $(X, \mathcal{F})$*, where* X *is a set and* $\mathcal{F}$ *is a* σ*-algebra on* X*.*

A measure space *is a triple* $(X, \mathcal{F}, \mu)$*, where* $(X, \mathcal{F})$ *is a measurable space and* μ *is a measure; that is,* $\mu : \mathcal{F} \to \mathbb{R}_+$ *satisfies*

$$\mu(\emptyset) = 0, \qquad \text{and} \qquad \mu\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mu(A_i) \quad \text{if } A_1, A_2, \dots \in \mathcal{F} \text{ are disjoint.}$$

It may seem strange to the reader to have the intermediate step of "measurable spaces." This is because it is often important to have two (or more) measures on the same measurable space. In fact, it is useful to work in metric spaces of measures defined on the same measurable space.

An important special case is that of a probability space.

Definition 4.3.8. *A probability space* $(X, \mathcal{F}, \mathbb{P})$ *is a measure space where*

$$\mathbb{P}(X) = 1.$$

This simply formalizes the idea that the probability that one of all possible events happening is one.

Notice that the measures in Example 4.1.1.(iii)-(v) are all probability measures. By the language we used, this is "obvious" for (iv) and (v), which are the coin flipping probabilities. But it is also true of (iii), the delta measure.

Exercise 4.3.4. *Choose a specific example of a set* X *and a point* x_0 *in Example 4.1.1.(iii), and give a heuristic explanation of the probabilistic meaning of* δ_{x_0}*.*

4.4. Measurable functions and the definition of the integral.

Remark 4.4.1. *It is often convenient in what follows to allow functions to take infinite values (as we did with convex optimization). Some references call this the "extended real numbers" and use a special notation for it. Here, we abuse notation and simply use* $\mathbb{R}$ *with the understanding that* $\pm\infty$ *are legal values for functions to take.*

Let us return back to the land of the Lebesgue measure and begin to build up the basic ideas for integration. First, let us look at simple functions

$$\mathbb{1}_S(x) = \begin{cases} 1 & \text{if } x \in S, \\ 0 & \text{if } x \notin S. \end{cases}$$

If there is any sense to the Lebesgue integral we should have that

$$\int \mathbb{1}_S(x) dx = \int_S 1 dx = m(S). \tag{4.4.1}$$

This is obvious when $S = [a, b]$. The above procedure, however, makes sense if and only if S is measurable. Indeed, if S is not measurable, $m(S)$ is not defined. Hence, we should think of $\mathbb{1}_S$ as

not being a "measurable function." How do we make sense of this for general $f : \mathbb{R} \to \mathbb{R}$? Let us notice something: for any $\alpha > 0$,

$$f^{-1}((\alpha, \infty)) = \{x \in \mathbb{R} : f(x) > \alpha\} = \begin{cases} \emptyset & \text{if } \alpha \geq 1, \\ S & \text{if } \alpha \in [0, 1), \\ \mathbb{R} & \text{if } \alpha < 0. \end{cases}$$

When S is measurable, this set is measurable for every α. When S is not measurable, this set is not measurable for some α. This motivates the following definition:

Definition 4.4.2.

- *A function $f : \mathbb{R} \to \mathbb{R}$ is Lebesgue (resp. Borel)* measurable *if, for every $\alpha \in \mathbb{R}$,*

  $$f^{-1}((\alpha, \infty)) = \{x \in \mathbb{R} : f(x) > \alpha\}$$

  is a Lebesgue (resp. Borel) measurable set.

- *In general, if we have a function*

  $$f : (\Omega, \mathcal{F}) \to (\Sigma, \mathcal{G})$$

  where $(\Omega, \mathcal{F})$ and $(\Sigma, \mathcal{G})$ are measurable spaces, then f is measurable if $f^{-1}(G) \in \mathcal{F}$ for every $G \in \mathcal{G}$.

Remark 4.4.3 (Terminology convention)**.** *Given a function $f : \mathbb{R} \to \mathbb{R}$, when we say that f is measurable without clarifying the σ-algebras involved, we always mean that it is Lebesgue measurable. In other words, we mean that*

$$f : (\mathbb{R}, \mathcal{L}) \to (\mathbb{R}, \mathcal{B})$$

is measurable.

Given an arbitrary measure space $(\Omega, \mathcal{F}, \mu)$, when we say a function $f : \Omega \to \mathbb{R}$ is measurable without clarifying the σ-algebra of the codomain, we always mean *that it is measurable as a function $f : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B})$. In other words, we always mean that it satisfies:*

$$f^{-1}((\alpha, \infty)) \in \mathcal{F} \qquad \text{for all } \alpha \in \mathbb{R}.$$

The main motivation for this is in how we develop integration theory below.

Example 4.4.4. *From the work above, we see that $\mathbb{1}_A$ is measurable if and only if A is measurable.*

Exercise 4.4.1. *Show that f is measurable if and only if any one of the following hold:*

(i) $f^{-1}([\alpha, \infty))$ is measurable for all α;

(ii) $f^{-1}((-\infty, \alpha))$ is measurable for all α;

(iii) $f^{-1}((-\infty, \alpha])$ is measurable for all α.

Example 4.4.5 (Popcorn function)**.** *Define*

$$f : \mathbb{R} \to \mathbb{R} \qquad \text{by} \quad f(x) = \begin{cases} \frac{1}{q} & \text{if } x = \frac{p}{q},\ p \in \mathbb{Z}, q \in \mathbb{N} \text{ (lowest terms)}, \\ 0 & \text{if } x \notin \mathbb{Q}. \end{cases}$$

Its name comes from the fact f has different "jumps" up at each rational number, like popcorn popping (and therefore jumping in the air).

Exercise 4.4.2. *Show that f is continuous at each irrational point and discontinuous at each rational point.*

If $\alpha \geq 0$, then $f^{-1}((\alpha, \infty)) \subset \mathbb{Q}$ has measure zero and is thus Lebesgue measurable. If $\alpha < 0$, then $f^{-1}((\alpha, \infty)) = \mathbb{R}$ is measurable.

Exercise 4.4.3. *If $f : \mathbb{R} \to \mathbb{R}$ is measurable, then $f^{-1}(U)$ is measurable for any open set U. Similarly, $f^{-1}(K)$ is measurable for any closed set K.*

A more general version of Exercise 4.4.3 is the following, which will be useful in the sequel.

Lemma 4.4.6. *Fix a measure space $(\Omega, \mathcal{F}, \mu)$. Suppose that $f : \Omega \to \mathbb{R}$ is measurable. Then $f^{-1}(B)$ is measurable for every Borel set B.*

Proof. Let $\mathcal{G} = \{E \subset \mathbb{R} : f^{-1}(E)\}$ is measurable[24]. We claim that $\mathcal{G}$ is a σ-algebra. Indeed, $f^{-1}(\mathbb{R}) = \Omega$, which is measurable, so $\mathbb{R} \in \mathcal{G}$. Additionally, if $E \in \mathcal{G}$, then

$$f^{-1}(E)^c = f^{-1}(E^c).$$

Due to the assumed measurability of $f^{-1}(E)$, the left hand side above must be measurable. We conclude that $E^c \in \mathcal{G}$. Finally, if $E_1, E_2, \cdots \in \mathcal{G}$, then

$$f^{-1}\Big(\bigcup_{i=1}^{\infty} E_i\Big) = \bigcup_{i=1}^{\infty} f^{-1}(E_i).$$

Again, due to the assumed measurability of $f^{-1}(E_i)$ for each i, the right hand side above must be measurable. We conclude that

$$\bigcup_{i=1}^{\infty} E_i \in \mathcal{G}.$$

On the other hand, clearly $\mathcal{G}$ contains sets of the form (α, ∞). It follows that $\mathcal{G}$ must contain the σ-algebra that these sets generate:

$$\mathcal{G} \supset \sigma\left(\{(\alpha, \infty) : \alpha \in \mathbb{R}\}\right) = \mathcal{B}.$$

It follows that $f^{-1}(B)$ is measurable for every $B \in \mathcal{B}$, which completes the proof. □

Proposition 4.4.7. *(i) If $f : \mathbb{R} \to \mathbb{R}$ is continuous, then it is Borel measurable.*

(ii) If $f : \mathbb{R} \to \mathbb{R}$ is Borel measurable, then it is Lebesgue measurable.

(iii) If $f : \mathbb{R} \to \mathbb{R}$ is Lebesgue measurable and $g : \mathbb{R} \to \mathbb{R}$ is Borel measurable, then $g \circ f$ is Lebesgue measurable.

(iv) If $c \in \mathbb{R}$, $f, g : \mathbb{R} \to \mathbb{R}$ are measurable, then so are cf, $f + g$, and fg.

Exercise 4.4.4. *A function $f : \mathbb{R} \to \mathbb{R}$ is continuous if and only if $f^{-1}(U)$ is open for every open set U.*

[24] In this lemma, we use measurable as shorthand for either Lebesgue or Borel measurable, depending on the assumptions.

Proof of Proposition 4.4.7.(i). This follows immediately from the exercise above and the measurability of open sets. □

Proof of Proposition 4.4.7.(ii). This is immediate from the fact that any Borel measurable set is Lebesgue measurable (see Lemma 4.3.5). □

"Proof" of Proposition 4.4.7.(iii). Since set operations commute with inverse images under functions, it is easy to check that $g^{-1}((\alpha,\infty))$ is Borel and that $f^{-1}(B)$ is Lebesgue measurable if B is Borel. The result follows from this. □

Proof of Proposition 4.4.7.(iv). We show the second is measurable and leave the first and third as an exercise. Indeed: let $\mathbb{Q} = \{q_1, q_2, \dots\}$, and notice that, for any $\alpha > 0$,

$$\begin{aligned}(f+g)^{-1}((\alpha,\infty)) &= \bigcup_{n=1}^{\infty} \{x : f(x) > \alpha + q_n \text{ and } g(x) > -q_n\} \\ &= \bigcup_{n=1}^{\infty} \left(f^{-1}((\alpha+q_n,\infty)) \cap g^{-1}((-q_n,\infty))\right)\end{aligned}$$

is measurable due to Proposition 4.2.10. □

Another operation that works well with measurability is limits:

Proposition 4.4.8. *Suppose that $f_1, f_2, \dots$ is a sequence of measurable functions. Then the functions $\liminf_n f_n$, $\limsup_n f_n$, $\inf_n f_n$, and $\sup_n f_n$ are all measurable.*

Exercise 4.4.5. *Prove this!*

4.4.1. **Random variables and the beginnings of integration.** An important "example" of a measurable function is a random variable X on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Let us consider some more explicit examples.

Example 4.4.9. *(Coin flipping) – Recall Example 4.1.1.(v), in which we flip a biased coin n times that takes the value tails with probability p and heads with probability $1-p$. Recall that our measure space is*

$$\Omega = \{0,1\}^n \qquad \text{and} \qquad \mathcal{F} = \mathcal{P}^{\{0,1\}^n},$$

and with $\mathcal{P}$ is defined by (4.1.5)*. Since this is a "discrete" space, we can assign measure to every set without an issues.*

Suppose we are gambling and receive \$2 *for every tails, and lose* \$3 *for every heads. We define a random variable X that represents our winnings:*

$$\begin{aligned}X(f_1, f_2, \dots, f_n) &= 2\#\{i : f_i = 0\} - 3\#\{i : f_i = 1\} = 2\#\{i : f_i = 0\} - 3(n - \#\{i : f_i = 0\}) \\ &= 5\#\{i : f_i = 0\} - 3n.\end{aligned}$$

Since all sets are measurable, all functions are also measurable. Hence, X is measurable.

How should we define the integral? In this case, we can simply decompose X into a collection of indicator functions:

$$X(\omega) = 5\sum_{k=0}^{n} k\mathbb{1}_{A_k}(\omega) - 3n.$$

Then, we have that the expected value *is*

$$\mathbb{E}[X] = \int_\Omega X(\omega) d\mathbb{P} = 5 \sum_{k=0}^{n} k\mathbb{P}(A_k) - 3n.$$

Using (4.1.5)*, we know that the probability of any string of heads and tails is* 2^{-n}*. Hence,*

$$\mathbb{P}(A_k) = 2^{-n} \# A_k.$$

Combinatorially, it is easy to check that this is $\binom{n}{k}$ *(there are* k *entries that you choose to put a* 0 *in). Hence*

$$\mathbb{E}[X] = 5 \cdot 2^{-n} \sum_{k=0}^{n} k \binom{n}{k} - 3n.$$

Using that

$$n(1+x)^{n-1} = \frac{d}{dx}(1+x)^n = \sum_{k=0}^{n} k \binom{n}{k} x^{k-1},$$

we get

$$\mathbb{E}[X] = \frac{5n}{2} - 3n = -\frac{n}{2}.$$

This is not a good outlook for our wallet...

4.4.2. **Integration theory.** Notice that, above, we computed the integral for the indicator function of any measurable set (recall (4.4.1)). Since we expect integrals to be linear, we can easily extend this definition to the linear combination of any indicator functions (see the notion of "simple function" in Definition 4.4.10), and, by approximation, to all measurable functions. Let us begin by setting the notion of simple function in stone.

Definition 4.4.10. *Let* $(\Omega, \mathcal{F}, \mu)$ *be a measure space. A* simple function *is a function* $f : \Omega \to \mathbb{R}$ *such that*

$$f = \sum_{i=1}^{n} c_i \mathbb{1}_{A_i}$$

for $n \in \mathbb{N}$*,* $c_1, \ldots, c_n \in \mathbb{R}$*, and* $A_1, \ldots, A_n \in \mathcal{F}$ *(i.e., are measurable) and satisfy* $\mu(A_i) < \infty$ *for all* $i = 1, 2, \ldots, n$*.*

We note that f is measurable because it is the linear combination of measure functions $\mathbb{1}_{A_i}$. In general, we only consider measurable simple functions because of their important role in defining the Lebesgue integral.

Remark 4.4.11. *(i) If* f *and* g *is simple and* $\alpha, \beta \in \mathbb{R}$*, then* $\alpha f + \beta g$ *is simple as well. Roughly, any "reasonable" modification of a simple function leads to another simple function (cf. Exercise 4.4.6).*

(ii) In the definition above, there is no uniqueness in the choice of c_i *and* A_i*. For example,*

$$\mathbb{1}_{[0,1]} = \mathbb{1}_{[0,1/2)} + \frac{1}{2}\mathbb{1}_{[1/2,3/4]} + \frac{1}{2}\mathbb{1}_{[1/2,1]} + \frac{1}{2}\mathbb{1}_{(3/4,1]}.$$

However, we can choose a canonical representative in the following way. If f *is simple, then the set* $f(\mathbb{R}) \setminus \{0\}$ *is finite. Let* n *be the number of elements and label them* $c_1, \ldots, c_n$*. Let*

$$A_i = \{x : f(x) = c_i\} \qquad \text{for each } i = 1, \ldots, n.$$

Then

$$f = \sum_{i=1}^{n} c_i \mathbb{1}_{A_i}$$

where the sets $A_1, \ldots, A_n$ are mutually disjoint and $c_i \neq c_j$ for all $i \neq j$. This choice of representative is unique, and it often makes computations easier.

Exercise 4.4.6. *If f is a simple function, so are $|f|$, $f_+ = \max\{0, f\}$, and $f_- = \max\{0, -f\}$.*

For such functions, the Lebesgue integral is easy to compute:

$$\int f \, dm = \sum_{i=1}^{n} c_i \int \mathbb{1}_{A_i} \, dm = \sum_{i=1}^{n} c_i m(A_i). \tag{4.4.2}$$

While we are "okay" with the integral being infinite, we start to get uneasy if we have $+\infty - \infty$ showing up in (4.4.2). This is our motivation for restricting to sets A_i with finite measure in Definition 4.4.10. In general, we usually work with absolutely integrable functions f such that

$$\int |f| \, dm < \infty$$

or, at the very least, functions such that f_- has a finite integral. Keep this in mind for when we define the general Lebesgue integral.

Another important thing to notice is that we change f on a set of measure zero, the integral is unaffected. To make this more explicit, let us consider the following: let

$$\tilde{f}(x) = f(x) + \mathbb{1}_{\mathbb{Q}}.$$

Clearly $\tilde{f}$ is a simple function, $\tilde{f} \neq f$, and

$$\int |\tilde{f}(x) - f(x)| \, dm = \int \mathbb{1}_{\mathbb{Q}} \, dm = 0. \tag{4.4.3}$$

The last equality shows us that, as far as the natural metric associated to the Lebesgue measure is concerned, $\tilde{f} = f$. So what gives!? As we discuss more later, we start to think of functions as being "the same" if they only differ on a set of measure zero.

A useful (and important!) property of the Lebesgue integral is its monotonicity.

Exercise 4.4.7. *If f and g are simple functions and $f \leq g$, then*

$$\int f \, dm \leq \int g \, dm. \tag{4.4.4}$$

What about if $f \leq g$ on a set $A \subset \mathbb{R}$ such that $m(\mathbb{R} \setminus A) = 0$? Does (4.4.4) still hold?

Definition 4.4.12. *If $f : \mathbb{R} \to \mathbb{R}_+$ is a measurable function, then we define*

$$\int f \, dm = \sup_{\substack{s \leq f, \\ s \text{ is a simple function}}} \int s \, dm.$$

If $f : \mathbb{R} \to \mathbb{R}$ is a measurable function and either

$$\int f_+ \, dm < \infty \qquad \text{or} \qquad \int f_- \, dm < \infty,$$

then we define

$$\int f\,dm = \int f_+\,dm - \int f_-\,dm.$$

In general, we say that $f \in L^1$ ("f is integrable") if

$$\int |f|\,dm = \int f_+\,dm + \int f_-\,dm < \infty.$$

Sometimes we write this as $L^1(dm)$ or $L^1(\mathbb{R})$ depending on whether we want to emphasize the measure or the domain of integration.

For the definition above to make sense, we need the following exercise:

Exercise 4.4.8. *If f is measurable, so are $|f|$, f_+, and f_-.*

Remark 4.4.13. *(i) One should always worry about consistency. If we apply Definition 4.4.12 to a simple function f, do we get the value given by Definition 4.4.10? Yes! This relies on the monotonicity of the Lebesgue integral* (4.4.4).

Exercise 4.4.9. *Convince yourself of this!*

(ii) Suppose that f is Riemann integrable on a set $[a,b]$. In this case, we see that the Lebesgue integral yields the same value as the Riemann integral as follows. Fix $N > 0$. Let

$$A_{n,N} = \left[a + n\frac{b-a}{N}, a + (n+1)\frac{b-a}{N}\right], \quad \ell_{n,N} = \inf_{x \in A_{n,N}} f(x) \quad \text{and} \quad u_{n,N} = \sup_{x \in A_{n,N}} f(x).$$

Using Definition 4.4.12 and the monotonicity of the Lebesgue integral on simple functions, one can check that

$$\sum_{i=1}^{N} \ell_{n,N} m(A_{n,N}) \leq \int \mathbb{1}_{[a,b]} f\,dm \leq \sum_{i=1}^{N} u_{n,N} m(A_{n,N}) \tag{4.4.5}$$

where the integral in the middle is the Lebesgue integral. But the terms on the left and right hand sides are the lower and upper Riemann sums, respectively. By Riemann integrability, they must converge to the Riemann integral. It follows that the Riemann and Lebesgue integrals are equal.

Exercise 4.4.10. *Work out the details! Notice that we did not show that f is measurable... show this!*

(iii) **Exercise 4.4.11.** *The monotonicity* (4.4.4) *extends to all Lebesgue measurable functions.*

Example 4.4.14. *The function $\mathbb{1}_{\mathbb{Q}}$ is not Riemann integrable. Indeed, using the notational of Remark 4.4.13.(ii), $u_{n,N} = 1$ and $\ell_{n,N} = 0$ for every n and N, so the left and right hand sides of* (4.4.5) *are, respectively, 0 and ∞ for every N. This disagreement between the upper and lower Riemann sums is exactly what it means to be not Riemann integrable.*

On the other hand, we see that

$$\int \mathbb{1}_{\mathbb{Q}}\,dm = m(\mathbb{Q}) = 0.$$

Remark 4.4.15. *We say that $f = g$* almost everywhere *(a.e.), or, equivalently, up to a set of measure zero, if $m(\{x : f(x) \neq g(x)\}) = 0$. As we first pointed out in* (4.4.3)*, we take the perspective that two functions are the same if they are equal a.e. In this way, we we can make sense of the L^p-norms being actual norms (otherwise, positive-definiteness would not be satisfied).*

The flip side to making this equivalence is that we can no longer assign a value to $f(x)$ for any x. This may seem like a strange thing to say. But it is unavoidable: if you fix you favorite function, which is $f = \pi^{-|x|}$, and tell me that $f(51) = \pi^{-51}$, I might tell you that you are wrong because my favorite equivalent *function to f is $g = f + \mathbb{1}_{\{57\}}$, which yields $g(51) = \pi^{-51} + 1$. Who is wrong here? Neither of us!*

We do not need to sink into a nihilistic despair, though, because we can still use the L^p-norms to distinguish functions. Indeed, in the example above,

$$\int f\,dm = \int g\,dm \qquad \text{and} \qquad \int |f - g|\,dm = 0,$$

but, for any measurable function h that is not equal to f a.e.,

$$\int |f - h|\,dm > 0.$$

More on this later.

Exercise 4.4.12. *Fix a complete*[25] *measure space $(\Omega, \mathcal{F}, \mu)$. Show that if $f = g$ μ-a.e. and f is measurable, then g is measurable as well. Is this true if $(\Omega, \mathcal{F}, \mu)$ is not complete?*

Lastly, let us recall how important the linearity of the integral is. Indeed, all of our work in L^p-spaces in Section 1.5 is built off of this property. We immediately get that the Lebesgue integral has this property as well. Indeed, it is immediate from the definition that if s_1, s_2 are simple and $a_1, a_2 \in \mathbb{R}$, then

$$\int (a_1 s_1 + a_2 s_2)\,dm = a_1 \int s_1\,dm + a_2 \int s_2\,dm. \tag{4.4.6}$$

This can be bootstrapped to general measurable functions with a bit of work. Let us be more precise below.

Proposition 4.4.16. *Fix any $a, b \in \mathbb{R}$ and suppose that f, g are measurable functions such that*

$$\int (af)_+\,dm, \int (bg)_+ dm < \infty \qquad \text{or} \qquad \int (af)_-\,dm, \int (bg)_- dm < \infty.$$

Then

$$\int (af + bg)\,dm = a \int f\,dm + b \int g\,dm.$$

Here we use the convention that $0 \cdot \infty = 0$ if necessary.

We prove this only for bounded functions whose support has finite measure. Once we learn the convergence theorems in Section 4.5.2, we can easily extend this to general measurable functions. To prove (our bounded version of) linearity, we need the following lemma:

[25] See Definition 4.4.18.

Lemma 4.4.17. *Suppose that* $m(\operatorname{supp} f) < \infty$ *and* $\sup |f| < \infty$. *Then*

$$\int f\,dm = \inf_{\substack{s \geq f, \\ s \text{ is a simple function}}} \int s\,dm.$$

Here $\operatorname{supp} f = \{x \in \mathbb{R} : f(x) \neq 0\}$.

Proof. Let us point out that the "$\leq$" direction follows from Exercise 4.4.11. We, thus, seek only to prove that

$$\int f\,dm \geq \inf_{\substack{s \geq f, \\ s \text{ is simple}}} \int s\,dm. \tag{4.4.7}$$

We prove this for $f \geq 0$. However, by writing $f = f^+ - f_-$, this is easily extended to the general case. Let $M = \sup f$ and fix any $n \in \mathbb{N}$. Let, for any $k = 1, \ldots, n$,

$$\Lambda_{k,n} = \left\{ x : M\frac{k-1}{n} < f(x) \leq M\frac{k}{n} \right\}.$$

We define the simple functions

$$\underline{s}(x) = \sum_{k=1}^{n} \frac{M(k-1)}{n} \mathbb{1}_{\Lambda_{k,n}} \qquad \text{and} \qquad \overline{s}(x) = \sum_{k=1}^{n} \frac{Mk}{n} \mathbb{1}_{\Lambda_{k,n}}.$$

It follows that $\underline{s} \leq f \leq \overline{s}$, so that, by Exercise 4.4.11,

$$\int \underline{s}\,dm \leq \int f\,dm \leq \int \overline{s}\,dm.$$

On the other hand, we have that, by (4.4.6),

$$\begin{aligned} \int \overline{s}\,dm - \int \underline{s}\,dm = \int (\overline{s} - \underline{s})\,dm &= \int \sum_{k=1}^{n} \frac{M}{n} \mathbb{1}_{\Lambda_{k,n}}\,dm \\ &= \frac{M}{n} \sum_{k=1}^{n} m(\Lambda_{k,n}) \leq \frac{M}{n} m(\operatorname{supp} f). \end{aligned}$$

We deduce that

$$\inf_{\substack{s \geq f, \\ s \text{ is simple}}} \int s\,dm - \frac{M}{n} m(\operatorname{supp} f) \leq \int \overline{s}\,dm - \frac{M}{n} m(\operatorname{supp} f) \leq \int \underline{s}\,dm \leq \int f\,dm.$$

Taking $n \to \infty$, we arrive at (4.4.7), which completes the proof. □

Proof of Proposition 4.4.16. It is clearly enough to prove this simply for $f, g \geq 0$ since the linearity of the positive and negative parts is built into Definition 4.4.12. We make the additional simplifying assumption that $a, b > 0$. When a or b is zero, the proof is obvious. If either of a or b is negative, the proof follows similarly. Finally, we assume f and g satisfy the additional hypotheses of Lemma 4.4.17 (see the discussion above Lemma 4.4.17 about how to remove this extra hypothesis).

Fix any $\varepsilon > 0$, and let s_f and s_g be nonnegative simple functions such that

$$s_f \leq f, \qquad s_g \leq g, \qquad \int f\,dm \leq \int s_f\,dm + \varepsilon \quad \text{and} \quad \int g\,dm \leq \int s_g\,dm + \varepsilon. \tag{4.4.8}$$

Then, we have that as_f and bs_g are simple, $as_f \le af$, and $bs_g \le bg$. Hence,

$$as_f + bs_g \le af + bg,$$

so that

$$\int (as_f + bs_g)\, dm \le \int (af + bg)\, dm.$$

Using then (4.4.8)

$$a\int f\, dm + b\int g\, dm - (a+b)\varepsilon \le \int (as_f + bs_g)\, dm \le \int (af + bg)\, dm.$$

Since ε is arbitrary, we deduce that

$$a\int f\, dm + b\int g\, dm \le \int (af + bg)\, dm.$$

To show the opposite inequality, fix $\varepsilon > 0$ again, and let s_f and s_g be nonnegative simple functions such that

$$s_f \ge f, \qquad s_g \ge g, \qquad \int f\, dm \ge \int s_f\, dm - \varepsilon \quad \text{and} \quad \int g\, dm \le \int s_g\, dm - \varepsilon. \tag{4.4.9}$$

This is possible by Lemma 4.4.17. Then, we have that as_f and bs_g are simple, $as_f \ge af$, and $bs_g \ge bg$. Hence, $as_f + bs_g \ge af + bg$, and, applying Lemma 4.4.17 again,

$$\int (as_f + bs_g)\, dm \ge \int (af + bg)\, dm.$$

Using then (4.4.9),

$$\begin{aligned} a\int f\, dm + b\int g\, dm + (a+b)\varepsilon &\ge a\int s_f\, dm + b\int s_g\, dm \\ &= \int (as_f + bs_g)\, dm \ge \int (af + bg)\, dm. \end{aligned}$$

Since ε is arbitrary, we deduce that

$$a\int f\, dm + b\int g\, dm \ge \int (af + bg)\, dm,$$

which completes the proof. □

Exercise 4.4.13. *Show that the definition of integral developed above works in general. More precisely, if $(\Omega, \mathcal{F}, \mu)$ is a measure space, then we may define*

$$\int f\, d\mu = \sup_{\substack{s \ge f, \\ s \text{ is a simple function}}} \int s\, d\mu$$

for any measurable $f : \Omega \to \mathbb{R}$. Here, we define

$$\int s\, d\mu = \sum_{i=1}^{N} c_i m(A_i) \qquad \text{if } s(x) = \sum_{i=1}^{N} c_i \mathbb{1}_{A_i}(x).$$

Show that all of the above results extend to this case.

Exercise 4.4.14. *Show that if $f = g$ a.e., then*

$$\int f dm = \int g dm.$$

Is the converse true?

4.4.3. **Arbitrary measure spaces, new measures, and a technical note about complete measure spaces.** We can define the integral associated to any arbitrary measure space $(\Omega, \mathcal{F}, \mu)$ exactly as in Definition 4.4.12. All theorems proved above hold in the general setting.

Actually, integration gives us a method by which to create new measures. Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. Let f be a nonnegative measurable function. Then $(\Omega, \mathcal{F}, \mu_f)$, where

$$\mu_f(A) = \int_A f \, d\mu = \int \mathbb{1}_A f \, d\mu.$$

for all $A \in \mathcal{F}$, is a measure space. In this case, we refer to f as the "density" of μ_f. Sometimes we write this as either

$$d\mu_f = f \, d\mu \quad \text{or} \quad \frac{d\mu_f}{d\mu} = f.$$

The existence of densities is the subject of the Radon-Nikodym theorem, which is outside the scope of these notes. We shall see the importance of densities in Section 5, when we study probability theory.

In the future, when we introduce measurable functions and develop some of the theory of L^p-spaces, it is less awkward to work with "nicer" measure spaces. This is not totally essential to the material presented here, so the reader may choose to simply ignore this technical detail.

Definition 4.4.18. *A measure space $(\Omega, \mathcal{F}, \mu)$ is* complete *if, for all $A \in \mathcal{F}$ (i.e. A is measurable), if $B \subset A$ and $\mu(A) = 0$, then $B \in \mathcal{F}$ (i.e., B is measurable).*

We see immediately that the Lebesgue measure space is complete because of how it is defined via outer measure (see Example 4.2.7.((i))). This is not true of the Borel sigma algebra paired with the Lebesgue measure.

Why might one want to do this? Rudin[26] points out the following motivation related to equivalence classes of functions. Recall that $f \sim g$ if $f = g$ almost everywhere. Suppose that we want to redefine f on a set E of measure zero in a complicated way. Call $\tilde{f}$ the resulting function. Then $f = \tilde{f}$ on E^c and, hence, almost everywhere. We would think of f and $\tilde{f}$ as the "same" function and, of course, the integral of f is not changed at all. However, if $B \subset E$ is a non-measurable set, and we accidentally define $\tilde{f}$ to take one value on B and a different value on $E \setminus B$, then $\tilde{f}$ will no longer be measurable. This is quite annoying! And, actually, this scenario does appear when one talks about "weak derivatives," which are central to many branches of modern analysis.

Luckily, one can "complete" any measure space in a unique way. This completed space will have the same integration properties of the original measure space. There is, thus, no loss of generality in working with complete measure spaces.

We always work with complete measure spaces from this point on. Actually, with some effort any measure space can be "completed," so this is not really a restriction.

Exercise 4.4.15. *Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. Then there is a unique complete measure space $(\Omega, \bar{\mathcal{F}}, \bar{\mu})$ such that*

(i) if $A \in \mathcal{F}$, then $A \in \bar{\mathcal{F}}$ and $\mu(A) = \bar{\mu}(A)$;

(ii) if $\bar{A} \in \bar{\mathcal{F}}$ then there is $A \in \mathcal{F}$ such that $\bar{\mu}(\bar{A} \Delta A) = 0$. Recall Δ from Exercise 4.2.3.

(Hint: develop an outer-measure $\bar{\mu}^*$ using μ.)

Exercise 4.4.16. *Show that the Lebesgue measure space $(\mathbb{R}, \mathcal{L}, m)$ is the completion of $(\mathbb{R}, \mathcal{B}, m)$.*

[26] "Real and complex analysis" 1987.

4.5. **Limits and convergence.** The real power of the Lebesgue measure comes from how well it works with limits. As we will see, the L^p-spaces are complete with the L^p-norm, and this seemingly modest fact plays an enormous role in the development of modern analysis. Completeness is a statement about how limits, so let us dive into understanding how limits work with measures.

4.5.1. **Sets and measures.** Given an increasing sequence of measurable sets $A_1 \subset A_2 \subset \cdots$, there is an obvious "limit" set:

$$\mathcal{A} = \bigcup_{i=1}^{\infty} A_i.$$

What happens with the measure?

Example 4.5.1. *(i) If $A_n = (0, 1 - 1/n)$, then A_i intuitively "approaches" $(0, 1)$, and we see that*

$$\mathcal{A} = \bigcup_{i=1}^{\infty} A_i = (0, 1).$$

Notice that $m(A_n) = 1 - 1/n \to 1 = m(\mathcal{A})$.

(ii) If $A_n = (0, n)$, then A_i intuitively "approaches" $(0, \infty)$, and we see that

$$\mathcal{A} = \bigcup_{i=1}^{\infty} A_i = (0, \infty).$$

Notice that $m(A_n) = n \to \infty = m(\mathcal{A})$.

(iii) If $A_n = \{q_1, \ldots, q_n\}$, where $\mathbb{Q} = \{q_1, q_2, \ldots\}$ then A_i intuitively "approaches" $\mathbb{Q}$, and we see that

$$\mathcal{A} = \bigcup_{i=1}^{\infty} A_i = \mathbb{Q}.$$

Notice that $m(A_n) = 0 \to 0 = m(\mathcal{A})$.

It seems that measures work very nicely with increasing sequences! Let us prove that:

Lemma 4.5.2. *Let $A_1 \subset A_2 \subset \cdots$ be an increasing sequence of measurable sets, then*

$$\mu\left(\mathcal{A} = \bigcup_{i=1}^{\infty} A_i\right) = \lim_{n\to\infty} \mu(A_n).$$

Proof. First, if $\mu(A_n) = \infty$ for any n, then the proof is trivial. Let us, therefore, assume that $\mu(A_n) < \infty$ for all n. Let use generate a disjoint countable collection that yields the same union:

$$B_1 = A_1, \quad B_2 = A_2 \setminus A_1 \quad \ldots \quad \text{and} \quad B_n = A_n \setminus A_{n-1}.$$

Then we have

$$\mu(A_n) = \sum_{k=1}^{n} \mu(B_k) \quad \text{and} \quad \mu(\mathcal{A}) = \sum_{k=1}^{\infty} \mu(B_k).$$

The conclusion thus follows from basic calculus results. □

Given a decreasing sequence of sets $A_1 \supset A_2 \supset \cdots$, there is an obvious "limit" set:

$$\mathcal{A} = \bigcap_{i=1}^{\infty} A_i.$$

What happens with the measure?

Example 4.5.3. *(i) If $A_n = (0, 1 + 1/n)$, then A_i intuitively "approaches" $(0, 1]$, and we see that*

$$\mathcal{A} = \bigcap_{i=1}^{\infty} A_i = (0, 1].$$

Notice that $m(A_n) = 1 + 1/n \to 1 = m(\mathcal{A})$.

(ii) If $A_n = (n, \infty)$, then A_i intuitively "approaches" $\emptyset$, and we see that

$$\mathcal{A} = \bigcap_{i=1}^{\infty} A_i = \emptyset.$$

Notice that $m(A_n) = \infty$, which does not *converge to $0 = m(\emptyset) = m(\mathcal{A})$.*

(iii) If $A_n = \mathbb{Q} \setminus \{q_1, \ldots, q_n\}$, where $\mathbb{Q} = \{q_1, q_2, \ldots\}$ then A_i intuitively "approaches" $\emptyset$, and we see that

$$\mathcal{A} = \bigcap_{i=1}^{\infty} A_i = \emptyset.$$

Notice that $m(A_n) = 0 \to 0 = m(\mathcal{A})$.

So we see that decreasing sequences are not as nice! Let us conjecture two things, though: (1) it seems that the "limit" works as long as some of the A_i are finite, and (2) it seems that you can only "lose" mass in the limit, not gain it.

Exercise 4.5.1. *Suppose that $A_1 \supset A_2 \supset \cdots$ are measurable sets and $\mu(A_N) < \infty$ for some $N \in \mathbb{N}$. Then*

$$\mu\left(\bigcap_{i=1}^{\infty} A_i\right) = \lim_{n\to\infty} \mu(A_n).$$

Let us test out these guesses with two non-increasing and non-decreasing sequences. Here we may not have a good notion of convergence, but in analogy with sequences, we should still be able to define a "lim inf" and "lim sup." Indeed

Definition 4.5.4. *Suppose that $A_1, A_2, \ldots$ is a countable collection of sets. Then we define*

$$\liminf_n A_n = \bigcup_{i=1}^{\infty} \bigcap_{j=i}^{\infty} A_j \qquad \text{and} \qquad \limsup_n A_n = \bigcap_{i=1}^{\infty} \bigcup_{j=i}^{\infty} A_j.$$

Roughly, $\liminf A_n$ is the set of points that are in all but finitely many of the A_n while $\limsup A_n$ is the set of points that are in A_n infinitely often.

Exercise 4.5.2. *(i) Show that, for ever x,*

$$\mathbb{1}_{\liminf_n A_n}(x) = \liminf_{n\to\infty} \mathbb{1}_{A_n}(x).$$

Similarly for $\limsup_n A_n$.

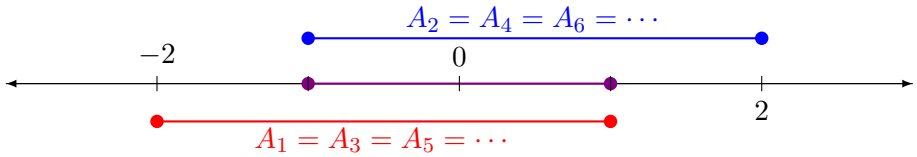
(ii) If A_n is an increasing sequence of sets, show that

$$\bigcup_{n=1}^{\infty} A_n = \liminf_n A_n = \limsup_n A_n.$$

(iii) If A_n is an decreasing sequence of sets, show that

$$\bigcap_{n=1}^{\infty} A_n = \liminf_n A_n = \limsup_n A_n.$$

Example 4.5.5. *(i) Let $A_n = (-1)^n[-1, 2]$.*



Then

$$\liminf_n A_n = \bigcup_{i=1}^{\infty} \bigcap_{j=i}^{\infty} A_j = [-1, 1].$$

Notice that

$$m\left(\liminf_n A_n\right) = 2 \leq 3 = \liminf_n m(A_n).$$

(ii) Let $A_n = [0, 1] \cup [n, 2n]$. Then

$$\liminf_n A_n = \bigcup_{i=1}^{\infty} \bigcap_{j=i}^{\infty} A_j = [0, 1].$$

Notice that

$$m\left(\liminf_n A_n\right) = 1 \leq \infty = \liminf_n m(A_n).$$

Lemma 4.5.6 (Fatou's Lemma)**.** *Suppose that $(\Omega, \mathcal{F}, \mu)$ is a measure space.*

(i) If $A_1, A_2, \ldots$ is a sequence of measurable sets, then

$$\mu\left(\liminf_n A_n\right) \leq \liminf_n \mu(A_n).$$

(ii) If $f_1, f_2, \ldots$ is a sequence of nonnegative measurable functions, then

$$\int \left(\liminf_n f_n\right)\, d\mu \leq \liminf_n \int f\, d\mu.$$

We point out that (i) corresponds to our observations above.

Proof of (i). This follows immediately from part (ii) and the correspondence between sets and their indicator functions.

Exercise 4.5.3. *Write down the steps for this!*

□

Proof of (ii). Let $f = \liminf f_n$ and fix any non-negative simple function

$$s = \sum_{i=1}^{n} c_i \mathbb{1}_{A_i}$$

such that $s \leq f$. Let $\varepsilon \in (0,1)$, and take, for any k,

$$B_k^{i,\varepsilon} = \{x \in A_i : (1-\varepsilon)s(x) = (1-\varepsilon)c_i \leq f_\ell(x) \text{ for all } \ell \geq k\}.$$

Notice that $B_k^{i,\varepsilon}$ is an increasing sequence in k such that

$$\bigcup_{k=1}^{\infty} B_k^{i,\varepsilon} = A_i.$$

Here is where we have used the ε. Thus, by Lemma 4.5.2,

$$\begin{aligned}\int (1-\varepsilon)s(x)\ d\mu &= \sum_{i=1}^{n}(1-\varepsilon)c_i\mu(A_i) = \lim_{k\to\infty}\sum_{i=1}^{n}(1-\varepsilon)c_i\mu(B_k^{i,\varepsilon}) \\ &= \lim_{k\to\infty}\sum_{i=1}^{n}\int_{B_k^{i,\varepsilon}}(1-\varepsilon)c_i\ d\mu \leq \liminf_k \sum_{i=1}^{n}\int_{B_k^{i,\varepsilon}} f_k(x)\ d\mu \\ &\leq \liminf_k \int_\Omega f_k(x)\ d\mu.\end{aligned}$$

Since this holds for all ε, we deduce that

$$\int s(x)\ d\mu \leq \liminf_k \int f_k(x)\ d\mu.$$

The claim then follows by the arbitrariness of s as well as the definition of the integral:

$$\int f\ d\mu = \sup_{\substack{s\leq f, \\ s \text{ is a simple function}}} \int s\ d\mu \leq \liminf_k \int f_k\ d\mu.$$

□

What about the limsup?

Exercise 4.5.4. *Borel-Cantelli Lemma: If A_k is a sequence of sets such that $\mu(A_1)+\mu(A_2)+\cdots < \infty$, then $\mu(\limsup A_n) = 0$. Prove this. Additionally, show that the conclusion is not necessarily true if the summability condition is not satisfied.*

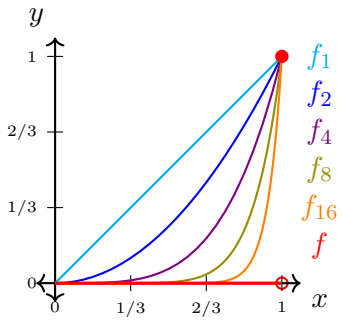
Let us now look at our first (and weakest) notion of convergence:

Definition 4.5.7 (Convergence almost everywhere)**.** *We say that a sequence of measurable functions f_n on a measure space $(\Omega, \mathcal{F}, \mu)$* converges to a function f almost everywhere *if there is a set $A \subset \Omega$ such that $\mu(\Omega \setminus A) = 0$ and, for every $x \in A$,*

$$\lim_{n\to\infty} f_n(x) = f(x).$$

Exercise 4.5.5. *Show that f is measurable in the above setting.*

Example 4.5.8. *Clearly if $f_n \to f$ pointwise (in the sense we learned in calculus class once upon a time), then $f_n \to f$ a.e. Let us look at some more exotic examples in action:*



(i) Let $f_n(x) = x^n$ on the measure space defined by the Lebesgue measure on $[0,1]$. Then $f_n(x) \to 0$ if $x \in [0,1)$ and $f_n(1) = 1$ for all n. Let $f(x) = \mathbb{1}_{\{1\}}$. It follows that $f_n \to f$ a.e. On the other hand, we may also define $\varphi(x) = 0$ for all x. We also see that $f_n \to \varphi$ a.e. Note that $f = \varphi$ a.e., so that does not cause issues with uniqueness of limits in a world where we only think about almost everywhere limits and where we identify two functions that are equal almost everywhere.

(ii) Let $g_n(x) = e^{x^2 - n|x|}$.

Exercise 4.5.6. *Give two examples of functions to which g_n converges almost everywhere.*

Exercise 4.5.7. *Show that if $f_n \to f$ a.e., $g_n \to g$ a.e., and $f_n = g_n$ a.e., then $f = g$ a.e.*

4.5.2. **Convergence of integrals.** In many branches of math, we solve problems by approximation or we ask questions about long-time limits, both of which involve limits. For example, let us say that $p_n(x)$ is the density of "heat" on a $\sqrt{n}$ spatial scale – roughly, it represents the temperature at time n and location $x\sqrt{n}$. Is it true that p_n converges to some other function? If it does, do "bulk" quantities like

$$\int p_n(x)^2 dm \quad \text{and} \quad \int x p_n(x) dm$$

converge? These might be thought of as the energy and the "center of heat" of the system, respectively.

To get our hands on this a little, let us look at two simple examples:

Example 4.5.9. *(i) Let $g_n(x) = n^{-2}\mathbb{1}_{[0,n]}$. Clearly $g_n \to 0$ pointwise in x. Notice also that, for any n,*

$$\int g_n \, dm = \frac{1}{n}.$$

Hence,

$$\lim_{n\to\infty} \int g_n \, dm = 0 = \int \left(\lim_{n\to\infty} g_n(x) \right) dm.$$

(ii) Let $f_n(x) = n^{-1}\mathbb{1}_{[0,n]}$. Clearly $f_n \to 0$ pointwise in x. On the other hand:

$$\int f_n \, dm = 1$$

for every n. Hence,

$$\lim_{n\to\infty} \int f_n \, dm = 1 \neq 0 = \int \left(\lim_{n\to\infty} f_n(x) \right) dm.$$

(iii) Let $h_n(x) = n\mathbb{1}_{[0,1/n]}$. Clearly $h_n \to 0$ a.e. in x. On the other hand:

$$\int h_n \, dm = 1$$

for every n. Hence,

$$\lim_{n\to\infty} \int h_n \, dm = 1 \neq 0 = \int \left(\lim_{n\to\infty} h_n(x) \right) \, dm.$$

So we have to ask ourselves... when can we move the limit under the integral?

Theorem 4.5.10 (Monotone convergence theorem)**.** *Fix any measure space $(\Omega, \mathcal{F}, \mu)$. Suppose that f_n is an a.e. non-decreasing sequence of a.e. nonnegative measurable functions: there is a set $A \subset \Omega$ such that $\mu(\Omega \setminus A) = 0$ and*

$$0 \leq f_1 \leq f_2 \leq \cdots \leq f_n \leq \cdots \qquad \text{on } A.$$

Then, defining

$$f(x) = \sup_n f_n(x) \qquad \text{for all } x \in \Omega,$$

which is equal to $\lim f_n$ for all $x \in A$, we have

$$\int f(x) \, d\mu = \lim_{n\to\infty} \int f_n(x) \, d\mu.$$

Proof. First, note that f is measurable by Proposition 4.4.8 and the last exercise in Section 4.4.2. Next, we note that we may assume that $A = \Omega$. Indeed, if not, let, for all n,

$$\tilde{f}_n(x) = \mathbb{1}_A(x) f_n(x) \qquad \text{and} \qquad \tilde{f}(x) = \mathbb{1}_A f(x).$$

Then $0 \leq \tilde{f}_1 \leq \tilde{f}_2 \leq \dots$ holds everywhere and, by Exercise 4.4.14, we have, for all n,

$$\int f \, d\mu = \int \tilde{f} \, d\mu \qquad \text{and} \qquad \int f_n \, d\mu = \int \tilde{f}_n \, d\mu.$$

The rest of the proof proceeds by Fatou's lemma (Lemma 4.5.6). Indeed, by Fatou's lemma

$$\int f \, d\mu \leq \liminf_{n\to\infty} \int f_n \, d\mu.$$

On the other hand, since $f \geq f_n$ a.e., we have, for all n,

$$\int f \, d\mu \geq \int f_n \, d\mu.$$

Thus,

$$\int f \, d\mu \geq \limsup_{n\to\infty} \int f_n \, d\mu,$$

which concludes the proof. □

Exercise 4.5.8. *Use the monotone convergence theorem to compute the integral*

$$\lim_{n\to\infty} \int_0^1 \left[1 - x + x^2 - x^3 + \cdots + x^{2n} - x^{2n+1} \right] \, dm.$$

Exercise 4.5.9. *Show that the nonnegativity can not be ignored by constructing an increasing sequence of measurable functions whose integral does not converge to the integral of its limit.*

The monotone convergence theorem is usually used when the functions f_n are either of the form $f_n = \mathbb{1}_{A_n} f$ with A_n an increasing sequence of sets and f nonnegative or $f_n(x) = \max\{\alpha_n, f(x)\}$ where $\alpha_n \to \infty$.

Theorem 4.5.11 (Lebesgue dominated convergence theorem)**.** *Fix any measure space $(\Omega, \mathcal{F}, \mu)$. Suppose that f_n is a sequence of measurable functions such that*

$$|f_n(x)| \leq g(x) \qquad \text{for all } x \in \Omega,$$

where g is measurable function satisfying

$$\int g(x)\, d\mu < \infty.$$

If $f_n \to f$ a.e., then

$$\int f(x)\, d\mu = \lim_{n\to\infty} \int f_n(x)\, d\mu.$$

Proof. We assume without loss of generality that $f_n \geq 0$. If not, one proceeds by arguing with $(f_n)_+$ and $(f_n)_-$.

By Fatou's lemma,

$$\int f\, d\mu \leq \liminf_{n\to\infty} \int f_n\, d\mu.$$

and

$$\begin{aligned}\int g\, d\mu - \int f\, d\mu = \int (g - f)\, d\mu &\leq \liminf_{n\to\infty} \int (g - f_n)\, d\mu \\ &= \int g\, d\mu - \limsup_{n\to\infty} \int f_n\, d\mu.\end{aligned}$$

Rearranging this, we find

$$\limsup_{n\to\infty} \int f_n\, d\mu \leq \int f\, d\mu \leq \liminf_{n\to\infty} \int f_n\, d\mu,$$

which completes the proof. □

Example 4.5.12. *(i) Let us revisit Example 4.5.9.(i). Let $g_n = n^{-2}\mathbb{1}_{[0,n]}$. We know that $g_n \to 0$ everywhere and the conclusion of the MCT and DCT holds. Let us show that the conditions of DCT holds as well (which would save us the computation done in Example 4.5.9.(i)!). Define the dominating function*

$$u(x) = \frac{1}{\max\{1, x^2\}} \qquad \text{for any } x \in \mathbb{R}.$$

By elementary calculus, we know that u is integrable. Additionally, we immediately see that $|g_n| \leq u$ for every n. Hence, the conditions of the DCT are satisfied.

(ii) We compute the value of

$$\lim_{n\to\infty} n^3 \int_0^\infty \sin(x/n) e^{-(nx)^2}\, dm(x).$$

First, we make the change $y = nx$ to find

$$\lim_{n\to\infty} n^2 \int_0^\infty \sin(y/n^2) e^{-y^2}\, dm(y).$$

We notice that

$$n^2 \sin(y/n^2) \le y \quad \text{and} \quad n^2 \sin(y/n^2) \to y \quad \text{as } n \to \infty.$$

for all y, n. Hence, we may apply the DCT if we let

$$f_n(y) = n^2 \sin(y/n^2) e^{-y^2} \quad \text{and} \quad f(y) = g(y) = y e^{-y^2}.$$

Since g is integrable, we conclude by the DCT that

$$\lim_{n\to\infty} n^2 \int_0^\infty \sin(y/n^2) e^{-y^2}\, dm(y) = \int_0^\infty \lim_{n\to\infty} n^2 \sin(y/n^2) e^{-y^2}\, dm(y) = \int_0^\infty y e^{-y^2} dy = \frac{1}{2}.$$

(iii) Define, for each n and $x \in \mathbb{R}$,

$$\delta_n(x) = n e^{-2n|x|}.$$

Let φ be any continuous and bounded function. Consider

$$\int \delta_n(x)\varphi(x)\, dm(x).$$

Changing variables, we find

$$\int e^{-2|y|}\varphi(y/n)\, dm(y).$$

Clearly

$$h_n(y) := e^{-2|y|}\varphi(y/n) \to \varphi(0) e^{-2|y|} \quad \text{as } n \to \infty.$$

We have, for every n and y,

$$|h_n(y)| \le \|\varphi\|_\infty e^{-2|y|} =: g(y).$$

Since g is integrable, we may apply the DCT to find

$$\int h_n(y)\, dm(y) \to \int \lim_{n\to\infty} \left(\varphi(y/n) e^{-2|y|}\right)\, dm(y) = \int \varphi(0) e^{-2|y|})\, dm(y) = \varphi(0).$$

Rephrased, we have shown that

$$\lim_{n\to\infty} \int \delta_n(x)\varphi(x)\, dm(x) = \varphi(0).$$

We, thus, call δ_n an "approximation of a delta function" as the "delta function" δ is the (as yet ill-defined) operator such that $\int \varphi\delta\, dm = \varphi(0)$.

Notice that Example 4.5.9 provides examples where the conditions of the monotonic convergence and dominated convergence theorems are not met. Roughly, we may characterize these cases as "mass escaping to infinity" ($f_n = n^{-1}\mathbb{1}_{[0,n]}$) and "mass concentrating at a single point" ($h_n = n\mathbb{1}_{[0,1/n]}$).

For the two examples above, let us point out why the DCT cannot apply. In general, the smallest dominating function is defined by taking the supremum over all functions. Doing that here, we see:

$$\sup_n f_n(x) = \mathbb{1}_{[0,\infty)}\frac{1}{\max\{1,x\}} \quad \text{and} \quad \sup_n h_n(x) = \mathbb{1}_{[0,\infty)}\frac{1}{\min\{1,x\}}.$$

The former is not integrable because it is $1/x$ as $x \to \infty$ and the latter is not integrable because it is $1/x$ as $x \searrow 0$.

The approach used in Example 4.5.12.(i) and (ii) is common in DCT problems. This usually arises when analyzing multi-scale problems; i.e., problems where there are small and large structures interaction[27]. Let us look at a few more examples.

Example 4.5.13. *(i) Suppose that f is integrable, nonnegative and satisfies the following: there exists C with*

$$f(x) \leq C/|x|^3 \quad \text{for any } |x| \geq C$$

Find the value of α for which the following has a non-trivial, non-infinite limit:

$$\lim_{n\to\infty} n^\alpha \int_0^\infty \frac{3}{1+x^2} f(n(x-n))\,dm.$$

Roughly, the integrand will be centered at $x \approx O(n)$ and will be nontrivial only on a set of width $O(1/n)$ (otherwise $f(n(x-n)) \approx 0$). Hence, we expect:

$$n^\alpha \int_0^\infty \frac{3}{1+x^2} f(n(x-n))\,dm \approx n^\alpha \frac{3}{1+O(n^2)} O(1/n) \approx n^{\alpha-3}.$$

We, thus, guess that $\alpha = 3$ is the correct answer.

Let us take the quantity of interest and change variables with $y = n(x-n)$ with the "guess" $\alpha = 3$:

$$\begin{aligned} n^{\alpha-1}\int_{-n^2}^\infty \frac{3}{1+\left(n+\frac{y}{n}\right)^2} f(y)\,dm &= n^2 \int_{-n^2}^\infty \frac{3}{1+\left(n+\frac{y}{n}\right)^2} f(y)\,dm \\ &= \int_{-n^2}^\infty \frac{3}{\frac{1}{n^2}+\left(1+\frac{y}{n^2}\right)^2} f(y)\,dm. \end{aligned}$$

We would like to use the DCT; however, at $y = -n^2$, the denominator is very small! To get around this, let us decompose the integral into the "small part" and the part where we can apply the DCT:

$$\begin{aligned} &\int_{-n^2}^\infty \frac{3}{\frac{1}{n^2}+\left(1+\frac{y}{n^2}\right)^2} f(y)\,dm \\ &\quad = \int \mathbb{1}_{[-n^2/2,\infty]} \frac{3}{\frac{1}{n^2}+\left(1+\frac{y}{n^2}\right)^2} f(y)\,dm + \int \mathbb{1}_{[-n^2,-n^2/2]} \frac{3}{\frac{1}{n^2}+\left(1+\frac{y}{n^2}\right)^2} f(y)\,dm. \end{aligned}$$

[27] For example, a classical problem in the area of "homoegenization" is to understand heat flow through a heterogeneous object like concrete. In this case, the large scale is the "block" of concrete and the small scale is the mix of cement and coarse aggregates that make up the concrete. Heat might move through each "piece" of the concrete at a different rate, but, on the scale of building, we expect these fluctuations to be averaged out. But averaged out in what sense? This is the fundamental problem in homogenization.

Let us call the first integral on the right I_n and the second integral on the right J_n. Notice that we can apply the DCT to I_n with the dominating function

$$\left| \mathbb{1}_{[-n^2/2,\infty]} \frac{3}{\frac{1}{n^2} + \left(1 + \frac{y}{n^2}\right)^2} f(y) \right| \leq 12 f(y).$$

Here, we used that $y/n^2 > -1/2$. It follows, then, from the DCT that

$$\lim_{n\to\infty} I_n = \int \lim_{n\to\infty} \mathbb{1}_{[-n^2/2,\infty]} \frac{3}{\frac{1}{n^2} + \left(1 + \frac{y}{n^2}\right)^2} f(y)\, dm = \int 3 f(y)\, dm.$$

On the other hand, when n is large enough, we have

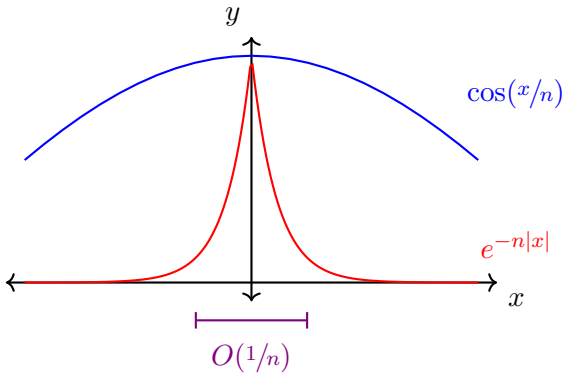
$$J_n \leq \int_{-n^2}^{-n^2/2} 3n^2 \frac{C}{|x|^3}\, dm \leq \int_{-n^2}^{-n^2/2} 3n^2 \frac{8C}{|n|^6}\, dm = \frac{12C}{n^2}.$$

We conclude that

$$\lim_{n\to\infty} n^3 \int_0^\infty \frac{3}{1+x^2} f(n(x-n))\, dm = \lim_{n\to\infty} (I_n + J_n) = 3 \int f(y)\, dm.$$

(ii) Find θ so that the following has a nontrivial, non-infinite limit:

$$\lim_{n\to\infty} n^\theta \int e^{-n|x|} \cos\left(\frac{x}{n}\right)\, dm.$$



Let us give a heuristic derivation for the value of θ. The exponential function clearly localizes the integrand to a set of width $O(1/n)$ – it is non-zero $O(1/n)$ away from the origin and approximately zero more than $O(1/n)$ away from the origin. On this set, $\cos(x/n)$ is essentially constant and equal to 1. Hence, we expect that the integral itself will be $O(1/n)$. Let us guess $\theta = 1$ and proceed from there.

The "localization" discussed above, suggests that we should change variables $y = nx$:

$$n \int e^{-n|x|} \cos\left(\frac{x}{n}\right)\, dm = \int e^{-|y|} \cos\left(\frac{y}{n^2}\right)\, dm.$$

Because $|\cos| \leq 1$, we note that a good dominating function is

$$\left| e^{-|y|} \cos\left(\frac{y}{n^2}\right) \right| \leq e^{-|y|}.$$

This is integrable. Hence, by the DCT,

$$\lim_{n\to\infty} n \int e^{-n|x|} \cos\left(\frac{x}{n}\right)\, dm = \lim_{n\to\infty} \int e^{-|y|} \cos\left(\frac{y}{n}\right)\, dm = \int \lim_{n\to\infty} e^{-|y|} \cos\left(\frac{y}{n}\right)\, dm$$
$$= \int e^{-|y|}\, dm = 2.$$

Let us make a small foray into antiderivatives or the "accumulation function." An important consequence of the DCT is that the antiderivatives of L^1 functions are continuous.

Proposition 4.5.14. *Let f be a measurable, absolutely integrable (i.e., L^1) function. Define, for any $x \in \mathbb{R}$,*

$$F(x) = \int_{-\infty}^{x} f \, dm = \int f \mathbb{1}_{(-\infty,x)} \, dm. \tag{4.5.1}$$

Then F is a continuous function.

Exercise 4.5.10. *Show that one cannot hope for a stronger theorem. Indeed, for each $\alpha \in (0,1)$, find an example f such that F, defined as in (4.5.1), satisfies*

$$\sup_{x \neq y} \frac{|F(x) - F(y)|}{|x-y|^\alpha} = \infty.$$

Proof. Fix any $x \in \mathbb{R}$ and $x_n \to x$. For ease, we assume that $x_n > x$ for each n. Then

$$F(x_n) - F(x) = \int f \mathbb{1}_{(-\infty,x_n)} \, dm - \int f \mathbb{1}_{(-\infty,x_n)} \, dm = \int f \mathbb{1}_{(x,x_n)} \, dm.$$

Notice that $|f\mathbb{1}_{(x,x_n)}| \leq |f|$, so that we may apply the DCT to find

$$\lim_{n\to\infty} (F(x_n) - F(x)) = \int \lim_{n\to\infty} f \mathbb{1}_{(x,x_n)} \, dm = \int 0 \, dm = 0.$$

□

4.6. L^p**-spaces.** Now that we have a notion of integral, it is tempting to define the L^p-spaces. We are nearly there! First, let us investigate what it means for the L^p-norms to be small. Afterwards, we use this to verify that (in a suitable sense) the L^p-norms are positive definite.

This result is called "Chebyshev's inequality" in the analysis community and "Markov's inequality" in the probability community. The fact that it is named, despite having an exceedingly simple proof, indicates that it is very useful.

Proposition 4.6.1 (Chebyshev's inequality / Markov's Inequality)**.** *Suppose that $(\Omega, \mathcal{F}, \mu)$ is a measure space and f is a measurable function. Then, for any $\lambda > 0$, we have*

$$\mu\left(\{x : |f(x)| \geq \lambda\}\right) \leq \frac{\int |f| \, d\mu}{\lambda}.$$

This inequality is used *very often* in probability.

Proof. Notice that

$$\lambda \mathbb{1}_{\{x:|f(x)|\geq\lambda\}} \leq |f|.$$

Hence, by the monotonicity of integration:

$$\lambda\mu(\{x : |f(x)| \geq \lambda\}) = \int \lambda \mathbb{1}_{\{x:|f(x)|\geq\lambda\}} \, d\mu \leq \int |f| \, d\mu.$$

Rearranging this yields the claim. □

The upshot to this is that any function whose integral is zero is equal to zero almost everywhere.

Lemma 4.6.2. *Let $(\Omega, \mathcal{F}, \mu)$ be a measure space and f a measurable function such that, for some $p \in [1, \infty)$,*

$$\int |f|^p \, d\mu = 0.$$

Then $f = 0$ a.e.[28]

Exercise 4.6.1. *Prove Lemma 4.6.2. You may find it helpful to apply Proposition 4.6.1 countably many times with $\lambda = 1/n$, using countable subadditivity (Proposition 4.2.4), and the fact that:*

$$\{x : f = 0\} = \bigcup_{n=1}^{\infty} \{x : |f| \geq 1/n\}.$$

Let us note, however, that Lemma 4.6.2 is not helpful in the case of the L^∞-norm. To this point, we have been using the sup-norm. This is problematic because

$$\sup_x \mathbb{1}_{\{\pi\}}(x) = 1,$$

but $\mathbb{1}_{\{\pi\}}$ is not equal to zero a.e. We need a better definition.

Actually, one thing we see in Proposition 4.6.1 is that

$$\mu\left(\{x : |f(x)| = \infty\}\right) \leq \liminf_{\lambda \to \infty} \mu\left(\{x : |f(x)| \geq \lambda\}\right) \leq \liminf_{\lambda \to \infty} \frac{\int |f| \, d\mu}{\lambda} = 0.$$

In other words, f has to be finite a.e. It might be that there is a bound for f a.e. This motivates the following definition:

Definition 4.6.3. *If f is a measurable function on a measure space $(\Omega, \mathcal{F}, \mu)$ with*[29] *$\mu(\Omega) > 0$, we define*

$$\operatorname{ess\,sup} f = \sup\left\{B : \mu\left(\{x : f(x) > B\}\right) > 0\right\}.$$

Let us note that this is not the "standard" definition. Usually, it is defined by

$$\operatorname{ess\,sup} f = \inf\left\{B : \mu\left(\{x : f(x) > B\}\right) = 0\right\}. \tag{4.6.1}$$

One can check that these two definitions are equivalent. This is clear when $\operatorname{ess\,sup} f < \infty$. However, if this is not the case, then, in (4.6.1),

$$\{B : \mu\left(\{x : f(x) > B\}\right) = 0\} = \emptyset$$

and one must recall that, by definition, $\inf \emptyset = +\infty$.

Exercise 4.6.2. *Show that* $\operatorname{ess\,sup} \mathbb{1}_{\{\pi\}} = 0$.

Remark 4.6.4. *It is easy to carry over our definitions of the L^p-spaces (see* (1.5.1)-(1.5.2)*) to our "better" setting here when $p \in [1, \infty)$. When $p = \infty$, we see that the* $\operatorname{ess\,sup} |f|$ *is the correct notation of L^∞-norm.*

[28] Not everywhere though! Indeed, $\int \mathbb{1}_{\mathbb{Q}} \, dm = 0$ but $\mathbb{1}_{\mathbb{Q}}$ is not the zero function.

[29] The essential supremum will always be $-\infty$ when $\mu(\Omega) = 0$. Actually, such trivial measure spaces should probably be analyzed via a different lens than that of this section. We, thus, do not lose generality in our discussions if we always assume our measure spaces to have positive measure.

Example 4.6.5. *For these examples, we use the Lebesgue measure.*

1. $\operatorname{ess\,sup} \mathbb{1}_{\mathbb{Q}} = 0$ – *Notice that* $m(\{x : \mathbb{1}_{\mathbb{Q}}(x) > a\}) = 0$ *if and only if* $a \geq 0$.
2. $\operatorname{ess\,sup} {}^{1}/_{|x|} = \infty$ – *Notice that, for any* $a > 0$, $m(\{x : \mathbb{1}_{\mathbb{Q}}(x) > a\}) = {}^{2}/_{a} > 0$.

The following definition is essentially that of (1.5.1)-(1.5.2), except that it is now phrased for arbitrary measurable functions (and arbitrary measures), so that we do not require continuity. This will be a very important point!

Definition 4.6.6. *Fix a measure space* $(\Omega, \mathcal{F}, \mu)$. *When* $p \in [1, \infty)$, *we let*

$$L^p = \Big\{ f : \Omega \to \mathbb{R} : f \text{ is measurable and } \int |f|^p \, d\mu < \infty \Big\}.$$

Additionally,

$$L^\infty = \Big\{ f : \Omega \to \mathbb{R} : f \text{ is measurable and } \operatorname{ess\,sup} |f| \, d\mu < \infty \Big\}.$$

These spaces come with the norms:

$$\|f\|_p = \left(\int |f|^p \, d\mu \right)^{\frac{1}{p}} \quad \text{when } p \in [1, \infty) \qquad \text{and} \qquad \|f\|_\infty = \operatorname{ess\,sup} |f|.$$

Often, when we want to emphasize the set Ω or the measure μ, we denote L^p by one of the following:

$$L^p(\mu), \quad L^p(\, d\mu), \quad L^p_\mu, \quad \text{or} \quad L^p(\Omega).$$

That $\| \cdot \|_p$ is a norm and L^p a linear space follows by adapting the proofs from Section 1.5. We summarize this in the following:

Proposition 4.6.7. *The space* L^p *with the metric* $\| \cdot \|_p$ *is a normed linear space when* $p \in [1, \infty]$.

Remark 4.6.8. *Consider the measure space* $(\mathbb{N}, \mathcal{P}^{\mathbb{N}}, \mu)$ *where*

$$\mu(X) = \begin{cases} n & \text{if } X \text{ has } n \in \mathbb{N} \cup \{0\} \text{ elements,} \\ \infty & \text{if } X \text{ is infinite.} \end{cases}$$

Then $L^p(\mu) = \ell^p(\mathbb{N})$, *where* ℓ^p *is defined in Section 1.5. Note that one can easily extend the measure* μ *to* $(\mathbb{R}, \mathcal{P}^{\mathbb{R}})$.

Our momentary goal is to show that the L^p-spaces are complete because this tells us that, roughly, they will have nice "convergence properties." Let us quick state a lemma that is useful in the proof and can be proved easily using basic calculus.

Exercise 4.6.3. *Show that there exists* $C_p > 0$, *depending only on* $p \in [1, \infty)$, *such that* $(a + b)^p \leq 2^{p-1}(a^p + b^p)$ *for any* $a, b \geq 0$.

We now state and prove the main step towards showing that the L^p-spaces are complete.

Proposition 4.6.9. *Fix any complete measure space $(\Omega, \mathcal{F}, \mu)$ and any $p \in [1, \infty]$. Let $f_n \in L^p$ be an absolutely summable sequence:*

$$\sum_{n=1}^{\infty} \|f_n\|_p < \infty.$$

Then there is a function $F \in L^p$ such that

$$\lim_{n\to\infty} \left\| F - \sum_{k=1}^{n} f_k \right\|_p = 0.$$

In other words,

$$\sum_{k=1}^{n} f_k \to F \qquad \text{in } L^p.$$

Exercise 4.6.4. *Prove the proposition when $p = \infty$.*

Proof. We show the proof for $p < \infty$. Let us first define a useful function. Let

$$h_n(x) = \sum_{k=1}^{n} |f_k(x)|.$$

Notice that this is a monotonic sequence of functions. By Chebyshev's inequality, for any $\lambda > 0$,

$$\mu\left(\{x : h_n(x) > \lambda\}\right) = \mu\left(\{x : |h_n(x)|^p > \lambda^p\}\right) \leq \frac{\|h_n\|_p^p}{\lambda^p} \leq \frac{C}{\lambda^p}, \tag{4.6.2}$$

where

$$C = \sup_n \left(\|f_1\|_p + \cdots + \|f_n\|_p\right) = \sum_{k=1}^{\infty} \|f_k\|_p < \infty.$$

Let

$$A = \left\{x : \lim_{n\to\infty} h_n(x) = \infty\right\}.$$

Notice that

$$A \subset \bigcap_{\lambda=1}^{\infty} \bigcup_{n=1}^{\infty} A_{n,\lambda} \tag{4.6.3}$$

where

$$A_{n,\lambda} = \left\{x : h_n(x) > \lambda^p\right\}.$$

Since h_n is an increasing sequence (it is the sum of nonnegative functions), $A_{n,\lambda}$ is an increasing sequence of sets:

$$A_{1,\lambda} \subset A_{2,\lambda} \subset A_{3,\lambda} \subset \cdots.$$

Then, using (4.6.2), we have

$$\mu\left(\bigcup_{n=1}^{\infty} A_{n,\lambda}\right) = \lim_{n\to\infty} \mu(A_{n,\lambda}) \leq \liminf_{n\to\infty} \frac{\int |h_n|^p \, d\mu}{\lambda} \leq \frac{C}{\lambda}. \tag{4.6.4}$$

Using (4.6.4) and similar reasoning, we then find

$$\mu\left(\bigcap_{\lambda=1}^{\infty} \bigcup_{n=1}^{\infty} A_{n,\lambda}\right) = \lim_{\lambda\to\infty} \mu\left(\bigcup_{n=1}^{\infty} A_{n,\lambda}\right) = 0.$$

By the completeness of $(\Omega, \mathcal{F}, \mu)$ and the inclusion (4.6.3), we deduce that A is measurable and satisfies

$$\mu(A) = 0. \tag{4.6.5}$$

Let us note that, were A not measurable, the remainder of the proof would not work. This is the only place in these notes where we use completeness of the measure space.

For each $x \in A^c$, we see that $h_n(x)$ is a bounded, increasing sequence, so it has a limit. For each such x, we let

$$h(x) = \lim_{n\to\infty} h_n(x).$$

We can define h to be $+\infty$ on A^c. Due to (4.6.5), we see that $h_n \to h$ a.e. By Fatou's lemma (or the Monotone Convergence Theorem), we see that

$$\int h^p \, d\mu \leq \liminf_n \int h_n^p \leq C.$$

In what follows, we use h^p as our dominating function, and the above reassures us that it is integrable.

Let us return to

$$F_n(x) := \sum_{k=1}^{n} f_n(x).$$

For all $x \in A^c$, this series is absolutely summable due to the work above. Hence, the series converges: let

$$F(x) = \lim_{n\to\infty} F_n(x). \tag{4.6.6}$$

We stress that this holds a.e. because $\mu(A) = 0$. We can define F to be zero on A^c. Since $|F_n|^p \leq h_n^p \leq h^p$, the DCT implies that

$$\int |F| \, d\mu = \lim_{n\to\infty} \int |F_n|^p \, d\mu \leq C.$$

Using the DCT again, with the dominating function (see Exercise 4.6.3)

$$|F - F_n|^p \leq C_p(|F|^p + |F_n|^p) \leq 2C_p h^p,$$

we find

$$\lim_{n\to\infty} \|F - F_n\|_p^p = \lim_{n\to\infty} \int |F - F_n|^p \, d\mu = \int \lim_{n\to\infty} |F - F_n|^p \, d\mu = 0.$$

The last equality holds because of (4.6.6). □

Exercise 4.6.5. *Use Proposition 4.6.9 to construct $f \in L^1(\mathbb{R})$ such that, for every $q \in \mathbb{Q}$,*

$$\lim_{x\to q} f(x) = \infty.$$

We now establish the completeness of the L^p-spaces. Note that this is the reason why the Lebesgue approach to integration is the *correct* one.

Theorem 4.6.10 (Riesz-Fischer theorem)**.** *Let $(\Omega, \mathcal{F}, \mu)$ be a complete measure space. For $p \in [1, +\infty]$, L^p is a complete normed linear space (i.e., it is a* Banach space*).*

Proof. The fact that L^p is a normed linear space follows from Proposition 4.6.7. We show completeness here. Take f_n to be a Cauchy sequence in L^p. Take any subsequence[30] f_{n_k} such that

$$\|f_{n_{k+1}} - f_{n_k}\|_p \leq \frac{1}{2^k}.$$

It is enough to show that f_{n_k} converges[31]. Let us rewrite this as

$$f_{n_{k+1}} = f_{n_1} + \sum_{i=1}^{k}(f_{n_{i+1}} - f_{n_i}) = f_{n_1} + \sum_{i=1}^{k} h_i,$$

where $h_i = f_{n_{i+1}} - f_{n_i}$. Since

$$\sum_{i=1}^{k} \|h_i\|_p \leq \sum_{i=1}^{k} \frac{1}{2^k} < 1,$$

we may apply Proposition 4.6.9 to conclude that there is f such that

$$f = \lim_{k\to\infty} \Big(f_{n_1} + \sum_{i=1}^{k} h_i\Big).$$

This limit is taken with respect to $\|\cdot\|_p$. Hence, we have

$$\begin{aligned} 0 &= \lim_{k\to\infty} \Big\|f - \Big(f_{n_1} + \sum_{i=1}^{k} h_i\Big)\Big\|_p \\ &= \lim_{k\to\infty} \Big\|f - \Big(f_{n_1} + \sum_{i=1}^{k}(f_{n_{i+1}} - f_{n_i})\Big)\Big\|_p = \lim_{k\to\infty} \Big\|f - f_{n_k}\Big\|_p. \end{aligned}$$

This concludes the proof. □

4.7. **One last notion of convergence.** Let us consider an illustrative example.

Example 4.7.1. *For any $\theta \in \mathbb{R}$, consider the sequence*

$$f_n = 2^{\theta k}\mathbb{1}_{\left[\frac{n-2^k}{2^k}, \frac{n-2^k+1}{2^k}\right]} \qquad \text{where } k \text{ is such that } 2^k \leq n < 2^{k+1}.$$

Especially when $\theta = 0$, this is often called the typewriter sequence. In a sense, it is clear that f_n "tends to zero." Indeed, the set on which f_n is different from zero is $O(1/n)$.

However, if $\theta \geq 0$, $f_n \not\to 0$ in an almost everywhere *sense. Indeed, given any x and any N, there is $n \geq N$ such that $f_n(x) \geq 1$! Moreover, if $\theta \geq 1$, then $f_n \not\to 0$ in any L^p-norm!*

Exercise 4.7.1. *Given $p \in [1, \infty]$, for which θ does $f_n \to 0$ in the L^p-norm?*

So we introduce a new notion of convergence that is very important in the study of probability. We note immediately that the typewriter sequence tends to zero in this sense.

[30] We showed this on a homework!

[31] We showed on a homework, that if a Cauchy sequence has a convergence subsequence, then the sequence is convergent as well and has the same limit.

Definition 4.7.2. *Given a measure space $(\Omega, \mathcal{F}, \mu)$, a sequence of measurable functions f_n converges to a measurable function f in measure (equivalently, in probability) if, for all $\varepsilon > 0$,*

$$\lim_{n\to\infty} \mu\left(\{x : |f(x) - f_n(x)| > \varepsilon\}\right) = 0.$$

We now have many notions of convergence, and it takes some practice to get used to how they relate. The following shows some of them:

Exercise 4.7.2. *Let $(\Omega, \mathcal{F}, \mu)$ be a measure space with measurable functions f, f_1, f_2, ...*

(i) If $f_n \to f$ in L^p then $f_n \to f$ in measure.

(ii) If $f_n \to f$ a.e. and $\mu(\Omega) < \infty$ then $f_n \to f$ in measure.

(iii) If $f_n \to f$ in measure then there is a subsequence f_{n_k} that converges to f a.e.

(iv) If $f_n \to f$ in L^p then there is a subsequence f_{n_k} that converges to f a.e.

(v) If $f_n \to f$ a.e. and there is a dominating function then $f_n \to f$ in L^p for any $p \in [1, \infty)$.

(vi) If $f_n \to f$ in L^p and $\mu(\Omega) < \infty$, then $f_n \to f$ in L^q for every $q < p$.

Note that some directions are not represented here – find counterexamples for them!

We note that, in practice, one often chooses a notion of convergence to work with because it (i) pairs well with the mathematical techniques associated to the problem, and (ii) it has physical relevance. These two points are sometimes in conflict! The strongest notion of convergence that has the best "real-world interpretation" might be very difficult to show, so one might instead compromise.

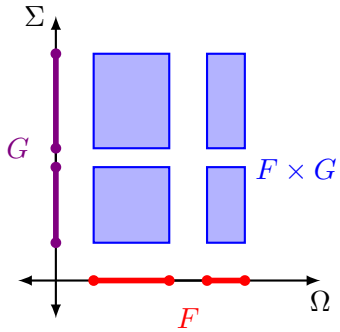
4.8. **Product measures: Fubini's theorem and Tonelli's theorem.** How might we define a measure on $\mathbb{R}^2$? Or, more generally, given $(\Omega, \mathcal{F}, \mu)$ and $(\Sigma, \mathcal{G}, \nu)$, how do we define a measure "$\mu \times \nu$ on the product space $\Omega \times \Sigma$?

We first begin by defining the "rectangles"

$$\mathcal{R} = \{F \times G : F \in \mathcal{F}, G \in \mathcal{G}\},$$

which make a natural basis for our product sigma algebra $\sigma(\mathcal{R})$. We can easily define the proto-measure of any rectangle $F \times G \in \mathcal{R}$ in the geometric natural way:

$$(\mu \times \nu)(F \times G) = \mu(F)\nu(G).$$



One can extend then this to a measure on $\mathcal{F} \times \mathcal{G}$ by our "outer measure" strategy from before. That is, let $(\mu \times \nu)^*$ be defined by, for all $A \subset \Omega \times \Sigma$,

$$(\mu \times \nu)^*(A) = \inf\left\{ \sum_{i=1}^{\infty} \mu(F_i)\nu(G_i) : A \subset \bigcup_{i=1}^{\infty} F_i \times G_i \right\}$$

and then let $\mathcal{F} \times \mathcal{G} \supset \sigma(\mathcal{R})$ be the set of all sets $A \subset \Omega \times \Sigma$ such that

$$(\mu \times \nu)^*(E) = (\mu \times \nu)^*(A \cap E) + (\mu \times \nu)^*(A^c \cap E) \qquad \text{for all } E \subset \Omega \times \Sigma.$$

In practice, most sets will be in $\sigma(\mathcal{R})$; however, we prefer $\mathcal{F} \times \mathcal{G}$ because it is complete (see Definition 4.4.18). Sifting through the details involved in constructing the product measure is a good exercise for those that be confident in their knowledge. For the purposes of these notes, though, we avoid the technical details as we only want access to the tools that are Fubini's and Tonelli's theorems.

Exercise 4.8.1. *Consider $(\mathbb{R}^2, \sigma(\mathcal{R}), m \times m)$, where $\mathcal{R}$ is the set of products of Lebesgue measurable sets. Show that this measure space is not complete.*

Hint: if A is any set and B is a Lebesgue measurable set of measure zero, then what is $(m \times m)^*(A \times B)$? Is $A \times B \in \sigma(\mathcal{R})$?

We note that this gives the basis for the Lebesgue measure on $\mathbb{R}^n$. In other words, we can now safely define measurable functions on $\mathbb{R}^n$ and integrate them.

Although we will skip the proofs, we state two important results about integration with respect to product measures.

Theorem 4.8.1 (Fubini's theorem)**.** *Suppose that $(\Omega, \mathcal{F}, \mu)$ and $(\Sigma, \mathcal{G}, \nu)$ are two complete*[32] *measure spaces. If f is an integrable function on $\mu \times \nu$ then*

(i) For a.e. x (resp. y) $f_x(\cdot) = f(x, \cdot)$ is a μ-integrable function of y (resp. $f_y(\cdot) = f(\cdot, y)$ is a ν-integrable function of x);

(ii) The function $\int f \, d\mu(x)$ is an ν-integrable function (resp. $\int f d\nu(y)$ is a μ-integrable function);

(iii) $\int \left(\int f(x, y) \, d\mu(x) \right) d\nu(y) = \int f d(\mu \times \nu) = \int \left(\int f(x, y) \, d\nu(y) \right) d\mu(x)$.

Theorem 4.8.2 (Tonelli's Theorem)**.** *The conclusion of Fubini's theorem holds if f is nonnegative and $(\Omega, \mathcal{F}, \mu)$ and $(\Sigma, \mathcal{G}, \nu)$ are σ-finite*[33]*.*

It is tempting not to check the hypotheses of these theorems, but one must be careful!

Example 4.8.3. *Consider the function $f(x, y) = \frac{x^2 - y^2}{(x^2 + y^2)^2}$ on the set $[0, 1]^2$. Let us consider the various integrals outlined in Fubini's theorem part (iii).*

(i) Consider $\int \left(\int f(x, y) \, dx \right) dy$. An elementary argument yields

$$\int_{[0,1]} \left(\int_{[0,1]} f(x, y) \, dx \right) dy = \int_{[0,1]} \frac{-1}{1 + y^2} \, dy < 0.$$

(ii) Consider $\int \left(\int f(x, y) \, dy \right) dx$. An elementary argument yields

$$\int_{[0,1]} \left(\int_{[0,1]} f(x, y) \, dy \right) dx = \int_{[0,1]} \frac{1}{1 + x^2} \, dx > 0.$$

[32] That is, every measure zero set is measurable

[33] That is, there exists $A_n \in \mathcal{F}$ (resp. $B_n \in \mathcal{G}$) such that $\mu(A_n) < \infty$ and $\cup_n A_n = \Omega$ (resp. $\nu(B_n) < \infty$ and $\cup_n B_n = \Sigma$).

(iii) Because f is odd with respect to reflection of $y = x$ (that is, $f(x, y) = -f(y, x)$) and $[0, 1]^2$ is preserved by this reflection, it is "clear" that

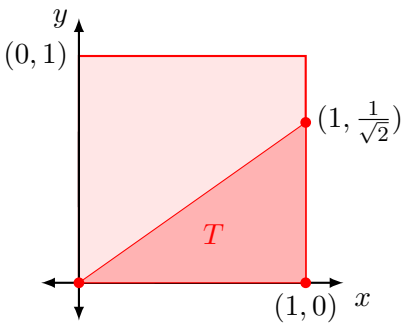
$$\int_{[0,1]^2} f(x, y)\, dm = 0.$$

Why do we get three different answers*?!? Well, we never checked the conditions of Fubini's theorem or Tonelli's Theorem. Clearly Tonelli's theorem does not apply because f is not nonnegative. Let us check the condition of Fubini's theorem; that is, let us see if f is integrable. Indeed,*

$$\int_{[0,1]^2} |f(x, y)|\, dm \geq \int_T |f(x, y)|\, dm.$$

where T is the triangle defined by

$$T = \{(x, y) \in [0, 1]^2 : x > \sqrt{2}y\}.$$



The light red box is $[0, 1]^2$ and the darker red box is T.

Notice that, for $(x, y) \in T$, we have

$$x^2 - y^2 = \frac{x^2}{4} + \frac{3x^2}{4} - y^2 \geq \frac{x^2}{4} + \frac{3y^2}{2} - y^2 \geq \frac{x^2 + y^2}{4}.$$

Hence

$$\int_T |f(x, y)|\, dm \geq \int_T \frac{1}{4} \frac{1}{x^2 + y^2}\, dm.$$

An elementary argument using polar coordinates shows that:

$$\int_T \frac{1}{4} \frac{1}{x^2 + y^2}\, dm = \infty.$$

So f is not integrable! This is why we get three different answers in (i), (ii), and (iii).

Exercise 4.8.2. *Compute $\displaystyle\int_0^\infty \frac{\sin x}{x}\, dm$. You may find it useful to use that $\displaystyle\frac{1}{x} = \int_0^\infty e^{-xt}\, dm(t)$.*

5. Probability theory

One advantage to the generality in which we have studied measure theory is that we may now apply it to the study of probability theory. In a sense, basic probability theory is applied measure theory, although, as we shall see, it can quickly get significantly more complex as one begins to think about random variables, processes, etc.

In this setting, we use $\mathbb{E}$ to denote the integral (expectation) and $\mathbb{P}$ to denote the probability measure. We clarify this further below when discussing random variables (i.e., measurable functions).

5.1. Densities, distributions functions, and the Stieltjes measure.

For a moment, let us play with a few different methods for obtaining measures and aspects of (finite) measures.

5.1.1. **Densities.** One can construct new measures from existing ones by using any non-negative measurable function: given a measure space $(\Omega, \mathcal{F}, \mu)$ and $f \geq 0$ that is μ-measurable, let

$$\mu_f(A) = \int_A f \, d\mu. \tag{5.1.1}$$

In this case, we often write $d\mu_f = f \, d\mu$. One can check that this is a measure defined on the same σ-algebra $\mathcal{F}$ as μ. If $f \in L^1$ and is non-zero we can "normalize" this to obtain a probability measure:

$$\nu(A) = \frac{1}{Z} \int_A f \, d\mu \qquad \text{where } Z = \int f \, d\mu.$$

In this case, we often write $d\mu_f = (f/z) \, d\mu$. Let us point out that not all probability measures have a density and that there is a general theory for how to know if your measure has a density[34]

Example 5.1.1. *Perhaps the most widely used example of this is the normal distribution with mean μ and variance σ^2:*

$$d\mathbb{P} = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} dm.$$

Roughly, an event has value in (a, b) if, for example,

$$\mathbb{P}((a,b)) = \int_a^b \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} dm.$$

This is used in a huge number of applications due to its "universality" – it arises any time a number of "not too correlated" random variables are summed. This is the central limit theorem! For example, exam scores!

5.1.2. **Stieltjes measures.** Suppose we have a function $F : \mathbb{R} \to \mathbb{R}$ that is non-decreasing that is "nice enough" (to be defined momentarily). Then we can define an associated measure on Borel sets by

$$\mu_F((a,b]) = F(b) - F(a).$$

If we want to assign a measure to an interval, then we should have, for any sequence $x_n \nearrow b$,

$$\mu_F((a,b)) = \mu_F\left(\bigcup_{n=1}^{\infty} (a, x_n]\right) = \lim_{n\to\infty} \mu_F((a,x_n]) = \lim_{n\to\infty} F(x_n) - F(a). \tag{5.1.2}$$

This will only be well defined if F has "limits on the left" (that is, $\lim_{n\to\infty} F(x_n)$ exists whenever $x_n \nearrow b$ for some b).

On the other hand, if we want μ_F to be a measure according to Definition 4.3.7, we must have the following: given any sequence $x_n \searrow a$,

$$\begin{aligned} F(x_1) - F(a) = \mu_F((a,x_1]) &= \mu_F\left(\bigcup_{n=1}^{\infty} (x_{n+1}, x_n]\right) \\ &= \sum_{n=1}^{\infty} (F(x_n) - F(x_{n+1}) = F(x_1) - \lim_{n\to\infty} F(x_n), \end{aligned}$$

[34] see the Radon-Nikodym theorem.

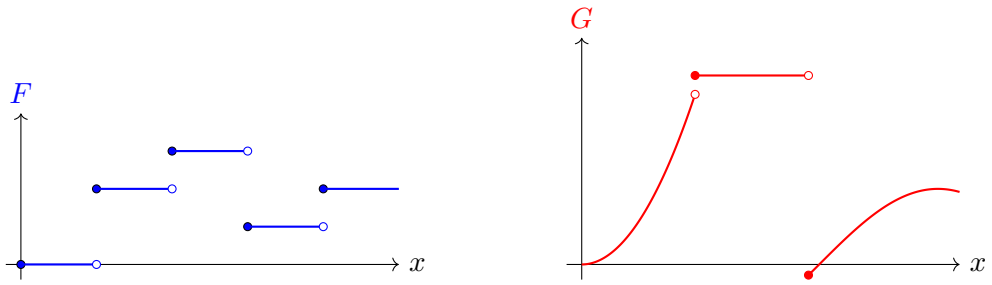


Figure 6: Two examples of càdlàg functions F and G.

where the last equality uses the definition of infinite series and the fact that this series is telescoping. We must, thus, have a limit on the right, call it $F(a^+)$. Plugging this in above, we see that

$$F(a) = F(a^+).$$

In other words, F must be "continuous on the right."

Definition 5.1.2. *A* cadlàg function $F : \mathbb{R} \to \mathbb{R}$ *is one that, for any x_0, is continuous on the right (***c***ontinue* **à** **d***roite) and has a limit on the left (***l***imite* **à** **g***auche); that is,*

$$\lim_{x \searrow x_0} F(x) = F(x_0) \qquad \text{and} \qquad \lim_{x \nearrow x_0} F(x) \text{ exists.}$$

Any continuous function is càdlàg; however, we point out that F need not be continuous! Take a look at the examples in Figure 6. In fact, one very important example of a càdlàg function is the paths of a Lévy process (which exhibits "random jumps" to move around). These arise in mathematical biology (e.g., modeling the movement of a bird or other animal that makes occasional large movements, modeling genetic change in a population), mathematical finance, and the microscopic behavior of any nonlocal PDE (e.g., the Boltzmann equation, modeling the density of particles in a diffuse gas).

We can create a measure space analogously to our construction of the Lebesgue measure. Let μ_F^* be the outer measure defined by

$$\mu_F^*(A) = \inf\left\{ \sum_{i=1}^{\infty} (F(b_i) - F(a_i)) : \bigcup_{i=1}^{\infty} (a_i, b_i] \supset A \right\}, \tag{5.1.3}$$

let $\mathcal{F}_F$ be the σ-algebra

$$\mathcal{F}_F = \{A : \mu_F^*(E) = \mu_F^*(A \cap E) + \mu_F^*(A^c \cap E) \text{ for all } E \subset \mathbb{R}\}, \tag{5.1.4}$$

and define $\mu_F(A) = \mu_F^*(A)$ for all $A \in \mathcal{F}_F$. Then $(\mathbb{R}, \mathcal{F}_F, \mu_F)$ is a measure space. Further, since we have defined μ_F on intervals $(a, b]$ and, by the work in (5.1.2), (a, b), it must be that $\mathcal{F}_F \supset \mathcal{B}$.

Actually, if our function is "nice" enough, we can generalize this procedure to non-monotonic functions F. In this case, we end up with a signed measure; that is, a measure that need not be positive. Actually, without the condition of nonnegativity, linear combinations of measures are measures. One can then define various norms on the linear spaces of measures and analyze these interesting spaces. We do not pursue this further here.

Example 5.1.3. *(i) $F(x) = x$: then $\mu_F = m$. Indeed, we see that, by* (5.1.2),

$$\mu_F((a, b)) = b - a$$

for all $a < b$. One can then check that the outer measures agree: $\mu_F^* = m^*$. It follows that they generate the same σ-algebra and the same measure.

(ii) $F(x) = \mathbb{1}_{[0,\infty)}$*: then* $\mu_F = \delta_0$*. In other words,* $\mu_F(A) = 0$ *if* $0 \notin A$ *and* $\mu_F(A) = 1$ *if* $0 \in A$*. Why? First, notice that, if* $b < 0$*, then*

$$\mu_F((a,b]) = F(b) - F(a) = 0 - 0 = 0,$$

and if $a \geq 0$*, then*

$$\mu_F((a,b]) = F(b) - F(a) = 1 - 1 = 0.$$

On the other hand, if $a < 0$ *and* $b \geq 0$*, we have*

$$\mu_F((a,b]) = F(b) - F(a) = 1 - 0 = 1.$$

Moreover,

$$\mu_F(\{0\}) = \lim_{n\to\infty} \mu_F((-1, 1/n]) + \mu_F(\{0\}) = \mu_F((-1,0)) + \mu_F(\{0\}) = \mu_F((-1,0]) = 1. \quad (5.1.5)$$

Hence, if $0 \in A$*,*

$$1 = \mu_F(\{0\}) \leq \mu_F(A) \leq \mu_F((-\infty,\infty]) = 1.$$

On the other hand, arguing as in (5.1.5)*, we find that, if* $0 \notin A$*,*

$$0 \leq \mu_F(A) \leq \mu_F((-\infty,0)) + \mu_F((0,\infty)) = 0 + 0 = 0.$$

Roughly, where F is continuous, we get something "like" the Lebesgue measure with a density, while if F has a "jump" then we get a constant multiple of the δ-function at that location. This is not quite true, see Exercise 5.2.3 and the discussion above it.

Remark 5.1.4. *Suppose we are trying to identify* μ_F *as a measure* μ *that we can write down explicitly. Ignoring the technical issues related to* σ*-algebras, we can deduce that* $\mu = \mu_F$ *if*

$$\mu_F((a,b]) = \mu((a,b]) \qquad \text{for every } a < b.$$

This is, roughly, because we use intervals of this form to define the measure of arbitrary sets via (5.1.3)*. The interested reader might hope for a rigorous statement. We include this in Appendix C.*

5.1.3. **Distribution functions.** We can also start from a measure space $(\mathbb{R}, \mathcal{F}, \mu)$ and, roughly, find the F that generates it. For this, we need $\mathcal{F} \supset \mathcal{B}$. If μ is a finite[35] measure (i.e., $\mu(\Omega) < \infty$), then we can define its distribution function:

Definition 5.1.5. *The* (cumulative) *distribution function of* μ *is*

$$F_\mu(x) = \int \mathbb{1}_{(-\infty,x]}\, d\mu.$$

Let us underline that F_μ is well-defined by the assumption that $\mathcal{F} \supset \mathcal{B}$, which ensures that $(-\infty, x]$ is measurable.

In fact, were one to put into action the procedure outlined in Section 5.1.2 that defines the Stieltjes measure associated to F_μ, call it $\tilde{\mu}$, one finds that $\tilde{\mu} = \mu$ on $\mathcal{B}$. A subtle technical note is that the σ-algebra of the measure space $(\mathbb{R}, \tilde{\mathcal{F}}, \tilde{\mu})$ generated by F_μ may differ from that of $(\mathbb{R}, \mathcal{F}, \mu)$; that is, it may be that $\mathcal{F} = \tilde{\mathcal{F}}$. See Exercise 5.1.1.

[35] If μ is not finite (e.g., the Lebesgue measure), one can define the distribution function from a particular point like 0: $F(t) = \mu((0,t])$, taking care to define F appropriately for $t < 0$ as well.

Exercise 5.1.1. *Let F be the distribution function associated to $(\mathbb{R}, \mathcal{B}, m_f)$, where $0 < f \in L^1$ and m_f has density f as in (5.1.1). Show that the Lebesgue-Stieltjes measure space generated by the process outlined in Section 5.1.2 is $(\mathbb{R}, \mathcal{L}, m_f)$. Note that this is not original measure space because the σ-algebras of measurable set differ.*

Recall that we showed in Proposition 4.5.14 that F_μ is continuous if $d\mu = f\,dm$ for some $f \in L^1$. However, F_μ need not be continuous if μ has a "δ component."

Example 5.1.6. *Given any nonnegative $f \in L^1$, we can define, for any Lebesgue measurable set A,*

$$\mu(A) := \delta_0(A) + \int_A f\,dm.$$

One can check that

$$F_\mu(x) = \mathbb{1}_{(-\infty,0]}(x) + \int \mathbb{1}_{(-\infty,x]} f\,dm,$$

which has a jump discontinuity at $x = 0$.

5.2. **Random variables.** At the highest level, a random variable on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is simply a measurable function:

Definition 5.2.1. *If $(\Omega, \mathcal{F}, \mathbb{P})$ is a measure space and $X : \Omega \to \mathbb{R}$, then X is a random variable if X is $\mathbb{P}$-measurable in the sense that $X^{-1}((\alpha, \infty))$ is measurable for all $\alpha \in \mathbb{R}$.*

We, however, think about them quite differently and use different notation, so let us rethink them a bit.

The elements $\omega \in \Omega$ are all the possible outcomes in an event. A random variable $X : \Omega \to \mathbb{R}$ is an "observable" – it is a quantitative reading of the event. In this sense, X being a well-defined function means we can quantify the outcome of the event and measurability of X with respect to $(\Omega, \mathcal{F}, \mathbb{P})$ means that we can understand the likelihood of X's value given a reasonable collection of events.

Example 5.2.2. *Let $\Omega = \{0,1\}^n$ be the outcome of n (independent) flips of a coin, where 1 represents heads and 0 represents tails. Here, $(1, 0, 1, 1, \ldots, 1)$ represents heads appearing on the first, third, fourth, etc. flips with tails appearing on the second, etc. flips.*

As discussed in Example 4.1.1.(iv), if the coin is fair, $(\Omega, \mathcal{P}^\Omega, \mathcal{P}_{1/2})$ is a probability space with

$$\mathbb{P}_{1/2}(\{\omega\}) = 2^{-n} \qquad \text{for any } \omega \in \Omega.$$

We can then extend $\mathbb{P}_{1/2}$ to any set A in a straightforward way because the $\{0,1\}^n$ is made up of finitely many elements. Indeed:

$$\mathbb{P}_{1/2}(A) = \sum_{\omega \in A} \mathbb{P}(\{\omega\}). \tag{5.2.1}$$

We can make the coin biased by fixing $p \in (0,1)$ that is the probability of taking the value 1 (heads). Then $1 - p$ is the probability of the value 0 (tails) on each flip. In this case, we have

$$\mathbb{P}_p(\{\omega\}) = p^{\#\text{ of heads}}(1-p)^{\#\text{ of tails}} = \prod_{i=1}^n p^{\omega_i}(1-p)^{1-\omega_i}.$$

We can then extend this to a measure on $\{0,1\}^n$ *as in* (5.2.1).

For the rest of this example, we fix $p \in (0,1)$ *and drop the* p *subscript in* $\mathbb{P}_p$. *Define the random variable* $X_n : \Omega \to \mathbb{R}$ *as*

$$X_n(\omega) = \sum_{i=1}^{n} \omega_i = \text{ the number of heads in a string.}$$

Here we use the notation $\omega = (\omega_1, \omega_2, \ldots, \omega_n)$. *For example, when* $n = 2$,

$$X_2(0,0) = 0, \quad X_2(1,1) = 2, \quad X_2(1,0) = 1, \quad \text{and} \quad X_2(0,1) = 2.$$

Let us note that we call X_1 *a Bernoulli random variable with parameter* p. *We denote this by* $X_1 \sim \text{Bernoulli}(p)$.

5.2.1. **Expectation and cumulative distribution functions.** One main operation we perform with random variables is to compute their expectation, as well as the expectation of their composition with Borel functions. The mean is

$$\mathbb{E}[X] = \int X(\omega)\, d\mathbb{P}(\omega).$$

The second moment is

$$\mathbb{E}[X^2] = \int X(\omega)^2\, d\mathbb{P}(\omega),$$

and, more generally, the nth moment is

$$\mathbb{E}[X^n] = \int X(\omega)^n\, d\mathbb{P}(\omega).$$

One can take the expectation using any Borel function $f : \mathbb{R} \to \mathbb{R}$:

$$\mathbb{E}[f(X)] = \int f(X(\omega))\, d\mathbb{P}(\omega). \tag{5.2.2}$$

As above, we refer to quantities such as this as the *statistics* of X. Note that another important statistic is the variance $\mathbb{E}[(X - \mathbb{E}[X])^2]$. It measures the "spread" of a random variable.

Exercise 5.2.1. *Show that the variance of a random variable* X *is given by* $\mathbb{E}[X^2] - \mathbb{E}[X]^2$ *if it is finite.*

Example 5.2.3. *(i) Returning to our coin flipping example from above with the choice* $p = 1/2$, *we have*

$$\begin{aligned}
\mathbb{E}[X_n] &= \sum_{\omega \in \Omega_n} X_n(\omega)\mathbb{P}(\{\omega\}) = 2^{-n} \sum_{\omega \in \Omega_n} X_n(\omega) = 2^{-n} \sum_{k=0}^{n} \sum_{\omega \in A_k} k \\
&= 2^{-n} \sum_{k=0}^{n} \binom{n}{k} k = 2^{-n} \sum_{k=1}^{n} \binom{n}{k} k = 2^{-n} \sum_{k=1}^{n} \frac{n \cdot (n-1)!}{k \cdot (k-1)!((n-1)-(k-1))!} k \\
&= 2^{-n} \sum_{k=1}^{n} n \binom{n-1}{k-1} = n 2^{-n} \sum_{k=1}^{n} \binom{n-1}{k-1} = n 2^{-n} \sum_{k=0}^{n-1} \binom{n-1}{k} = n 2^{-n} 2^{n-1} = \frac{n}{2},
\end{aligned}$$

where

$$A_k = \{\omega : X_n(\omega) = k\} \quad \text{which has } \binom{n}{k} = \frac{n!}{k!(n-k)!} \text{ elements.}$$

Here we used that the sum[36] *over* k *of* $\binom{m}{k} = 2^m$.

(ii) A uniform random variable X *on an interval* $[a,b]$ *(denoted* $X \sim \text{Unif}(a,b)$*) is one whose probability of being any number in* $[a,b]$ *is "equally likely." This, of course, does not make sense as there are infinitely many values in* $[a,b]$*, so the probability of any one particular value occurring is zero. Let us write this down rigorously. We take the probability space* $(\mathbb{R}, \mathcal{L}, \mathbb{P})$ *where*

$$\mathbb{P}(A) = \frac{1}{b-a} m(A \cap [a,b]) \qquad \text{for all } A \in \mathcal{L},$$

and the random variable $X : \mathbb{R} \to \mathbb{R}$ *by* $X(x) = x$*. One can easily check that*

$$\mathbb{P}(X \in [\alpha, \beta]) = \frac{\beta - \alpha}{b-a}, \tag{5.2.3}$$

for any $a \le \alpha \le \beta \le b$*. Since*

$$d\mu = \mathbb{1}_{[a,b]}\, dm,$$

we can easily compute the expected value of X*:*

$$\mathbb{E}[X] = \int x\, d\mu = \frac{1}{b-a} \int x \mathbb{1}_{[a,b]}(x)\, dm = \frac{b+a}{2}.$$

This makes sense since it is the "average value" in $[a,b]$*.*

At this point, we see that there are many "different" random variables $\sim \text{Unif}(a,b)$*. Indeed, let us present two more.*

- *Define the probability space* $\big([a,b], \mathcal{L}_{[a,b]}, \mathbb{P}_Y\big)$*, where*

  $$\mathcal{L}_{[a,b]} = \{A \in \mathcal{L} : A \subset [a,b]\} \quad \text{and} \quad \mathbb{P}_Y(A) = \frac{1}{b-a} m(A) \quad \text{for all } A \in \mathcal{L}_{[a,b]}.$$

  Then we define $Y : [a,b] \to [a,b]$ *as* $Y(x) = x$*. It is easy to see that* Y *is measurable and* (5.2.4) *holds for* Y *as well; that is,*

  $$\mathbb{P}_Y(Y \in [\alpha, \beta]) = \frac{\beta - \alpha}{b-a}, \tag{5.2.4}$$

  for any $a \le \alpha \le \beta \le b$*. Hence,* $Y \sim \text{Unif}(a,b)$*, as well.*
- *Define the probability space* $\big([2a, 2b], \mathcal{L}_{[2a,2b]}, \mathbb{P}_Z\big)$*, where*

  $$\mathcal{L}_{[2a,2b]} = \{A \in \mathcal{L} : A \subset [2a,2b]\} \quad \text{and} \quad \mathbb{P}_Z(A) = \frac{1}{2(b-a)} m(A) \quad \text{for all } A \in \mathcal{L}_{[2a,2b]}.$$

  Then we define $Z : [2a,2b] \to \mathbb{R}$ *as* $Z(x) = x/2$*. It is easy to see that* Z *is measurable and* (5.2.4) *holds for* Z *as well; that is,*

  $$\mathbb{P}_Z(Z \in [\alpha, \beta]) = \mathbb{P}_Z([2\alpha, 2\beta]) = \frac{2\beta - 2\alpha}{2(b-a)} = \frac{\beta - \alpha}{b-a},$$

  for any $a \le \alpha \le \beta \le b$*. Hence,* $Z \sim \text{Unif}(a,b)$*, as well.*

[36] One can show this by using the binomial expansion of $(1+1)^n$. Otherwise, one can also prove it by induction on n.

We will see below that one can tell X, Y, and Z are the "same" because they have the same "c.d.f." See Example 5.2.5.

(iii) An exponential random variable $X \sim \text{Exp}(\lambda)$ with rate $\lambda > 0$ is one where

$$\mathbb{P}(X \leq t) = \begin{cases} 0 & \text{if } t \leq 0, \\ 1 - e^{-\lambda t} & \text{if } t > 0. \end{cases}$$

A typical application of exponential random variables is to measure whether an event, which occurs at rate λ, has taken place before some time t. For example, the event could be cell-division – it happens at a "random" time that gets "much more likely" as time goes.

One can rigorously construct this by defining the probability space $(\mathbb{R}, \mathcal{L}, \mu_f)$, where $d\mu_f = f\, d\mu$ with

$$f(x) = \lambda e^{-\lambda x} \mathbb{1}_{(0,\infty)}(x),$$

and letting $X(x) = x$. We then see that

$$\mathbb{E}[X] = \int x\, d\mu_f = \int x f(x)\, dm = \lambda \int_0^\infty x e^{-\lambda x}\, dm = \frac{1}{\lambda}.$$

Let us note that in this and the previous example, the random variable is the identity on $\mathbb{R}$. This is not a coincidence, and it is explored further in (5.2.5).

In practice, the definition of a probability space can be complicated or inconvenient for computations. Another approach is related to the theory we developed above in Section 5.1.2-Section 5.1.3. Let us begin by defining the nondecreasing, càdlàg that will be the basis for this.

Definition 5.2.4. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and X be a random variable on it. The cumulative distribution function (cf. Definition 5.1.5) of X is*

$$F_X(t) = \mathbb{P}(\{\omega : X(\omega) \leq t\}) = \mathbb{P}(\{X \leq t\}).$$

Notice the introduction of notation in the last equality above. In probability, we often suppress the ω's.

Before discussing exactly how we connect this to computations involving the expectation of $f(X)$, let us compute a few simple examples. We also show that, by computing the Stieltjes measure associated to each *c.d.f.*, one obtains another probability space related to the random variable.

Example 5.2.5. *The examples here are defined in Example 5.2.3 with the same numbering.*

(i) Let us consider the coin flipping example with $n = 1$. We see that its cumulative distribution function is

$$F_X(t) = \mathbb{P}(X \leq t) = \begin{cases} 0 & \text{if } t < 0, \\ 1 - p & \text{if } t \in [0, 1), \\ 1 & \text{if } t \geq 1. \end{cases}$$

Let μ_X be the Stieltjes measure generated from F_X. As in Example 5.1.3.((ii)), *we see that*

$$\mu_X = (1 - p)\delta_0 + p\delta_1.$$

It follows that $(\mathbb{R}, \mathcal{F}_X, \mu_X)$ is a probability space, where $\mathcal{F}_X$ is the σ-algebra generated by F_X.

We see that we can then easily obtain $Y \sim \text{Bernoulli}(p)$ using the space $(\mathbb{R}, \mathcal{F}_X, \mu_X)$. Indeed, let $Y : \mathbb{R} \to \mathbb{R}$ be defined by $Y(x) = x$. One can check that $Y \sim \text{Bernoulli}(p)$.

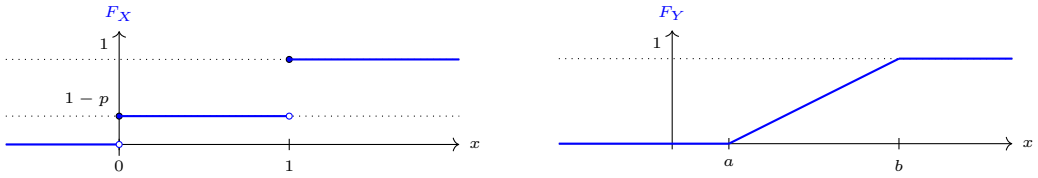


Figure 7: The c.d.f.'s of, respectively, random variables $X \sim \text{Bernoulli}(p)$ *and* $Y \sim \text{Unif}(a,b)$.

(ii) One sees that the c.d.f. is

$$F_X(t) = \begin{cases} 0 & \text{if } t < a, \\ \frac{t-a}{b-a} & \text{if } a \le t \le b, \\ 1 & \text{if } t \ge b. \end{cases}$$

(iii) An exponential random variable $X \sim \text{Exp}(\lambda)$ *with rate* $\lambda > 0$ *is one where*

$$F_X(t) = \mathbb{P}(X \le t) = \begin{cases} 0 & \text{if } t \le 0, \\ 1 - e^{-\lambda t} & \text{if } t > 0. \end{cases}$$

A typical application of exponential random variables is to measure whether an event has occurred before some time t*. For example, the event could be cell-division – it happens at a "random" time that gets "much more likely" as time goes.*

We notice immediately that F_X is a nondecreasing, càdlàg function on $\mathbb{R}$. Indeed, we use Lemma 4.5.2 (see also the computation (5.1.2)) to obtain the "càd" part, while "làg" follows form its monotonicity). It, thus, generates a measure space $(\mathbb{R}, \mathcal{F}_X, \mu_X)$. Of course, since

$$\mu_X(\mathbb{R}) = F_X(+\infty) - F_X(-\infty) = \mathbb{P}(\{X \le +\infty\}) - \mathbb{P}(\{X \le -\infty\}) = 1,$$

then $(\mathbb{R}, \mathcal{F}_X, \mu_X)$ is, itself, a probability space.

What we see is that the c.d.f. of X will hold all of the information about X. Indeed, (5.2.2) can be computed with respect to the Stieltjes measure μ_X defined purely by F_X. This is the *Law of the unconscious statistician*: if X is a random variable and g is a Borel function, then

$$\mathbb{E}[g(X)] = \int g(x)\, d\mu_{F_X} = \int g(x)\, dF_X. \tag{5.2.5}$$

Here we introduced the somewhat standard notation dF_X for the integral with respect to the measure μ_{F_X} generated by the c.d.f. of X. In general, we avoid this notation and use μ_{F_X} instead.

This law is so-named because it is tempting to think of the above as the definition of the expected value; however, it is not! Indeed,

$$\mathbb{E}[g(X)] = \int_\Omega g(X(\omega))\, d\mathbb{P}(\omega).$$

So... why is (5.2.5) true? Let us look at the simple case where $g = \mathbb{1}_{(a,b]}$. Then

$$\begin{aligned} \mathbb{E}[g(X)] &= \mathbb{E}[\mathbb{1}_{\{X \in (a,b]\}}] = \mathbb{P}\left(\{X \in (a,b]\}\right) \\ &= \mathbb{P}\left(\{X \in (-\infty, b]\}\right) - \mathbb{P}\left(\{X \in (-\infty, a]\}\right) = F_X(b) - F_X(a). \end{aligned}$$

It may not be surprising that one can then "perform the usual routine" to get that this holds for any Borel measurable function (i.e., do the above for simple functions and then use simple functions to approximate Borel functions). This is an arduous process so we omit it here. See Appendix D for the full explanation.

Remark 5.2.6. *Suppose we have a nondecreasing càdlàg function F. If we wish to construct a random variable X that has a particular c.d.f. $F_X = F$, we can do so simply letting our probability space be $(\mathbb{R}, \mathcal{F}_F, \mu_{F_X})$ and $X(x) := x$ for all $x \in \mathbb{R}$. One can easily check that this is a probability space and $F_X = F$.*

5.2.2. **Classifying random variables.** It is not hard to show that F_X, being monotonic, can be decomposed as

$$F_X = F_X^{(d)} + F_X^{(c)}, \tag{5.2.6}$$

where $F_X^{(c)}$ is continuous and $F_X^{(d)}$ is constant apart from a discrete collection of jumps. This decomposition is unique if we impose the condition

$$F_X^{(d)}(-\infty) = F_X^{(c)}(-\infty) = 0.$$

Here we use $f(-\infty)$ as shorthand for the limit of $f(x)$ as $x \to -\infty$. We use this (5.2.6) to classify random variables.

Definition 5.2.7. *Let X be a random variable on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We say that X is:*

- discrete *if $F_X = F_X^{(d)}$; that is, F is constant apart from a countable collection of jumps;*
- continuous *if $F_X = F_X^{(c)}$; that is, F_X is continuous;*
- absolutely continuous *if there is a nonnegative $f \in L^1_m(\mathbb{R})$ such that $d\mu_{F_X} = f\, dm$.*

The most basic case, which is covered in most undergraduate probability courses, is the discrete case. This is due to the fact that everything can be done without measure theory. On the other hand, the continuous case is more complicated. In particular, there are continuous, but absolutely continuous random variables. The interested reader should look at the Exercise 5.2.3 to see such an example and should consult measure theory textbooks about the Lebesgue decomposition theorem. In these notes, however, we will mainly avoid delving into the subtleties of this.

Exercise 5.2.2. *(i) Show that if X is discrete, there is a finite or countable set A such that $\mathbb{P}(X \in A) = 1$ and $P(X = a) > 0$ for all $a \in A$.*

(ii) Show that if $X \sim$ Bernoulli(p), then X is discrete. Find the set A from part (i).

(iii) Do the same for X_n as in Example 5.2.2.

Exercise 5.2.3. *[The Cantor function] Define a function $F : \mathbb{R} \to \mathbb{R}$ as follows. If $x \leq 0$, let $F(x) = 0$. If $x \geq 1$, let $F(x) = 1$. If $x \in (0, 1)$, write*

$$x = \sum_{i=1}^{\infty} \frac{t_i}{3^i},$$

where $t_i \in \{0,1,2\}$ and, for all N, there is $n \geq N$ such that $t_n \neq 2$. Then let

$$F(x) = \sum_{i=1}^{\infty} \frac{f(\bar{t})_i}{2^i},$$

where, if $\bar{t}$ contains no 1's, $f(\bar{t})_i = {}^{t_i}/_2$, and, if the first 1 in $\bar{t}$ is at index N, we let

$$f(\bar{t}) = ({}^{t_1}/_2, {}^{t_2}/_2, \dots, {}^{t_{N-1}}/_2, 1, 0, 0, \dots).$$

Show that F is continuous and is constant on any interval $(a,b) \subset C^c$, where C is the Cantor set of Exercise 4.2.7. Deduce that there is no such $g \in L^1_m$ such that $d\mu_F = g\, dm$.

An example of this is our coin flipping example. In a sense, this is the simplest case. It corresponds to the case where F_X is constant apart from a discrete collection of jumps.

Example 5.2.8. *(i) $X \sim \text{Bernoulli}(p)$ – Notice that F_X is discrete. Each "jump" corresponds to a new outcome becoming possible and is represented by a δ measure. Indeed, one can check directly that, defining*

$$dF_X = (1-p)\delta_0 + p\delta_1,$$

then

$$F_X(x) = \int \mathbb{1}_{(-\infty,x]} dF_X.$$

In particular, one sees that the Lebesgue-Stieltjes integral this defines is

$$\int g\, dF_X = (1-p)g(0) + pg(1).$$

What does this give us as an expected value?

$$\int x\, dF_X = (1-p)\cdot 0 + p \cdot 1 = p.$$

(ii) $X \sim \text{Unif}(a,b)$ – This is piecewise differentiable and is continuous, so we find

$$dF_X = \frac{1}{|b-a|}\mathbb{1}_{[a,b]}.$$

Hence

$$\int g\, dF_X = \frac{1}{|b-a|}\int_{[a,b]} g\, dm.$$

What does this give us as an expected value?

$$\int x\, dF_X = \frac{1}{|b-a|}\int_{[a,b]} x\, dm = \frac{1}{|b-a|}\left(\frac{b^2}{2} - \frac{a^2}{2}\right) = \frac{b+a}{2}.$$

(iii) $X \sim \text{Exp}(\lambda)$ – Just as in the previous example, we can simply differentiate to find

$$dF_X = \lambda e^{-\lambda x}\mathbb{1}_{[0,\infty)}.$$

Hence

$$\int g\, dF_X = \lambda \int_0^{\infty} g e^{-\lambda x}\, dm.$$

What does this give us as an expected value?

$$\int x\, dF_X = \lambda \int_0^{\infty} x e^{-\lambda x}\, dm = \frac{1}{\lambda}\int_0^{\infty} (\lambda x) e^{-\lambda x}\, d(\lambda x) = \frac{1}{\lambda}.$$

5.2.3. **Convergence results and inequalities for random variables.** The usual results (in this context) hold:

Lemma 5.2.9 (Fatou's Lemma)**.** *Suppose that X_n is a sequence of nonnegative random variables such that $X_n \to X$ almost surely*[37]*, then*

$$\mathbb{E}[X] \leq \liminf_{n\to\infty} \mathbb{E}[X_n].$$

Theorem 5.2.10 (Monotone Convergence Theorem)**.** *Suppose that X_n is a sequence of nonnegative random variables such that $X_n \to X$ almost surely and $X_1 \leq X_2 \leq X_3 \leq \cdots$, then*

$$\mathbb{E}[X] = \lim_{n\to\infty} \mathbb{E}[X_n].$$

Theorem 5.2.11 (Dominated Convergence Theorem)**.** *Suppose that X_n is a sequence of random variables and Y be a random variable such that $X_n \to X$ almost surely, $|X_n| \leq Y$ almost surely, and $\mathbb{E}[|Y|] < \infty$. Then*

$$\mathbb{E}[X] = \lim_{n\to\infty} \mathbb{E}[X_n].$$

Theorem 5.2.12 (Hölder's inequality)**.** *Suppose that X and Y are random variables and $p, q \in (1, \infty)$ are conjugate exponents. Then*

$$\mathbb{E}[XY] \leq \left(\mathbb{E}[|X|^p]\right)^{\frac{1}{p}} \left(\mathbb{E}[|Y|^q]\right)^{\frac{1}{q}}.$$

Also

$$\mathbb{E}[XY] \leq \mathbb{E}[|X|] \operatorname{ess\,sup}_\Omega |Y|.$$

Theorem 5.2.13 (Jensen's inequality)**.** *Suppose that X is a random variables and $\varphi : \mathbb{R} \to \mathbb{R}$ is a convex function. Then*

$$\varphi\left(\mathbb{E}[X]\right) \leq \mathbb{E}[\varphi(X)].$$

Theorem 5.2.14 (Markov's inequality)**.** *Suppose that X is a random variables and $\lambda > 0$. Then*

$$\mathbb{P}(X \geq \lambda) \leq \frac{\mathbb{E}[|X|]}{\lambda}.$$

Let us comment mostly on Markov's inequality. This is used to understand "tail probability" – how often a random variable takes a large value. Additionally, it can be used with a function: take an increasing function φ and apply Markov's inequality with $\varphi(X)$ to deduce the more general inequality

$$\mathbb{P}(X \geq \varphi^{-1}(\lambda)) = \mathbb{P}(\varphi(X) \geq \lambda) \leq \frac{\mathbb{E}[|\varphi(X)|]}{\lambda}.$$

Example 5.2.15. *Let us consider applying this to $X \sim \mathrm{Norm}(0, 1)$. Note a "normal random variable with mean μ and variance σ^2," denoted by $\mathrm{Norm}(\mu, \sigma^2)$ is one whose c.d.f. is given by the density*

$$dF_X = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\, dm.$$

Then, by Markov's inequality, we get for $\lambda \gg 1$

$$\mathbb{P}(X \geq \lambda) \leq \frac{\mathbb{E}[|X|]}{\lambda}.$$

[37] This is the probabilistic equivalent of almost everywhere: there is a set A such that $\mathbb{P}(A) = 1$ and, for all $\omega \in A$, $X_n(\omega) \to X(\omega)$.

Notice that

$$\mathbb{E}[|X|] = \int \frac{|x|}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\, dm = 2\int_0^\infty \frac{x}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\, dm = \frac{2}{\sqrt{2\pi}} = \sqrt{\frac{2}{\pi}}.$$

Hence

$$\mathbb{P}(X \geq \lambda) \leq \frac{\sqrt{2}}{\sqrt{\pi}\lambda}.$$

This is not great! The density for X decays like $e^{-x^2/2}$, which is much faster! We should expect a better estimate on the tail probability. Let's try something better:

$$\mathbb{P}(X \geq \lambda) = \mathbb{P}(e^X \geq e^\lambda) \leq \frac{\mathbb{E}[e^X]}{e^\lambda}.$$

Since[38]

$$\mathbb{E}[e^X] = \int \frac{e^x}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\, dm = e^{\frac{1}{2}}.$$

we deduce that

$$\mathbb{P}(X \geq \lambda) \leq \frac{e^{1/2}}{e^\lambda}.$$

This is much better!

Exercise 5.2.4. *Get a better estimate by using e^{tx} for a different choice t.*

Notice that this is especially helpful for estimating "rare" events like when $\lambda \gg 1$.

5.3. **Joint distributions.** Often, we have multiple random variables on the *same probability space*. In this case, we can obtain their joint distribution.

Definition 5.3.1. *Let X, Y be random variables on $(\Omega, \mathcal{F}, \mathbb{P})$. Their* joint distribution $F_{X,Y} : \mathbb{R} \times \mathbb{R} \to [0,1]$ *is defined by*

$$F_{X,Y}(x,y) = \mathbb{P}\{\omega : X(\omega) \leq x, Y(\omega) \leq y\} = \mathbb{P}\{X \leq x, Y \leq y\}.$$

This can easily be extended analogously to any finite number of random variables: if $X_1, \ldots, X_n$ are random variables and $t_1, \ldots, t_n \in \mathbb{R}$, then the joint distribution $F_{X_1,\ldots,X_n} : \mathbb{R} \times \cdots \times \mathbb{R} \to [0,1]$ *is defined as*

$$F_{X_1,\ldots,X_n}(t_1, \ldots, t_n) = \mathbb{P}(X_1 \leq t_1, \ldots, X_n \leq t_n).$$

We note that the joint distribution encodes what "type" of random variables X and Y are (e.g., uniform, normal, etc) and how "related" they are.

Example 5.3.2. *(i) Independent coins – Let $(\Omega, \mathcal{F}, \mathbb{P})$ be the probability space of two coin flips –see Example 4.4.9 with $n = 2$. Here $\Omega = \{0,1\}^2$, $\mathcal{F} = \mathcal{P}^\Omega$, and $\mathbb{P}(\omega) = 1/4$ for all $\omega \in \Omega$.*

Define two random variables X and Y by

$$X(\omega) = a_1 \quad \text{and} \quad Y(\omega) = a_2,$$

for any $\omega = (a_1, a_2) \in \{0,1\}^2$. In other words, X gives 1 if the first flip is heads and Y gives 1 if the second flip is heads. Note that these do not interact – having knowledge about the value of X does not impart any knowledge about the value of Y.

[38] This computation can be done by completing the square... try it!

We now compute the joint distribution:

$$F_{X,Y}(t,s) = \mathbb{P}\{X \le t, Y \le s\} = \begin{cases} 0 & \text{if } t \text{ or } s < 0, \\ 1/4 & \text{if } t, s \in [0,1), \\ 1/2 & \text{if } t \in [0,1) \text{ and } s \ge 1, \\ 1/2 & \text{if } s \in [0,1) \text{ and } t \ge 1, \\ 1 & \text{if } t, s \ge 1. \end{cases}$$

On the other hand, notice that

$$F_X(t)F_Y(s) = \mathbb{P}\{X \le t\}\mathbb{P}\{Y \le s\} = \begin{cases} 0 & \text{if } t \text{ or } s < 0, \\ 1/4 & \text{if } t, s \in [0,1), \\ 1/2 & \text{if } t \in [0,1) \text{ and } s \ge 1, \\ 1/2 & \text{if } s \in [0,1) \text{ and } t \ge 1, \\ 1 & \text{if } t, s \ge 1. \end{cases}$$

In other words, $F_{X,Y}(t,s) = F_X(t)F_Y(s)$. In fact, this is the hallmark of independent random variables. There are many equivalent definitions of "independence," one of which is the following:

Definition 5.3.3. *Two random variables X and Y on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ are independent if $F_{X,Y}(x,y) = F_X(x)F_Y(y)$ for all $x, y \in \mathbb{R}$.*

This can easily be extended to finite sets of random variables analogously: $X_1, \ldots, X_n$ are independent if, for all $t_1, \ldots, t_n \in \mathbb{R}$,

$$F_{X_1,\ldots,X_n}(t_1, \ldots, t_n) = F_{X_1}(t_1) \cdots F_{X_n}(t_n).$$

Exercise 5.3.1. *Show that the following definition is equivalent: $\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B)$ for all $X, Y \in \mathcal{F}$.*

Exercise 5.3.2. *If X_1, X_2, and X_3 are pairwise independent, is it true that X_1 and $X_2 + X_3$ are independent?*

Exercise 5.3.3. *If X and Y are independent and integrable, then $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$.*

(ii) Now define $X(\omega) = Y(\omega) = a_1$. In other words, X and Y are both 1 if the first flip is heads and are 0 if the first flip is tails. The second flip is ignored. Clearly, X and Y are not "independent" (in the usual non-mathematical sense of the word). We can compute that

$$F_{X,Y}(t,s) = \begin{cases} 0 & \text{if } \min\{t,s\} < 0, \\ 1/2 & \text{if } \min\{t,s\} \in [0,1), \\ 1 & \text{if } \min\{t,s\} \ge 1. \end{cases}$$

In other words,

$$F_{X,Y}(t,s) = \min\{F_X(t), F_Y(s)\}.$$

Actually, this is more apparent by just realizing that if $X \le t$ and $Y \le s$ then $X \le \min\{t,s\}$. Hence,

$$\begin{aligned} \mathbb{P}\{X \le t, Y \le s\} &= \mathbb{P}\{X \le \min\{t,s\}\} = F_X(\min\{t,s\}) \\ &= \min\{F_X(t), F_X(s)\} = \min\{F_X(t), F_Y(s)\}. \end{aligned}$$

The second-to-last equality follows from the fact that F_X is an increasing function, while the last equality follows from the fact that $X = Y$.

Let us take two random variables, X and Y, on the same probability space. Assume that they their joint distribution is given by a density $f_{X,Y}$:

$$\mathbb{P}((X,Y) \in B) = \int_B f_{X,Y}\, dm.$$

We can recover the density for each random variable by "forgetting" the other:

$$\mathbb{P}(X \in A) = \mathbb{P}((X,Y) \in A \times \mathbb{R}) = \int_{A\times\mathbb{R}} f_{X,Y}(x,y)\, dm(x,y).$$

After using Tonelli's theorem, this is precisely the statement that

$$f_X(x) = \int f_{X,Y}(x,y)\, dm(y). \tag{5.3.1}$$

A similar computation shows that

$$f_Y(y) = \int f_{X,Y}(x,y)\, dm(x).$$

Exercise 5.3.4. *Suppose that X and Y are absolutely continuous random variables. Is (X,Y) absolutely continuous? In other words, is there a density $f_{X,Y} : \mathbb{R}^2 \to \mathbb{R}$ such that, for all $A, B \in \mathcal{B}$,*

$$\mathbb{P}(X \in A, Y \in B) = \int_{A\times B} f_{X,Y}(x,y)\, d(m\times m)?$$

Example 5.3.4. *A general Gaussian random variable in n-dimensions is given by the following: let Σ be a symmetric, positive definite[39] $n \times n$ matrix and $\mu \in \mathbb{R}^n$, then define $\bar{X}$ to be a random variable with density*

$$f_{\bar{X}}(x) = \frac{1}{(2\pi)^{n/2}\sqrt{\det \Sigma}} e^{-\frac{1}{2}(x-\mu)\cdot(\Sigma^{-1}(x-\mu))}.$$

One can check that $\mathbb{E}[\bar{X}] = \mu$. We call Σ the "covariance matrix" of the "multivariate normal random variable" $\bar{X}$.

When $n = 2$, Σ is diagonal

$$\Sigma = \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix},$$

and $\mu = 0$, we get

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma_1\sigma_2} e^{-\frac{x^2}{2\sigma_1^2} - \frac{y^2}{2\sigma_2^2}}.$$

Here we are writing $\bar{X} = (X,Y)$. It is easy to check that

$$f_X(x) = \int f_{X,Y}(x,y)\, dm(y) = \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{x^2}{2\sigma_1^2}},$$

and

$$f_Y(y) = \int f_{X,Y}(x,y)\, dm(x) = \frac{1}{\sqrt{2\pi\sigma_2^2}} e^{-\frac{y^2}{2\sigma_2^2}},$$

We observe that $f_{X,Y}(x,y) = f_X(x) f_Y(y)$, which belies the independence of the two Guassian random variables X, Y. In fact, this independence is really due to the fact that we chose a diagonal Σ.

[39] In other words, $\Sigma^T = \Sigma$ and $x \cdot (\Sigma x) > 0$ for all $x \neq 0$. Equivalently, all of Σ's eigenvalues are positive.

Note that above we were considering a collection of random variables on a single probability space and then "projecting" down to find the distribution of each random variable. We can always go in the other direction: if we have two random variables X and Y on the probability spaces $(\Omega, \mathcal{F}, \mathbb{P})$ and $(\Sigma, \mathcal{G}, \mathbb{Q})$, respectively, we can consider the product space $(\Omega \times \Sigma, \mathcal{F} \times \mathcal{G}, \mathbb{P} \times \mathbb{Q})$ on which X and Y are (still) random variables with a joint distribution. In this case, X and Y will be independent.

Exercise 5.3.5. *If Σ has row vectors $[2, 1]$ and $[1, 2]$, what are f_X and f_Y? Are X and Y independent?*

5.4. **Functions of random variables.** Given any Borel measurable $g : \mathbb{R} \to \mathbb{R}$ and random variable X on $(\Omega, \mathcal{F}, \mathbb{P})$, we can compose them to find a new random variable on the same probability space

$$Y = g(X) \qquad (\text{that is, } Y(\omega) = g \circ X(\omega) = g(X(\omega)) \text{ for all } \omega \in \Omega).$$

One can usually compute the c.d.f. for Y from that of X.

Example 5.4.1. *(i) Let X be any random variable, and define*

$$Y = X^2.$$

This is composition with the continuous function $g : \mathbb{R} \to \mathbb{R}$ defined by $g(x) = x^2$.

Let us first consider the case where F_X is continuous (instead of just càdlàg). Then, if $t < 0$, we have, since $X^2 \geq 0$,

$$F_Y(t) = \mathbb{P}(Y \leq t) = \mathbb{P}(X^2 \leq t) = 0.$$

On the other hand, if $t \geq 0$, we have

$$\begin{aligned} F_Y(t) &= \mathbb{P}(Y \leq t) = \mathbb{P}(X^2 \leq t) = \mathbb{P}(-\sqrt{t} \leq X \leq \sqrt{t}) \\ &= \mathbb{P}(X \leq \sqrt{t}) - \mathbb{P}(X < -\sqrt{t}) = F_X(\sqrt{t}) - F_X(-\sqrt{t}). \end{aligned}$$

Here we used the continuity of F_X to get that

$$\mathbb{P}(X < -\sqrt{t}) = \lim_{z \nearrow -\sqrt{t}} \mathbb{P}(X < z) = \lim_{z \nearrow -\sqrt{t}} F_X(z) = F_X(-\sqrt{t}). \tag{5.4.1}$$

If F_X is only càdlàg, then the last step does not work, but we can still all but the last step in (5.4.1) to conclude that

$$F_Y(t) = F_X(\sqrt{t}) - \lim_{z \nearrow -\sqrt{t}} F_X(z).$$

If F_X is given by a density f_X (i.e., $dF_X = f_X\, dm$), then, for all $t \geq 0$,

$$f_Y(t) := dF_Y = dF_X(\sqrt{t}) - dF_X(-\sqrt{t}) = \frac{1}{2\sqrt{t}}(f_X(\sqrt{t}) + f_X(-\sqrt{t})).$$

Exercise 5.4.1. *In the case where $X \sim \text{Unif}(-1, 1)$, work out what f_Y is.*

(ii) Inverse transform sampling. Let $Y = F_X(X)$. A helpful way to process this is to remember that we are really saying $Y(\omega) = F_X(X(\omega))$ for $\omega \in \Omega$. Indeed, the subscript "X" in F_X just says that F_X depends on the statistics of X (what kind of values it takes and how often it takes them) while the input of F_X is $X(\omega)$, the outcome/value of an event.

One thing that we see immediately is that Y must take values only in $[0,1]$. Indeed, $F_X(t) = \mathbb{P}(X \leq t) \in [0,1]$. Hence,

$$F_Y(t) = \begin{cases} 0 & \text{if } t < 0, \\ 1 & \text{if } t \geq 1. \end{cases}$$

When $t \in (0,1)$, we perform the following computation:

$$F_Y(t) = \mathbb{P}(Y \leq t) = \mathbb{P}(F_X(X) \leq t) = \mathbb{P}(X \leq F_X^{-1}(t)) = F_X(F_X^{-1}(t)) = t.$$

The above is justified if F_X is strictly increasing.

Exercise 5.4.2. *ADD EXERCISE TO JUSTIFY THIS IF F_X IS NOT STRICTLY INCREASING.*

This is exactly the distribution function of $\mathrm{Unif}(0,1)$*! Hence, $Y \sim \mathrm{Unif}(0,1)$.*

This is precisely why this is useful: if you develop a good sampling algorithm for $U \sim \mathrm{Unif}(0,1)$ and you know the distribution F_X of another random variable X, then you can sample X via the formula:

$$X \sim F_X^{-1}(U).$$

5.4.1. **Moment generating function.** We defined above the moments:

$$\mathbb{E}[X^n] \qquad \text{is the } n\text{th moment of } X.$$

The first moment is the mean ('average value') of the random variable, while the second moment is related to variance, which measures the 'spread' of a random variable:

$$\begin{aligned} \mathrm{Var}(X) &= \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \mathbb{E}\left[X^2 - 2X\mathbb{E}[X] + \mathbb{E}[X]^2\right] \\ &= \mathbb{E}\left[X^2\right] - 2\mathbb{E}[X]^2 + \mathbb{E}[X]^2 = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2. \end{aligned}$$

Additionally, moments can be useful when used with Markov's inequality:

$$\mathbb{P}(X \geq \lambda) \leq \frac{\mathbb{E}[X^{2n}]}{\lambda^{2n}}.$$

In general, moments hold a lot of information about a random variable. Let us then wrap them into one quantity.

Definition 5.4.2. *The moment generating function of a random variable is*

$$M_X(t) = \mathbb{E}[e^{tX}].$$

Notice that this is only finite if X is "not too large".

Where does this name come from? Notice that

$$M_X(t) = \mathbb{E}\left[\sum_{n=0}^{\infty} \frac{(tX)^n}{n!}\right] = \sum_{n=0}^{\infty} \frac{t^n}{n!}\mathbb{E}[X^n].$$

Thus, the moments of X are the coefficients in the Taylor series (around 0) of $M_X(t)$. As a consequence, knowing the moments of X and knowing M_X are equivalent.

It may not seem believable at this moment, but it can sometimes be easier to compute the moment generating function than the moments themselves. Additionally, the moment generating function can be useful to show convergence to a particular random variable.

Let us make note of a few asides. First, the interested reader may enjoy reading about 'generating functions' in general – they show up in number theory, combinatorics, ordinary and partial differential equations, and many other fields. In other words, we have the following result:

Proposition 5.4.3. *Let X and Y be two random variables such that, for some $C > 0$, we have*

$$|\mathbb{E}[X^n]|, |\mathbb{E}[Y^n]| \leq C^n n!. \tag{5.4.2}$$

Then $M_X(t)$ and $M_Y(t)$ are well-defined on an interval $(-R, R)$ for some $R > 0$, possibly infinite.

1. *If $M_X(t) = M_Y(t)$ for all $t \in (-R, R)$, then X and Y are identically distributed; that is, $F_X = F_Y$.*
2. *For every $k \in \mathbb{N}$, $M_X^{(k)}(0) = \mathbb{E}[X^k]$.*

ADD PROOF IN AN APPENDIX AT A LATER DATE.

Remark 5.4.4. *Reading Proposition 5.4.3, one might be tempted to say that if two random variables X and Y have all of the same moments, i.e.,*

$$\mathbb{E}[X^n] = \mathbb{E}[Y^n] \qquad \text{for all } n \geq 0,$$

then $X \sim Y$. This, of course, is not true. There are counterexamples to this when (5.4.2) *is does not hold; see Exercise 5.4.4.*

To give some perspective on this, let us consider an analogous problem. Fix a smooth function $f : \mathbb{R} \to \mathbb{R}$. Recall that if there are a constants $R, C > 0$ such that

$$|f^{(n)}(x)| \leq C^n n! \qquad \text{for all } |x| \leq R, \tag{5.4.3}$$

then, for all $|x| < \min\{R, 1/C\}$,

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

Again, one might be tempted to remove the condition (5.4.3)*; however, one cannot. Let*

$$f(x) = \begin{cases} 0 & \text{if } x \leq 0, \\ e^{-1/x} & \text{if } x > 0. \end{cases}$$

It is easy to check that f is smooth and $f^{(n)}(0) = 0$ for all n. Unfortunately, for any $x > 0$,

$$\sum_{n=0}^{\infty} \frac{x^n}{n!} f^{(n)}(0) = \sum_{n=0}^{\infty} \frac{x^n}{n!} \cdot 0 = 0 \neq f(x).$$

Exercise 5.4.3. *Suppose that X_1, X_2, and X_3 are independent random variables on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and satisfy* (5.4.2)*. Show that if $F_{X_1+X_2} = F_{X_1+X_3}$ (in other words, $X_1 + X_2$ and $X_1 + X_3$ are* equal in distribution*), then $F_{X_1} = F_{X_2}$ (i.e., X_1 and X_2 are equal in distribution.*

Exercise 5.4.4. *Show that random variables X_1 and X_2, each with p.d.f.*

$$f_1(x) = \mathbb{1}_{\mathbb{R}_+}(x)\frac{1}{x\sqrt{2\pi}}e^{-\frac{(\log x)^2}{2}} \quad \text{and} \quad f_2(x) = f_1(x)\left(1+\frac{1}{2}\sin\left(2\pi\log(x)\right)\right),$$

respectively, have the same moments. Deduce that two random variables with the same moments need not be equal.

Roughly, Proposition 5.4.3 holds because the moment generating function is the Laplace transform of the Stieltjes measure μ_X associated to X. This is particularly clear if X is an absolutely continuous random variable with density f_X, whence

$$M_X(t) = \mathbb{E}\left[e^{tX}\right] = \int e^{tx} f_X(x)\, dm(x) = \mathcal{L}(f_X)(-t), \tag{5.4.4}$$

where $\mathcal{L}$ is the Laplace transform. The Laplace transform is invertible, although we do not discuss this further.

Finally, recall that the moment generating function was very useful in getting a good estimate (via Markov's inequality) on the tail of the normal distribution. See Example 5.2.15.

Example 5.4.5. *(i) Let $X \sim \mathrm{Unif}(0,1)$. Then due to (5.4.4) and the fact that $f_X = \mathbb{1}_{[0,1]}$, we find*

$$M_X(t) = \int e^{tx}\mathbb{1}_{[0,1]}\, dm(x) = \frac{1}{t}\left[e^{tx}\right]_0^1 = \frac{e^t-1}{t}.$$

(ii) Let $X \sim \mathrm{Exp}(\lambda)$. Again using (5.4.4) and the fact that $f_X = \mathbb{1}_{\mathbb{R}_+}\lambda e^{-\lambda x}$, we find, for $t \in (0,\lambda)$,

$$M_X(t) = \int e^{tx}\mathbb{1}_{\mathbb{R}_+}\lambda e^{-\lambda x}\, dm(x) = \frac{\lambda}{\lambda - t}.$$

(iii) **Exercise 5.4.5.** *Compute $M_X(t)$ for $X \sim \mathrm{Norm}(\mu,\sigma^2)$.*

(iv) Note that the moment generating function need not be defined as random variables may not have higher moments. The "most" pathological example of this is perhaps a Cauchy random variable whose p.d.f. is

$$\frac{1}{\pi(1+x^2)}.$$

We might be tempted to say the mean of this random variable is 0 because it is even; however,

$$\mathbb{E}[X] = \int \frac{x}{\pi(1+x^2)}\, dm,$$

and the integrand on the right is not L^1. Hence X has no moments!

A typical usage of Markov's inequality to estimate the rare events $\{X \geq \mu\}$ for $\mu \gg 1$. Indeed:

$$\mathbb{P}(X \geq \mu) \leq \frac{M_X(t)}{e^{\mu t}}.$$

For example, returning to Example 5.4.5.(ii), we see that, for $X \sim \mathrm{Exp}(\lambda)$,

$$\mathbb{P}(X \geq \mu) \leq \frac{e\lambda}{(\lambda - t)}e^{-\lambda\mu}.$$

Optimizing this estimate in t leads to

$$\mathbb{P}(X \geq \mu) \leq \frac{\lambda\mu}{e}e^{-\mu\lambda}$$

This is, of course, a silly example because we can compute $\mathbb{P}(X \geq \mu)$ directly using the density.

5.4.2. **Characteristic function.**

Definition 5.4.6. *The characteristic function of a random variable X is*

$$\Phi_X(t) = \mathbb{E}[e^{itX}].$$

Before continuing on, let us make note of the fact that we have not discussed complex variables above. In these notes, we only use a few facts about them: (1) $i^2 = -1$, (2) $e^{i\theta} = \cos(\theta) + i\sin(\theta)$ for all $\theta \in \mathbb{R}$, (3) $\int if(x)\,dm = i\int f(x)\,dm$, and (4) $|a+ib| = a^2 + b^2$. We point out that

$$\frac{1}{i} = -i \qquad \text{because } i\frac{1}{i} = 1 = i(-i).$$

Notice that, unlike M_X, we do not require infinite moments to define Φ_X. Indeed,

$$|\Phi_X(t)| \leq \mathbb{E}[|e^{itX}|] = \mathbb{E}[1] = 1.$$

In fact, if X has a p.d.f. f_X, then

$$\Phi_X(t) = \int f_X(x)e^{itx}\,dx.$$

This is the Fourier transform of f_X. If you have read about the Fourier transform, then you know that it is unbelievably useful in a both pure and applied math. It should not come as a surprise, then, that Φ_X is very useful as well.

Actually, this leads to a simple and important observation: the Fourier transform is invertible. Hence, if $\Phi_{X_1} = \Phi_{X_2}$ then X_1 and X_2 must be equal in distribution. Actually, one does not need X_1 and X_2 to have densities to come to this conclusion; however, we will not get into the ins and outs of defining the Fourier transform of measures that do not correspond to densities.

Second, one might be tempted to think that $\Phi_X(t) = M_X(it)$. By that I mean, if we have a formula for M_X we can simply "plug in" it to obtain Φ_X. This is true with a caveat – M_X must be finite for this two hold. We see this in Example 5.4.8 below.

Proposition 5.4.7. *Let X be a random variable and $n \in \{0, 1, 2, \dots\}$. If $X \in L^n$ then Φ_X is n-times differentiable and*

$$\Phi_X^{(k)}(t) = \mathbb{E}[i^k X^k e^{itX}]. \tag{5.4.5}$$

If, instead, n is even and $\Phi_X \in C^n$, that is, it is n-times continuously differentiable, then $X \in L^n$.

Proof. We complete this in two parts.

Step one: integrability of X implies differentiability of Φ_X. Let us show the differentiability by inducting on $k \in \{0, \cdots, n\}$ to establish (5.4.5). We first consider the case $k = 0$. In this case, we must show that Φ_X is continuous. We use the Dominated Convergence theorem for this. Note that, for all t,

$$|e^{itX}| \leq 1 \qquad \text{and} \qquad 1 \in L^1_{\mathbb{P}}.$$

Then, we simply note that, for any t,

$$\begin{aligned}\lim_{s\to t}|\Phi_X(t) - \Phi_X(s)| &\leq \liminf_{s\to t}\mathbb{E}\left[\left|e^{itX} - e^{isX}\right|\right] \leq \liminf_{s\to t}\mathbb{E}\left[\left|e^{i(t-s)X} - 1\right|\right] \\ &\leq \mathbb{E}\left[\liminf_{s\to t}\left|e^{i(t-s)X} - 1\right|\right] = 0.\end{aligned}$$

The last equality uses the continuity of the exponential function. Let us note that the above argument actually establishes *uniform* continuity of Φ_X. This completes the proof in the case $k = 0$.

Now assume we have shown it for n and we show it for $n+1$. We focus only on establishing (5.4.5), the continuity of which follows from exactly the argument as in the case $n = 0$. Indeed,

$$\begin{aligned}
&\lim_{h\to 0} \frac{\Phi_X^{(n)}(t+h) - \Phi_X^{(n)}(t) - h\mathbb{E}[i^{n+1}X^{n+1}e^{itX}]}{h} \\
&\quad = \lim_{h\to 0} \frac{\mathbb{E}[i^n X^n \left(e^{i(t+h)X} - e^{itX}\right)] - h\mathbb{E}[i^{n+1}X^{n+1}e^{itX}]}{h} \\
&\quad = \lim_{h\to 0} \mathbb{E}\left[X^n \frac{\cos(hX) + i\sin(hX) - 1 - ihX}{h}\right] \\
&\quad = \lim_{h\to 0} \mathbb{E}\left[X^{n+1}\frac{\cos(hX)-1}{Xh}\right] + i\lim_{h\to 0}\mathbb{E}\left[X^{n+1}\frac{\sin(hX)-hX}{Xh}\right].
\end{aligned}$$

If $X \in L^{k+1}$, then the limit is zero by the dominated convergence theorem and the fact that

$$\lim_{h\to 0} X^{k+1}\frac{\cos(hX)-1}{Xh},\ \lim_{h\to 0} X^{k+1}\frac{\sin(hX)-hX}{Xh} = 0.$$

Step two: differentiability of Φ_X implies integrability of X. Now we discuss why differentiability yields integrability. Let us note that there is no content to the case $n = 0$. We show the proof for $n = 2$ first, and then discuss how it can be generalized. By Fatou's lemma, we have

$$\begin{aligned}
\mathbb{E}\left[X^2\right] &= 2\mathbb{E}\left[\lim_{h\to 0}\frac{1-\cos(hX)}{h^2}\right] \le 2\liminf_{h\to 0}\mathbb{E}\left[\frac{1-\cos(hX)}{h^2}\right] \\
&= \liminf_{h\to 0}\frac{\Phi_X(h) - 2\Phi_X(0) + \Phi_X(-h)}{h^2} = -\Phi_X''(0).
\end{aligned}$$

Hence $X \in L^2$.

The main idea of the above proof is to approximate the second derivative by a symmetric finite difference

$$\Delta_{2,h}f(x) = \frac{f(x+h) + f(x-h) - 2f(x)}{h^2}.$$

This is symmetric because $\Delta_{2,h} = \Delta_{2,-h}$. Let us show that we can do this for all even n by induction. The base case $n = 2$ is finished. In general, if we have $\Delta_{n,h}$ defined so that, for all smooth f,

$$\lim_{h\to 0}\Delta_{n,h}f(x) = f^{(n)}(x)$$

and $\Delta_{n,h} = \Delta_{n,-h}$, we define

$$\Delta_{n+2,h}f(x) := \frac{\Delta_{n,h}f(x+h) + \Delta_{n,h}f(x-h) - 2\Delta_{n,h}f(x)}{h^2}.$$

It is easy to check that, for all smooth f,

$$\lim_{h\to 0}\Delta_{n+2,h}f(x) = f^{(n+2)}(x)$$

and $\Delta_{n+2,h} = \Delta_{n+2,-h}$. One can prove by induction that

$$\Delta_{n+2,h}\Phi_X(0) = i^{n+2}\mathbb{E}\left[\frac{\sin(hX/2)^{n+2}}{(h/2)^{n+2}}\right].$$

Since this integrand is nonnegative, we may apply Fatou's lemma to obtain

$$\begin{aligned}\mathbb{E}[X^{n+2}] &= \mathbb{E}\left[\lim_{h\to 0}\frac{\sin(hX/2)^{n+2}}{(h/2)^{n+2}}\right]\\ &\leq \liminf_{h\to 0}\mathbb{E}\left[\frac{\sin(hX/2)^{n+2}}{(h/2)^{n+2}}\right]\\ &= \frac{1}{i^{n+2}}\liminf_{h\to 0}\Delta_{n+2,h}\Phi_X^{(n+2)}(0) = i^{n+2}\Phi_X^{(n+2)}(0).\end{aligned}$$

Here we also used that, since n is even, $i^{-(n+2)} = i^{n+2}$. □

Example 5.4.8. *(i) Let $X \sim \mathrm{Unif}(0,1)$. From Example 5.4.5.(i), we expect that*

$$\Phi_X(t) = M_X(it) = \frac{e^{it}-1}{it}.$$

Let us check this honestly. Indeed,

$$\begin{aligned}\Phi_X(t) &= \int \mathbb{1}_{[0,1]}e^{itx}\,dm = \int \mathbb{1}_{[0,1]}(\cos(tx)+i\sin(tx))\,dm\\ &= \int \mathbb{1}_{[0,1]}\cos(tx)\,dm + i\int \mathbb{1}_{[0,1]}\sin(tx)\,dm = \frac{\sin(t)}{t} + i\frac{1-\cos(t)}{t}\\ &= i\frac{1-\cos(t)-i\sin(t)}{t} = \frac{\cos(t)+i\sin(t)-1}{it} = \frac{e^{it}-1}{it}.\end{aligned}$$

(ii) Let $X \sim \mathrm{Exp}(\lambda)$. Again, from Example 5.4.5.(ii), we expect

$$\Phi_X(t) = M_X(it) = \frac{\lambda}{\lambda - it}.$$

Note that

$$(\lambda - it)\frac{1}{\lambda - it} = 1 = (\lambda - it)\frac{\lambda + it}{\lambda^2 + t^2}, \tag{5.4.6}$$

so that

$$\frac{\lambda}{\lambda - it} = \lambda\frac{\lambda + it}{\lambda^2 + t^2}.$$

Let us compute this:

$$\begin{aligned}\Phi_X(t) &= \int e^{itx}\mathbb{1}_{\mathbb{R}_+}\lambda e^{-\lambda x}\,dm(x) = \lambda\int_0^\infty(\cos(tx)+i\sin(tx))e^{-\lambda x}\,dm(x)\\ &= \lambda\left(\frac{t}{t^2+\lambda^2} + i\frac{\lambda}{t^2+\lambda^2}\right) = \frac{\lambda}{\lambda - it}.\end{aligned}$$

In the last equality, we used (5.4.6).

(iii) **Exercise 5.4.6.** *Compute $\Phi_X(t)$ for $X \sim \mathrm{Norm}(\mu, \sigma^2)$.*

(iv) Recall from Example 5.4.5 the Cauchy random variable X whose p.d.f. is given by

$$\frac{1}{\pi(1+x^2)}.$$

Using complex analysis that is outside the scope of this course, one can show that

$$\varphi_X(t) = e^{-|t|}.$$

Notice that Φ_X is not even differentiable at $t = 0$, which is consistent with the fact that the Cauchy random variable does not have any moments.

In the multidimensional case, $\bar{X} = (X_1, \ldots, X_n)$, we can defined the moment generating function and characteristic function analogously: for any $\bar{t} \in \mathbb{R}^n$, let

$$M_{\bar{X}}(\bar{t}) = \mathbb{E}\left[e^{\bar{t}\cdot\bar{X}}\right] \qquad \text{and} \qquad \Phi_{\bar{X}}(\bar{t}) = \mathbb{E}\left[e^{i\bar{t}\cdot\bar{X}}\right].$$

All of our results above generalize to this case.

If the coordinates are independent, then we have

$$M_{\bar{X}}(\bar{t}) = M_{X_1}(t_1)M_{X_2}(t_2)\cdots M_{X_n}(t_n) \qquad \text{and} \qquad \Phi_{\bar{X}}(\bar{t}) = \Phi_{X_1}(t_1)\Phi_{X_2}(t_2)\cdots\Phi_{X_n}(t_n). \quad (5.4.7)$$

Actually, the fact that the moment generating function "splits" in such a way is equivalent to independence.

Exercise 5.4.7. *Show this!*

5.5. **Some further aspects of independence.** We have introduced independence of random variables before independence of events, which is backwards compared to standard courses. We now fill in the gaps.

Definition 5.5.1. *Two events $A, B \in \mathcal{F}$ are* independent *if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$.*

In other words, the probability that A and B occur is simply the product of the probability that A occurs and the probability that B occurs.

Notice that, by the exercise after Definition 5.3.3, two random variables are independent if they, for any Borel sets A and B, the events $X^{-1}(A)$ and $Y^{-1}(B)$ are independent events in $(\Omega, \mathcal{F}, \mathbb{P})$.

Example 5.5.2. *(i) Let us consider our favorite example: coin flipping. Let $\Omega_n = \{0,1\}^n$ with the σ-algebra $\mathcal{P}^{\Omega_n}$ and the probability measure $\mathbb{P}_n$.*

Consider any $k \in \{1, \ldots, n-1\}$. Let $\mathcal{A} = A \times \{0,1\}^k$ and $\mathcal{B} = \{0,1\}^{n-k} \times B$ for any$A \subset \{0,1\}^{n-k}$ and $B \subset \{0,1\}^k$. Plainly, $\mathcal{A}$ is an event depending only on the first $n-k$ flips while $\mathcal{B}$ is an event depending only on the last k flips. Intuitively, we know that $\mathcal{A}$ and $\mathcal{B}$ are independent. Let us see that with the definition.

Exercise 5.5.1. *Check that $\mathcal{A} \cap \mathcal{B} = A \times B$.*

Then

$$\mathbb{P}_n(\mathcal{A} \cap \mathcal{B}) = \mathbb{P}_n(A \times B) = \mathbb{P}_{n-k}(A)\mathbb{P}_k(B) = \frac{|A|}{2^{n-k}}\frac{|B|}{2^k}.$$

Here we use $|K|$ to denote the number of elements in a set K. On the other hand,

$$\mathbb{P}_n(\mathcal{A}) = \sum_{\omega\in\mathcal{A}} 2^{-n} = \sum_{a\in A}\sum_{\tilde{\omega}\in\{0,1\}^k} 2^{-n} = \sum_{a\in A} 2^{-(n-k)} = \frac{|A|}{2^{n-k}}.$$

Similarly, $\mathbb{P}_n(\mathcal{B}) = |B|/2^k$. We conclude that

$$\mathbb{P}_n(\mathcal{A} \cap \mathcal{B}) = \mathbb{P}_n(\mathcal{A})\mathbb{P}_n(\mathcal{B}).$$

Hence, any event of the first $n-k$ flips is independent of the last k flips.

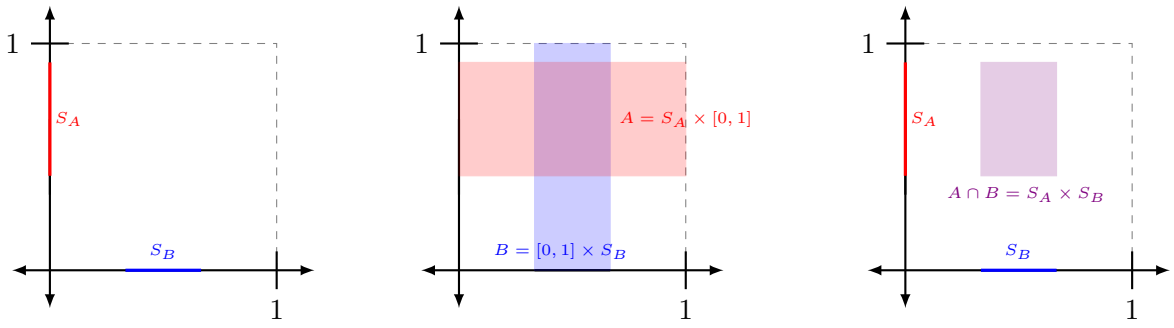
(ii) Let us look at a silly example by thinking of m as a probability measure on $[0,1]\times[0,1]$. Fix any measurable sets S_A and S_B. Take $A = [0,1]\times S_A$ and $B = S_B \times [0,1]$. It is easy to see that

$$m(A) = m([0,1]\times S_A) = m(S_A) \qquad \text{and, similarly, } m(B) = m(S_B).$$

where we are abusing notation by letting m refer to both the one- and two-dimensional Lebesgue measures. On the other hand, $A\cap B = S_B \times S_A$, so

$$m(A\cap B) = m(S_B\times S_A) = m(S_B)m(S_A),$$

since $S_B \times S_A$ is a rectangle.



Why is this a useful example? Well, A essentially only has anything "interesting" in the second coordinate and B essentially only has anything "interesting" in the first coordinate, and the first and second coordinates are orthogonal and, in a heuristic sense, orthogonality is "like" independence. If I tell you the first coordinate of a vector in $[0,1]\times[0,1]$, you have no information about the second coordinate!

Contrast that with the situation where B is replaced by $\{(x,x) : \operatorname{dist}(x, S_B) < 1/2\}$. This is, roughly, along $(1,1)$. Notice that, if tell you that a vector v satisfies $v\cdot(1,1)\approx 0$, then you know that the first coordinate of v is ≈ 0. Hence, in this case, we would not expect B to be independent of A.

5.5.1. **Independent σ-algebras.** Let us mention an interpretation of a σ-algebra as containing information. This will be especially useful for us in Section 5.8. Indeed, the events in a σ-algebra are those that we can assess the probability of. This means that we have access to the information in them (e.g., whether a coin is heads), and that we do not have information about events not in the σ-algebra (e.g., whether I am wearing a blue shirt while flipping that coin).

Given this interpretation, we should be able to define a notion of sub-σ-algebras being independent; that is, whether the information that one contains informs the other. This is what we do now.

Definition 5.5.3. *Fix a probability space $(\Omega,\mathcal{F},\mathcal{P})$. We say that two σ-algebras $\mathcal{G}_1,\mathcal{G}_2\subset\mathcal{F}$ are independent if, for every pair of random variables X_1 that is $\mathcal{G}_1$-measurable and X_2 that is $\mathcal{G}_2$-measurable, X_1 and X_2 are independent.*

Exercise 5.5.2. *Show that this is equivalent to the following: for all $A\in\mathcal{G}_1$ and $B\in\mathcal{G}_2$, the events A and B are independent.*

Example 5.5.4. *(i) Product space: Let $(\Omega,\mathcal{F},\mathbb{P})$ and $(\Sigma,\mathcal{G},\mathbb{Q})$ be probability spaces. Then take $(\Omega\times\Sigma,\mathcal{F}\times\mathcal{G},\mathbb{P}\times\mathbb{Q})$ to be the product space.*

Define

$$\mathcal{F}_1 = \{A \times \Sigma : A \in \mathcal{F}\} \qquad \text{and} \qquad \mathcal{F}_2 = \{\Omega \times A : A \in \mathcal{G}\}.$$

One can readily see that, for $A \times \Sigma \in \mathcal{F}_1$ and $\Omega \times B \in \mathcal{F}_2$,

$$\mathbb{P} \times \mathbb{Q}\left((A \times \Sigma) \cap (\Omega \times B)\right) = \mathbb{P} \times \mathbb{Q}\left(A \times B\right) = \mathbb{P}(A)\mathbb{Q}(B),$$

by definition of the product measure. See Example 5.5.2 for an explicit example of this.

To be more explicit take

$$(\Omega, \mathcal{F}, \mathbb{P}) = (\Sigma, \mathcal{G}, \mathbb{Q}) = (\{H, T\}, \mathcal{P}^{\{H,T\}}, \mathbb{P}),$$

where $\mathbb{P}(\{\omega\}) = 1/2$ for $\omega = H, T$. Then $\mathcal{F}_1$ is the "information" about the first coin flip and $\mathcal{F}_2$ is the information about the second coin flip. It is clear that they should be independent (as long as our modeling is good), which is what we see mathematically.

(ii) σ-algebras generated by independent random variables: Let X and Y be independent random variables on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Then let

$$\mathcal{G}_1 = X^{-1}(\mathcal{B}) \qquad \text{and} \qquad \mathcal{G}_2 = Y^{-1}(\mathcal{B})$$

where $\mathcal{B}$ is the set of Borel subsets of $\mathbb{R}$. It is not difficult to show that $\mathcal{G}_i$ are σ-algebras and, by the measurability of X and Y, these are subsets of $\mathcal{F}$. The independence of $\mathcal{G}_1$ and $\mathcal{G}_2$ follows then from the definition of the independence of X and Y.

5.5.2. **Independence of countably many sets, random variables, and σ-algebras.** The above concepts can easily be generalized to countable sets by requiring independence of any finite subcollection. We show this definition below for events, but it is defined analogously for random variables and σ-algebras.

Definition 5.5.5. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. A collection of events $A_1, A_2, \ldots$ are independent if, for any finite collection $n_1, \ldots, n_k$, we have*

$$\mathbb{P}\left(\bigcap_{i=1}^{k} A_{n_i}\right) = \prod_{i=1}^{k} \mathbb{P}(A_{n_i}).$$

5.6. **Convergence of random variables.** Let us look at a simple, motivating question: given independent identically distributed ("i.i.d.") random variables $X_1, \ldots, X_n$, say with mean μ and variance σ^2, what happens to the follow "average" quantities

$$\frac{X_1 + \cdots + X_n}{n} \qquad \text{and} \qquad \frac{X_1 + \cdots + X_n - \mu n}{\sqrt{n}}? \tag{5.6.1}$$

Let us mention that this is a natural question in any sort of repeated event (e.g., exam scores). It also shows up in random walks (which are, then, used to model things like population dynamics, genetic variation, heat flow, or really anything that "moves randomly").

The first quantity in (5.6.1) is clearly an average, but why do we refer to the second as an average? In a sense, the mean is a first moment average because

$$\mathbb{E}\left[\frac{X_1 + \cdots + X_n}{n}\right] = \frac{\mathbb{E}[X_1] + \cdots + \mathbb{E}[X_n]}{n} = \frac{\mu + \cdots + \mu}{n} = \mu.$$

The second quantity in (5.6.1) is a second moment average because

$$\begin{aligned}\mathbb{E}\left[\left(\frac{X_1+\cdots+X_n-\mu n}{\sqrt{n}}\right)^2\right] &= \mathbb{E}\left[\left(\sum_{i=1}^n \frac{X_i-\mu}{\sqrt{n}}\right)^2\right] \\ &= \frac{1}{n}\left(\sum_{i=1}^n \mathbb{E}\left[(X_i-\mu)^2\right] + \sum_{i\neq j}\mathbb{E}\left[(X_i-\mu)(X_j-\mu)\right]\right) \\ &= \frac{1}{n}\left(\sum_{i=1}^n \sigma^2 + \sum_{i\neq j} 0\right) = \sigma^2.\end{aligned}$$

Here we used that, due to independence, if $i \neq j$ then

$$\mathbb{E}[X_iX_j] = \mathbb{E}[X_i]\mathbb{E}[X_j] = \mu^2.$$

Let us point out that the two quantities in (5.6.1) are related in that, if we refer to $\bar{X}_n$ as the former, then $\sqrt{n}(\bar{X}_n - \mu)$ is the latter. In this sense, the latter measures the "fluctuations" of the former.

In order to evaluate these questions, we need to understand better what we mean by convergence. As with L^p functions, there are several standard notions that one might use. Here are some other notions of convergence.

5.6.1. **Convergence in distribution (weak convergence).**

Definition 5.6.1 (Convergence in distribution)**.** *A sequence of random variables X_n* converges in distribution *to X if $F_{X_n}(x) \to F_X(x)$ for every x at which F_X is continuous.*

This is also called weak convergence *and* convergence in law*. It can be denoted in many ways, including:*

$$X_n \xrightarrow{d} X, \qquad X_n \xrightarrow{\mathcal{D}} X, \qquad X_n \xrightarrow{\mathcal{L}} X, \qquad \mathcal{L}(X_n) \to \mathcal{L}(X), \qquad \text{and} \qquad X_n \Rightarrow X.$$

This is the weakest notion of convergence that we will cover.

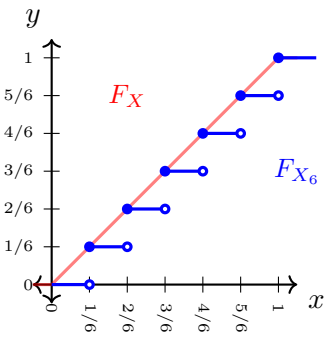
Exercise 5.6.1. *Show that if $X_n \to X$ in probability, then $X_n \to X$ in distribution.*

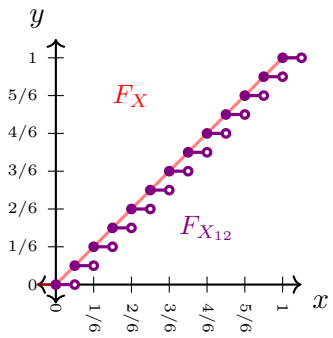
Notice that it *does not* require that X_n and X are defined on the same probability space!

Example 5.6.2. *(i) Let X_k be a sequence of random variables such that*

$$\mathbb{P}(X_k = y) = \begin{cases} \frac{1}{k} & \text{if } y = \frac{j}{k} \text{ for } j = 1,\dots,k, \\ 0 & \text{otherwise.} \end{cases} \tag{5.6.2}$$

Exercise 5.6.2. *Explicitly write down a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and the function X_k such that* (5.6.2) *holds.*





It is then clear that

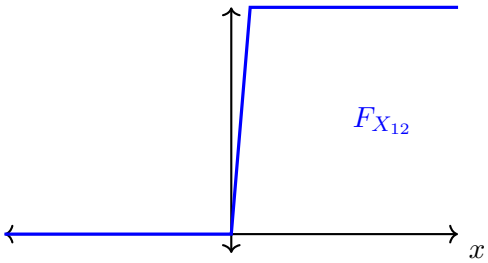
$$F_{X_k}(y) = \begin{cases} \frac{j}{k} & \text{if } y \in [j/k, j/k + 1/k) \text{ for } j = 1, \dots, k-1, \\ 1 & \text{if } y \geq 1, \\ 0 & \text{if } y < 1/k. \end{cases}$$

Hence, $F_{X_k} \to F_X$ where $X \sim \mathrm{Unif}(0,1)$.

Let us point out that F_{X_k} does not have a density, so there is no convergence of densities here.

(ii) Let $X_k \sim \mathrm{Unif}(0, 1/k)$. Then

$$F_{X_k}(y) = \begin{cases} yk & \text{if } y \in [0, 1/k], \\ 1 & \text{if } y \geq 1/k, \\ 0 & \text{if } y < 0. \end{cases}$$

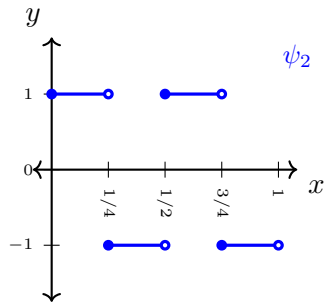


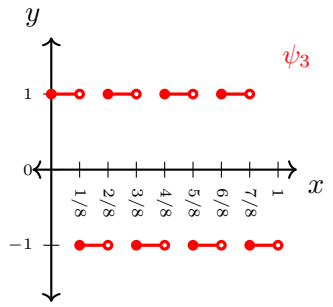
Clearly $F_{X_k}(y) \to \mathbb{1}_{[0,\infty)}$ for all $y \neq 0$. This corresponds to $X_k \to X$ in distribution, where X is the random variable $X(\omega) = 0$ for all $\omega \in \Omega$. Notice that $F_{X_k}(0) = 0$ and $F_X(0) = 1$, so that everywhere convergence of distribution functions cannot be expected.

(iii) Let $\psi_n : [0,1) \to \mathbb{R}$ be the random variable

$$\psi_n(x) = \sum_{\ell=0}^{2^k} (-1)^\ell \mathbb{1}_{\left[\frac{\ell}{2^k}, \frac{\ell+1}{2^k}\right)} \qquad \text{where } k \text{ is such that } 2^k \leq n < 2^{k+1}.$$

on the probability space $([0,1), \mathcal{B}, m)$.





Notice that, for every k,

$$F_{X_k}(t) = \mathbb{P}(X_k \leq t) = \begin{cases} 0 & \text{if } t < -1, \\ 1/2 & \text{if } t \in [-1, 1), \\ 1 & \text{if } t \geq 1. \end{cases}$$

In other words, the X_k are identically distributed. Hence, for any fixed n_0, we have $X_n \xrightarrow{d} X_{n_0}$ as $n \to \infty$. This is a bit silly... but it reflects how weak of a notion of convergence this is.

Exercise 5.6.3. *Show that convergence in distribution is equivalent to* weak convergence*: $\mathbb{E}[h(X_n)] \to \mathbb{E}[h(X)]$ for any continuous bounded function h. Deduce that $\Phi_{X_n}(t) \to \Phi_X(t)$ for any fixed t as $n \to \infty$.*

Note that the above exercise does *not* apply to computing moments! Why? Because $h(x) = x^p$ is not a bounded function.

This leads to the key tool in proving the Central Limit Theorem.

Theorem 5.6.3 (Lévy's continuity theorem)**.** *A sequence of random variables X_n (not necessarily on the same probability space) converges in distribution if and only if Φ_{X_n} converges pointwise to a function Φ that is continuous at $t = 0$.*

Further, if X is the (weak) limit of the X_k, then $\Phi_X = \Phi$.

Proof sketch. One direction is due to Exercise 5.6.3. The other direction is a bit more technical. Let us discuss it in the case where Φ is assumed to be everywhere continuous.

Let Φ be the pointwise limit of Φ_{X_n}. The Fourier transform:

$$\mathcal{F} : \mathcal{M} \to C(\mathbb{R})$$

is a bijection. Here $\mathcal{M}$ is the set of finite (signed) measures on $\mathbb{R}$. By properties of the Fourier transform (that we do not cover in these notes, unfortunately), $\mathcal{F}$ is a bijection with inverse $\mathcal{F}^{-1}$. Thus, there is a measure

$$\mu := \mathcal{F}^{-1}(\Phi).$$

We then let X be the random variable on $(\mathbb{R}, \mathcal{B}, \mu)$ defined by $X(t) = t$. One can then check that $X_n \xrightarrow{d} X$. □

Let us show a major application of such a theorem.

Theorem 5.6.4 (Central Limit Theorem)**.** *Let $X_1, X_2, \ldots$ be i.i.d. random variables on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with mean μ and variance σ^2. For any n, let*

$$S_n = \frac{X_1 + \cdots + X_n - \mu n}{\sqrt{n}}.$$

Then there is $S \sim \text{Norm}(0, \sigma^2)$ such that $S_n \Rightarrow S$ as $n \to \infty$.

Before continuing let us just mention that there is an enormous industry related to the Central Limit Theorem. First, there are many proofs of it, and these show that the assumptions of "identically distribution" and "independence" can be relaxed (although at the expense of strengthening other

assumptions such as the number of moments). Second, the Central Limit Theorem is the most basic example of the notion of *universality*, which is the idea that large, seemingly different classes of random processes all have the same behavior at some larger scale. A very active field of research related to this (and yielding a few Fields Medals in recent years).

Proof. Up to shifting (that is, changing to new random variables $\tilde{X}_n = X_n - \mu$), we may assume that $\mu = 0$. By Proposition 5.4.7, we thus have, for each k,

$$\Phi_{X_k}(0) = 1, \qquad \Phi'_{X_k}(0) = \mathbb{E}[iX_k] = 0, \qquad \text{and} \qquad \Phi''_{X_k}(0) = \mathbb{E}[i^2 X_k^2] = -\sigma^2. \tag{5.6.3}$$

Notice that, for each k and n,

$$\Phi_{X_k/\sqrt{n}}(t) = \mathbb{E}\left[\exp\left\{\frac{itX_k}{\sqrt{n}}\right\}\right] = \mathbb{E}\left[\exp\left\{i\frac{t}{\sqrt{n}}X_k\right\}\right] = \Phi_{X_k}\left(t/\sqrt{n}\right).$$

Additionally, note that $\Phi_{X_k} = \Phi_{X_1}$ for all k. Hence, using the independence of $(X_1, \ldots, X_k)$, we find

$$\Phi_{S_n}(t) = \prod_{k=1}^{n} \Phi_{X_k}(t/\sqrt{n}) = \Phi_{X_1}(t/\sqrt{n})^n. \tag{5.6.4}$$

In the first step, we essentially used (5.4.7) with the choice $\bar{t} = (t/\sqrt{n}, t/\sqrt{n}, \cdots, t/\sqrt{n})$.

It follows from (5.6.3) that[40]

$$\Phi_{X_1}(t/\sqrt{n}) = \Phi_{X_1}(0) + \frac{t}{\sqrt{n}}\Phi'_{X_1}(0) + \frac{1}{2}\left(\frac{t}{\sqrt{n}}\right)^2 \Phi''_{X_1}(0) + o(t^2/n) = 1 - \frac{t^2\sigma^2}{2}\frac{1}{n} + o(t^2/n). \tag{5.6.5}$$

Putting (5.6.4) and (5.6.5) together, we find

$$\Phi_{S_n}(t) = \left(1 - \frac{t^2\sigma^2}{2n} + o\left(\frac{t^2}{n}\right)\right)^n. \tag{5.6.6}$$

Recall[41] that, for all a,

$$\lim_{n\to\infty}\left(1 + \frac{a}{n}\right)^n = e^a,$$

Hence, (5.6.6) becomes

$$\lim_{n\to\infty} \Phi_{S_n}(t) = e^{-\frac{t^2\sigma^2}{2}} =: \Phi(t).$$

Clearly Φ is continuous at $t = 0$. Thus, by Theorem 5.6.3, there is S with characteristic function Φ such that $S_n \Rightarrow S$.

We note that Φ is the characteristic function of a random variable with distribution $\text{Norm}(0, \sigma^2)$. By the uniqueness of characteristic functions, this implies that $S \sim \text{Norm}(0, \sigma^2)$. This completes the proof. □

PERHAPS ADD EXAMPLE WITH ADDING INFINITE COIN FLIPS TO NORMAL AND RANDOM WALK TO BROWNIAN MOTION

[40] We write $o(t^2/n)$ to mean a quantity satisfying $\lim_{t^2/n\to 0} (n/t^2)\, o(t^2/n) = 0$ for any t.

[41] This can be shown by using Taylor polynomials with the remainder term to show that, for each a and n, there is ξ between 0 and a such that $n\log(1 + a/n) = a - \xi^2/2n$. Hence, *displaystyle* $\lim_{n\to\infty} n\log(1 + a/n) = a$.

5.6.2. **Convergence in probability.** We have already covered convergence in probability (convergence in measures, see Definition 4.7.2). Recall $X_k \to X$ in probability if, for all $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| > \varepsilon) = 0.$$

Also recall the exercise above that states that convergence in probability implies convergence in distribution. The opposite is not true.

Exercise 5.6.4. *Show that the sequences in Example 5.6.2.(i) and (ii) converge in probability as well, but the sequence in Example 5.6.2.(iii) does not. Note that this is a little ill-defined because, in (i) and (ii), the random variables are not explicitly defined on the same probability space, so do this as well.*

5.6.3. **Almost sure convergence.**

Definition 5.6.5. *A sequence of random variables X_n and X defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$ converges almost surely if*

$$\mathbb{P}(\{\omega : \limsup_{n\to\infty} |X_n(\omega) - X| > 0\}) = 0.$$

In other words, there is an event A such that $\mathbb{P}(A) = 1$ and $X_n \to X$ on A.

This is the same as almost everywhere convergence. Hence, from above, we know that almost sure convergence implies convergence in probability.

5.6.4. **Convergence in the pth mean.**

Definition 5.6.6. *A sequence of random variables X_n and X defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$ converges in the p-th mean if*

$$\mathbb{E}[|X_n - X|^p] = 0.$$

In other words, there is an event A such that $\mathbb{P}(A) = 1$ and $X_n \to X$ on A.

This is the same as L^p-convergence. Hence, from above, we know that convergence in the pth mean implies convergence in the qth mean for any $q \leq p$, convergence in probability, and convergence in distribution.

5.6.5. **Law of large numbers.** Let us return to the motivating example at the beginning of Section 5.6. We show that the first quantity in (5.6.1) converges to the mean of X_n.

Exercise 5.6.5. *Show that if X_n are all defined on the same probability space and X_n converges in distribution to a constant random variable, then X_n converges in probability as well.*

Theorem 5.6.7 (Weak law of large numbers)**.** *If X_n is a sequence of i.i.d. random variables defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with mean μ and variance σ^2, then*

$$S_n := \frac{X_1 + \cdots + X_n}{n}$$

converges in probability to μ.

The proof is essentially a simpler version of Theorem 5.6.4, and, hence, is omitted. Let us suggest that a more direct proof can be made using Markov's inequality.

Exercise 5.6.6. *Show both of these proofs.*

We note that the assumption of independence is not necessary (we only need "uncorrelated," that is, $\mathbb{E}[X_iX_j] = \mathbb{E}[X_i]\mathbb{E}[X_j]$ when $i \neq j$) as is the assumption of finite variance (we only need an L^1 bound). Finally, with significantly more work, we can deduce the *Strong law of large numbers*, which states that S_n converges almost surely to μ. We do not cover that here.

5.6.6. **Poisson convergence theorem and rare events.** For next year!

5.7. **Conditioning.** In the aim of simplicity, we almost exclusively consider continuous random variables in this section. For (purely) discrete random variables, it is clear how to to generalize everything we cover. The general case is messier, so we omit it.

5.7.1. **Conditional probability.** Conditional probability allows us to quantify independence. In the most basic example, where we have two events A and B on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, the conditional probability of A given B is

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

Note that this formula only makes sense if $\mathbb{P}(B) > 0$. This lets us know how much "information" B gives about A – is A more or less likely to occur given that we know B occurred? Notice that if A and B are independent, we immediately have that

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

In other words, knowing that B occurred tells us nothing about whether A will occur.

Example 5.7.1. *Let us take our three-coin flipping probability space $(\{0,1\}^3, \mathcal{P}^{\{0,1\}^3}, \mathbb{P})$, where $\mathbb{P}(\omega) = 1/8$ for all ω. Let $A = \{(1,1,1),(1,1,0),(1,0,0)\}$, that is, the first flip is heads and each flip after that is nonincreasing in value, and $B = \{(1,1,1),(0,0,0)\}$, that is, all flips have the same outcome. Then*

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A,B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(\{(1,1,1)\})}{1/2} = \frac{1/8}{1/4} = \frac{1}{2}.$$

In other words, if we are told that all flips are the same, then the probability that the first flip is heads and the rest are nonincreasing jumps up to $1/2$ from $3/8$.

Let X and Y be two absolutely continuous random variables such that (X,Y) is absolutely continuous as well. Fix any $A, B \in \mathcal{B}$, then

$$\mathbb{P}(X \in A | Y \in B) = \frac{\mathbb{P}(X \in A, Y \in B)}{\mathbb{P}(Y \in B)} = \frac{\int_{A\times B} f_{X,Y}(x,y)\, d(m\times m)}{\int_B f_Y(y)\, dm} \tag{5.7.1}$$

5.7.2. **Conditional expectation: conditioning on a random variable.** In this and the next section, we consider a random variable X on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Let us first consider the case where X is a continuous random variable. Suppose that Y is another absolutely continuous random variable and that (X, Y) is absolutely continuous as well.

If X and Y are continuous random variables with joint density $f_{X,Y}$, then the density of Y given that $X = x$ is

$$f_{X|Y=y}(y), f_{X|Y}(x|y) := \frac{f_{X,Y}(x,y)}{\int f_{X,Y}(x,y)\,dx} = \frac{f_{X,Y}(x,y)}{f_Y(y)}.$$

Intuitively, one can motivate this definition by looking at (5.7.1) and "taking the limit" $B \to \{y\}$. Notice that this is a probability density because

$$\int f_{X|Y=y}(x)\,dx = \int \frac{f_{X,Y}(x,y)}{f_Y(y)}\,dx = \frac{1}{f_Y(y)} \int f_{X,Y}(x,y)\,dx = \frac{1}{f_Y(y)} f_X(x) = 1.$$

In the second-to-last inequality, we used (5.3.1). Note that, since Y is continuous, the event $Y = y$ has probability zero! So we are conditioning upon an event that "will never" occur, in a sense. This may seem silly, but we will see its usefulness in a moment.

Then we define

$$\mathbb{E}[X|Y = y] = \int x f_{X|Y=y}(x)\,dm(x) = \int x \frac{f_{X,Y}(x,y)}{f_X(y)}\,dm(x).$$

This is, roughly, the average value of X given that we know $Y = y$; even more heuristically, it is the "best guess" for X given that $Y = y$ has occurred.

Let us point out that this is a function of y. In particular, we can define a function

$$Z : \Omega \to \mathbb{R}$$

by

$$Z(\omega) = \mathbb{E}[X|Y = Y(\omega)] = \int x \frac{f_{X,Y}(x, Y(\omega))}{f_Y(Y(\omega))}\,dm(x). \tag{5.7.2}$$

This will be measurable, making Z a random variable. Undoing this notation, we find that

$$\mathbb{E}[X|Y] := Z$$

is a random variable on our probability space!

We may also do the same in the discrete case. Let X, Y take values only in S_X and S_Y, which are countable subsets of $\mathbb{R}$. Then

$$\mathbb{E}[X|Y](\omega) = \sum_{s \in S_X} s \frac{\mathbb{P}(X = s, Y = Y(\omega))}{\mathbb{P}(Y = Y(\omega))}.$$

Example 5.7.2. *Consider two independent coin flips, in which our probability space is* $(\Omega, \mathcal{F}, \mathbb{P})$ *with*

$$\Omega = \{H, T\}^2, \quad \mathcal{F} = \mathcal{P}^\Omega, \quad \text{and} \quad \mathbb{P}(\{\omega\}) = 1/4 \qquad \text{for all } \omega \in \Omega.$$

Let us consider three random variables

$$X = Y + Z, \quad Y(\omega) = \begin{cases} 1 & \text{if } \omega_1 = H, \\ 0 & \text{if } \omega_1 = T, \end{cases} \qquad \text{and} \quad Z(\omega) = \begin{cases} 1 & \text{if } \omega_2 = H, \\ 0 & \text{if } \omega_2 = T. \end{cases}$$

Roughly, X is the number of heads, while Y and Z tell us if we get heads on the first or second flips, respectively.

First we compute $\mathbb{E}[X|Y]$. There are two cases. First consider the case where $Y(\omega) = 1$; that is $\omega_1 = H$. Then

$$\begin{aligned}\mathbb{E}[X|Y](\omega) &= \sum_{i=0}^{2} i\frac{\mathbb{P}(X=i, Y=1)}{\mathbb{P}(Y=1)} = \sum_{i=1}^{2} i\frac{\mathbb{P}(X=i, Y=1)}{1/2} \\ &= 2\left(1\mathbb{P}(X=1, Y=1) + 2\mathbb{P}(X=2, Y=1)\right) = 2\left(1\cdot 1/4 + 2\cdot 1/4\right) = \frac{3}{2}.\end{aligned}$$

Above we used that $\mathbb{P}(X=1, Y=1) = \mathbb{P}(\{(H,T)\}) = 1/4$ and $\mathbb{P}(X=2, Y=1) = \mathbb{P}(\{(H,H)\}) = 1/4$.

The second case is when $Y(\omega) = 0$; that is $\omega_1 = T$. Then

$$\begin{aligned}\mathbb{E}[X|Y](\omega) &= \sum_{i=0}^{2} i\frac{\mathbb{P}(X=i, Y=0)}{\mathbb{P}(Y=0)} = \sum_{i=1}^{2} i\frac{\mathbb{P}(X=i, Y=0)}{1/2} \\ &= 2\left(1\mathbb{P}(X=1, Y=0) + 2\mathbb{P}(X=2, Y=0)\right) = 2\left(1\cdot 1/4 + 2\cdot 0\right) = \frac{1}{2}.\end{aligned}$$

Above we used that $\mathbb{P}(X=1, Y=0) = \mathbb{P}(\{(T,H)\}) = 1/4$ and $\mathbb{P}(X=2, Y=0) = \mathbb{P}(\emptyset) = 0$. Hence, $\mathbb{E}[X|Y]$ is the random variable such that

$$\mathbb{E}[X|Y](\omega) = \begin{cases} 3/2 & \text{if } \omega_1 = H, \\ 1/2 & \text{if } \omega_1 = T. \end{cases}$$

Let us point out something important. The random variable $\mathbb{E}[X|Y]$ is measurable with respect to

$$\sigma(Y) := Y^{-1}(\mathcal{B}) = \{A \times \{H,T\} : A \subset \{H,T\}\},$$

which is the smallest σ-algebra on which Y is measurable. In other words, once we "know" what value Y takes, we know what value $\mathbb{E}[X|Y]$ takes. This will factor into our general definition below.

Next we compute $\mathbb{E}[Z|Y]$. Let us notice that Z and Y are independent. Hence, for any ω,

$$\begin{aligned}\mathbb{E}[Z|Y](\omega) &= \sum_{i=0}^{2} i\frac{\mathbb{P}(Z=i, Y=Y(\omega))}{\mathbb{P}(Y=Y(\omega))} = \sum_{i=0}^{2} i\frac{\mathbb{P}(Z=i)\mathbb{P}(Y=Y(\omega))}{\mathbb{P}(Y=Y(\omega))} \\ &= \sum_{i=0}^{2} i\mathbb{P}(Z=i) = \mathbb{E}[Z].\end{aligned}$$

In the second equality, we used the independence of Z and Y.

Why does this happen? Recall the intuition that $\mathbb{E}[Z|Y]$ is the best guess for Z once we know Y. Because Z and Y are independent, knowing the value that Y takes does not change our guess for the value Z should take. Hence, our best guess for Z remains $\mathbb{E}[Z]$, regardless of the knowledge of Y.

How do we define conditional expectation in general? Unfortunately, when (X, Y) is neither absolutely continuous nor discrete, it becomes a slightly more technical question, one that we do not have the tools to do properly. As such, we will give the definition, but do not worry too much about the details, beyond the intuition.

Definition 5.7.3. *Fix a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. If $X, Y \in L^2_{\mathbb{P}}$, we define*

$$\mathbb{E}[X|Y] = \underset{g(Y):\ g \text{ Borel measurable}}{\operatorname{argmin}} \mathbb{E}[(X - g(Y))^2]. \tag{5.7.3}$$

Let us rephrase this in useful ways. Let

$$\sigma(Y) = Y^{-1}(\mathcal{B})$$

be the smallest σ-algebra on which Y is measurable. The only other random variables that are $\sigma(Y)$ measurable are those of the form $g(Y)$. Let $S_Y = \{g(Y) : g \text{ is Borel measurable}\}$. Then S_Y is a closed subspace of $L^2_{\mathbb{P}}$. Then $\mathbb{E}[X|Y]$ is the closest element of S_Y to X. Another way to view this is as the projection of the random variable X onto the subspace S_Y.

The technical issue with this definition, for us, is that we do not know that such a minimum exists! On the other hand, this representation bakes-in the intuition that $\mathbb{E}[X|Y]$ is the best guess for X given information about Y. We do note that it is not hard to prove the uniqueness of the minimizer using this definition. Indeed, $g \mapsto \mathbb{E}[(X - g(Y))^2]$ is strictly convex in the sense that, for all $g \neq h$ and $\theta \in (0, 1)$,

$$\mathbb{E}[(X - \theta g(Y) - (1-\theta)h(Y))^2] < \theta \mathbb{E}[(X - g(Y))^2] + (1-\theta)\mathbb{E}[(X - h(Y))^2]. \tag{5.7.4}$$

Exercise 5.7.1. *Prove* (5.7.4) *and use it to show that there is a unique minimizer in Definition 5.7.3.*

Definition 5.7.4. *Fix a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Suppose that $X, Y \in L^1_{\mathbb{P}}$. Then $\mathbb{E}[X|Y]$ is the $L^1_{\mathbb{P}}$ random variable such that*

(i) $\mathbb{E}[X|Y] \in L^1_{\mathbb{P}}$;

(ii) $\mathbb{E}[X|Y]$ *is $\sigma(Y)$-measurable;*

(iii) $\mathbb{E}[g(Y)\mathbb{E}[X|Y]] = \mathbb{E}[g(Y)X]$ *for all g that are Borel measurable.*

Let us recall that $\sigma(Y) = Y^{-1}(\mathcal{B})$ is the smallest σ-algebra on which Y is measurable.

Seeing a definition like this, one must worry that such an object exists. Let us justify it briefly. One can check that the

$$\mu_{\pm}(A) = \int_A X_{\pm}\, d\mathbb{P}, \qquad \text{for all } A \in \sigma(Y),$$

are measures. By the Radon-Nikodym theorem[42], there is $f_{\pm}$ *that is $\sigma(Y)$ measurable* such that

$$\mu_{\pm}(A) = \int_A f_{\pm}\, d\mathbb{P}.$$

We define

$$\mathbb{E}[X|Y] = f_+ - f_-.$$

We obtain (*iii*) for the special case $g = \mathbb{1}_{Y^{-1}(A)}$ immediately. The remaining g may be obtained by approximation (although we skip this step). Clearly, (ii) follows immediately. To find (i), we define

$$g(y) = \operatorname{sign} \mathbb{E}[X|Y = y],$$

[42]Not covered in these notes.

and find

$$\mathbb{E}[|\mathbb{E}[X|Y]|] = \mathbb{E}[g(Y)\mathbb{E}[X|Y]] = \mathbb{E}[g(Y)X] \leq \mathbb{E}[|X|] < \infty. \tag{5.7.5}$$

Obviously, the downside to this is that it requires the Radon-Nikodym theorem, which we have not covered. The upside is that it requires less "regularity" than Definition 5.7.3: it is "easier" to be in $L^1_{\mathbb{P}}$ than $L^2_{\mathbb{P}}$ since $L^2_{\mathbb{P}} \subset L^1_{\mathbb{P}}$. In these notes, we use both definitions, opting for whichever is more convenient in the circumstances. The only difference that *you should be aware of* is that Definition 5.7.3 is only available if the $X, Y \in L^2_{\mathbb{P}}$.

Let us say a few brief words about their equivalence.

Sketch of the proof of the equivalence of Definition 5.7.4 and Definition 5.7.3. Suppose that $\bar{g}(Y) := \mathbb{E}[X|Y]$ is given by Definition 5.7.3. Fix any $\sigma(Y)$ measurable set A. Fix any h, and let

$$F_h(t) = \mathbb{E}[(X - (g(Y) + th(Y)))^2]. \tag{5.7.6}$$

Since $\bar{g}$ is the location of a minimum, we have

$$0 = \frac{1}{2}F_h'(0) = -\mathbb{E}[(X - (\bar{g}(Y) + 0 \cdot h(Y)))h(Y)] = -\mathbb{E}[Xh(Y)] + \mathbb{E}[\bar{g}(Y)h(Y)].$$

Rearranging this is exactly Definition 5.7.4.(iii). We omit (i) and (ii) since they are, respectively, addressed in (5.7.5) and are true by construction (note that $\mathbb{E}[X|Y] = \bar{g}(Y)$).

Now let us suppose that $\mathbb{E}[X|Y]$ is given by Definition 5.7.4. Since $\mathbb{E}[X|Y]$ is $\sigma(Y)$ measurable, there must exist $\bar{g}$ such that $\bar{g}(Y) = \mathbb{E}[X|Y]$. Fix any h and let F_h be as in (5.7.6). Then

$$\frac{1}{2}F_h'(0) = -\mathbb{E}[Xh(Y)] + \mathbb{E}[\bar{g}(Y)h(Y)] = 0.$$

The last inequality follows from Definition 5.7.3. Notice that the strict convexity of the functional in (5.7.3) (see (5.7.4)) implies F_h is strictly convex as well. It follows that F_h has a strict minimum at $t = 0$. Hence, for any g,

$$\mathbb{E}[(X - \bar{g}(Y))^2] = F_{g-\bar{g}}(0) < F_{g-\bar{g}}(1) = \mathbb{E}[(X - g(Y))^2].$$

Thus $\bar{g}$ is the minimized in Definition 5.7.3, as desired. □

Let us point out a few important properties:

(i) Suppose that X and Y are independent. Let us first compute assuming that all random variables are absolutely continuous. Then, $f_{X,Y}(x, y) = f_X(x)f_Y(y)$. It follows from (5.7.2) that

$$\mathbb{E}[X|Y] = \int x \frac{f_{X,Y}(x, Y)}{f_Y(Y)}\, dm(x) = \int x \frac{f_X(x)f_Y(Y)}{f_Y(Y)}\, dm(x) = \mathbb{E}[X].$$

In the general case, let us use Definition 5.7.3. Let us make a calculus style argument, in which we identify $\mathbb{E}[X|Y]$ by finding a critical point. Fix any Borel measurable g, h. Then, X and $g(Y)$ are independent, as are X and $h(Y)$. Consider the function

$$F(t) = \mathbb{E}[(X - (g(Y) + th(Y)))^2].$$

If $g(Y) = \mathbb{E}[X|Y]$, then

$$0 = \frac{1}{2}F'(0) = -\mathbb{E}[(X - (g(Y) + 0 \cdot h(Y)))h(Y)] = -\mathbb{E}[X]\mathbb{E}[h(Y)] + \mathbb{E}[g(Y)h(Y)].$$

It is clear that this is satisfied with the choice $g(Y) = \mathbb{E}[X]$. By the convexity of F (see Exercise 5.7.1), this "critical point" is the location of the minimum in Definition 5.7.3.

(ii) Let us find the expected value of the random variable $\mathbb{E}[X|Y]$. Since $\mathbb{E}[X|Y]$ is a function of Y, we know that its expectation is given by integrating against the density of Y (recall (5.2.5)):

$$\begin{aligned}\mathbb{E}[\mathbb{E}[X|Y]] &= \int \left(\int x \frac{f_{X,Y}(x,y)}{f_Y(y)} \, dm(x) \right) f_Y(y) \, dm(y) = \int \int x f_{X,Y}(x,y) \, dm(x) \, dm(y) \\ &= \int \int x f_{X,Y}(x,y) \, dm(y) \, dm(x) = \int \int x f_X(x) \, dm(x) = \mathbb{E}[X].\end{aligned}$$

Here we used Tonelli's theorem and (5.3.1).

Now let us see this in general. We use Definition 5.7.4. We have that, for all $A \in \sigma(Y)$,

$$\int_A \mathbb{E}[X|Y] \, d\mathbb{P} = \int_A X \, d\mathbb{P}.$$

Applying this with the choice $A = \Omega$, we have

$$\mathbb{E}\left[\mathbb{E}[X|Y]\right] = \int_A \mathbb{E}[X|Y] \, d\mathbb{P} = \int_\Omega X \, d\mathbb{P} = \mathbb{E}[X].$$

(iii) **Exercise 5.7.2.** *Show that* $\mathbb{E}[\mathbb{E}[X|Y,Z]|Y] = \mathbb{E}[X|Y]$. *Here it is useful to think of* $\mathbb{E}[X|Y,Z] = \mathbb{E}[X|(Y,Z)]$ *in order to understand how to define* $f_{X|Y,Z}(x|y,z)$.

5.7.3. **Conditional expectation: conditioning on a σ-algebra.** In an analogous way to the above, we can also obtain a random variable $\mathbb{E}[X|\mathcal{F}_1]$ for any σ-algebra $\mathcal{F}_1 \subset \mathcal{F}$. To begin, let us observe that the only reasonable way to define this should allow

$$\mathbb{E}[X|\sigma(Y)] = \mathbb{E}[X|Y].$$

What does this mean? Well, this represents the best guess for X if we knew Y. Thus, we should have that

$$\mathbb{E}[X|\mathcal{F}_1]$$

represents the best guess for X given all of the information in $\mathcal{F}_1$. Of course, we should think of this as a random variable, as we did above. In this sense, we should think of it as follows. Once a trial ω occurs, it gives us "information" according to $\mathcal{F}_1$ (although this does not tell us *everything*), and then we make our best guess for X given this. To illustrate this, let us consider a simple example:

Example 5.7.5. *Let us take* $([0,1]^2, \mathcal{B} \times \mathcal{B}, m \times m)$ *as a probability space,* $X(x,y) = x + y$, *and*

$$\mathcal{F}_1 = \{B \times [0,1] : B \in \mathcal{B}\}.$$

We should roughly think of $\mathcal{F}_1$ *as being the information about the first variable only. Then*

$$\mathbb{E}[X|\mathcal{F}_1]$$

should be the random variable expected value of X *as a function of knowing the first variable. It would only make sense to have:*

$$\mathbb{E}[X|\mathcal{F}_1] = x + \frac{1}{2}$$

because we simply average over all possibilities for y. *Hence, were we to "know" the "information" contained in* $\mathcal{F}_1$, *then the best guess for* X *would be* $\mathbb{E}[X|\mathcal{F}_1] = x + 1/2$.

We will check this rigorously below.

Let us give the analogues of Definition 5.7.3 and Definition 5.7.4, from above.

Definition 5.7.6. *Fix a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a sub-σ-algebra $\mathcal{F}_1 \subset \mathcal{F}$. If $X \in L^2_{\mathbb{P}}$, then*

$$\mathbb{E}[X|\mathcal{F}_1] = \underset{Z \text{ is } \mathcal{F}_1\text{-measurable}}{\operatorname{argmin}} \mathbb{E}[(X-Z)^2].$$

In other words, if we

Definition 5.7.7. *Let X be an L^1 random variable and $\mathcal{F}_1 \subset \mathcal{F}$. Then $\mathbb{E}[X|\mathcal{F}_1]$ is the random variable such that*

(i) $\mathbb{E}[X|\mathcal{F}_1] \in L^1$*;*

(ii) $\mathbb{E}[X|\mathcal{F}_1]$ *is $\mathcal{F}_1$-measurable;*

(iii) $\mathbb{E}[\mathbb{1}_A \mathbb{E}[X|\mathcal{F}_1]] = \mathbb{E}[\mathbb{1}_A X]$ *for all $A \in \mathcal{F}_1$.*

Note that (ii) says that, were we to know all of $\mathcal{F}_1$, then we would know the value of $\mathbb{E}[X|\mathcal{F}_1]$. To better understand (iii), let us consider the example $\mathcal{F}_1 = \sigma(Y)$ and $A = \{a \le Y \le b\}$ and "guess" that $\mathbb{E}[X|\mathcal{F}_1] = \mathbb{E}[X|Y]$. Then, we verify (iii):

$$\begin{aligned}
\mathbb{E}[\mathbb{1}_A \mathbb{E}[X|\mathcal{F}_1]] &= \mathbb{E}\left[\mathbb{1}_A \int x \frac{f_{X,Y}(x,Y)}{f_X(Y)}\, dm(x)\right] \\
&= \int_a^b \left(\int x \frac{f_{X,Y}(x,y)}{f_Y(y)}\, dm(x)\right) f_Y(y)\, dm(y) \\
&= \int_a^b \int x f_{X,Y}\, dm(x)\, dm(y) = \mathbb{E}[\mathbb{1}_A X].
\end{aligned}$$

Actually, notice that we used nothing here about the particular form of A; we simply used that A is defined only by Y.

In general, one should think of $\mathbb{E}[X|\mathcal{F}_1]$ as being defined by the following process. Given ω, look at the value $Y(\omega)$ for every random variable Y that is $\mathcal{F}_1$-measurable. Given this extra information (which may not be complete!), $\mathbb{E}[X|\mathcal{F}_1]$ is the best guess for X.

Before we get into technical details about the existence and uniqueness of $\mathbb{E}[X|\mathcal{F}_1]$, let us look at a few examples to get our bearings:

Example 5.7.8. *(i) If $\mathcal{F}_1 = \{\emptyset, \Omega\}$, which is the trivial σ-algebra, then $\mathbb{E}[X|\mathcal{F}_1] = \mathbb{E}[X]$. Notice that this is a constant random variable, so (i) and (ii) hold immediately. On the other hand, we need only check (iii) for the choices $A = \emptyset$ and Ω:*

$$\mathbb{E}[\mathbb{1}_\emptyset \mathbb{E}[X]] = 0 = \mathbb{E}[\mathbb{1}_\emptyset X],$$

and

$$\mathbb{E}[\mathbb{1}_\Omega \mathbb{E}[X]] = \mathbb{E}[X] = \mathbb{E}[\mathbb{1}_\Omega X].$$

Hence $\mathbb{E}[X|\mathcal{F}_1] = \mathbb{E}[X]$.

Roughly, this is saying that if we have "no information," our best guess of X is just $\mathbb{E}[X]$. Let's draw this out a bit more. A random variable Y is $\mathcal{F}_1$-measurable if and only if it is constant. How much "information" does a constant random variable contain? For example, if we flip two coins and have a random variable Y that always returns π, does knowing that $Y(\omega) = \pi$ tell us anything about whether $\omega = (H, H)$? Of course not!

(ii) *If $\mathcal{F}_1 = \mathcal{F}$, then $\mathbb{E}[X|\mathcal{F}_1] = X$. Roughly, this is saying that if we have "all of the information," then we know X so our best guess of it is just X itself (since we already know it!). It is easy to check that Definition 5.7.7.(i)-(iii) are satisfied.*

(iii) *Suppose that $\mathcal{F}_X$ and $\mathcal{F}_1$ are independent σ-algebras, with X being $\mathcal{F}_X$-measurable. Roughly, this tells us that $\mathcal{F}_X$ has all of the information to determine X and $\mathcal{F}_1$ has none of it. In this case, we expect*

$$\mathbb{E}[X|\mathcal{F}_1] = \mathbb{E}[X];$$

that is, knowing $\mathcal{F}_1$ does not improve our best guess for X.

Again, (i) and (ii) follow immediately because $\mathbb{E}[X|\mathcal{F}_1] = \mathbb{E}[X]$ is constant. As for (iii), notice that, if $A \in \mathcal{F}_1$, then $\mathbb{1}_A$ and X are independent[43] *by the independence of $\mathcal{F}_1$ and $\mathcal{F}_X$. Hence,*

$$\mathbb{E}[\mathbb{1}_A \mathbb{E}[X|\mathcal{F}_1]] = \mathbb{E}[\mathbb{1}_A \mathbb{E}[X]] = \mathbb{E}[\mathbb{1}_A]\mathbb{E}[X] = \mathbb{E}[\mathbb{1}_A X].$$

(iv) *Consider Example 5.7.5 again. We check that, for all $\omega = (x, y)$,*

$$\mathbb{E}[X|\mathcal{F}_1](\omega) = x + \frac{1}{2}.$$

Let

$$g(\omega) = g(x, y) = x + \frac{1}{2}.$$

Our goal, then, is to show that $\mathbb{E}[X|\mathcal{F}_1] = g$.

It is easy to see that Definition 5.7.7.(i) is satisfied by g. To check (ii), notice that, for any Borel set B,

$$g^{-1}(B) = \tilde{B} \times [0, 1] \qquad \text{where } \tilde{B} = \{x \in [0, 1] : x + 1/2 \in B\}.$$

Clearly, $g^{-1}(B) \in \mathcal{F}_1$. Finally, we check (iii). A set $A \in \mathcal{F}_1$ if and only if there is $\tilde{A} \subset [0, 1]$ such that $A = \tilde{A} \times [0, 1]$. Then

$$\begin{aligned}
\mathbb{E}[\mathbb{1}_A g] = \mathbb{E}[\mathbb{1}_{\tilde{A}\times[0,1]} g] &= \int_{\tilde{A}\times[0,1]} g(x, y)\, d(m \times m) \\
&= \int_{\tilde{A}} \left(\int_0^1 \left(x + \frac{1}{2} \right) dm \right) dm = \int_{\tilde{A}} \left(x + \int_0^1 y\, dm \right) dm \\
&= \int_{\tilde{A}} \left(\int_0^1 (x + y)\ dm \right) dm = \int_{[0,1]^2} \mathbb{1}_A(x, y)(x + y)\, d(m \times m) \\
&= \mathbb{E}[\mathbb{1}_A X].
\end{aligned}$$

In the third equality, we used Tonelli's theorem.

(v) *Let $A, B \in \mathcal{F}$ with $\mathbb{P}(B) > 0$. Consider $\sigma(B) = \{\emptyset, B, B^c, \Omega\}$. This is the σ-algebra containing the information of whether B has occurred or not. Then*

$$\mathbb{E}[\mathbb{1}_A|\sigma(B)] = \mathbb{P}(A|B)\mathbb{1}_B + \mathbb{P}(A|B^c)\mathbb{1}_{B^c} = \frac{\mathbb{P}(A, B)}{P(B)}\mathbb{1}_B + \frac{\mathbb{P}(A, B^c)}{\mathbb{P}(B^c)}\mathbb{1}_{B^c}.$$

[43] Let us check this: given a Borel set B, $\mathbb{1}_A^{-1}(B)$ is either $\emptyset$, A, or Ω. It is clear that each of these events is independent from $X^{-1}(B)$, the first and third trivially and the second by the independence of $\mathcal{F}_1$ and $\mathcal{F}_X$. Recall the Definition 5.5.3.

Exercise 5.7.3. *Verify that this satisfies Definition 5.7.7.*

By the definition of integration, any random variable can be approximated by simple functions, which are linear combinations of indicator functions. Hence, after rewriting the above as

$$\mathbb{E}[\mathbb{1}_A|\sigma(B)] = \frac{\mathbb{E}[\mathbb{1}_A\mathbb{1}_B]}{P(B)}\mathbb{1}_B + \frac{\mathbb{E}[\mathbb{1}_A\mathbb{1}_{B^c}]}{\mathbb{P}(B^c)}\mathbb{1}_{B^c}.$$

we see that

$$\mathbb{E}[X|\sigma(B)] = \frac{\mathbb{E}[\mathbb{1}_B X]}{\mathbb{P}(B)}\mathbb{1}_B + \frac{\mathbb{E}[\mathbb{1}_{B^c} X]}{\mathbb{P}(B^c)}\mathbb{1}_{B^c}.$$

Let us consider a simple case of this. Let $X \sim \text{Norm}(0,1)$ and $B = X^{-1}(\mathbb{R}_+)$. Clearly $\mathbb{P}(B) = \mathbb{P}(B^c) = 1/2$. We have

$$\mathbb{E}[X|\sigma(B)] = 2\mathbb{E}[\mathbb{1}_B X]\mathbb{1}_B + 2\mathbb{E}[\mathbb{1}_{B^c} X]\mathbb{1}_{B^c}.$$

Then, by symmetry and a direct computation,

$$-\mathbb{E}[\mathbb{1}_{B^c} X] = \mathbb{E}[\mathbb{1}_B X] = \frac{1}{\sqrt{2\pi}}\int_0^\infty x e^{-\frac{x^2}{2}}\,dx = \frac{1}{\sqrt{2\pi}}.$$

We thus have

$$\mathbb{E}[X|\sigma(B)] = \sqrt{\frac{2}{\pi}}\left(\mathbb{1}_B - \mathbb{1}_{B^c}\right).$$

(vi) Suppose that Ω can be partitioned by disjoint sets $\Omega_1, \ldots, \Omega_n$ with $\mathbb{P}(\Omega_i) > 0$ for all $i = 1, \ldots, n$ and $n \geq 2$. Then let $\mathcal{F}_1 = \sigma(\Omega_1, \ldots, \Omega_n)$, which contains all of the information of whether Ω_i or Ω_i^c have occurred. Recall Example 4.3.6.

Then, arguing exactly as in the previous example (where $\Omega_1 = B$ and $\Omega_2 = B^c$), we find

$$\mathbb{E}[\mathbb{1}_A|\mathcal{F}_1] = \sum_{i=1}^n \frac{\mathbb{P}(A,\Omega_i)}{\mathbb{P}(\Omega_i)}\mathbb{1}_{\Omega_i} = \sum_{i=1}^n \frac{\mathbb{E}[\mathbb{1}_A\mathbb{1}_{\Omega_i}]}{\mathbb{P}(\Omega_i)}\mathbb{1}_{\Omega_i}$$

For random variables, this becomes

$$\mathbb{E}[X|\mathcal{F}_1] = \sum_{i=1}^n \frac{\mathbb{E}[\mathbb{1}_{\Omega_i} X]}{\mathbb{P}(\Omega_i)}\mathbb{1}_{\Omega_i}$$

It is not apparent from Definition 5.7.7, but conditional expectation is well-defined. We state this in a proposition below, although we do not present the proof in its entirety. It is based on the Radon-Nikodym theorem[44], which is a kind of Fundamental Theorem of Calculus for measures.

Proposition 5.7.9. *Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, a random variable $X \in L^1$, and a sub-σ-algebra $\mathcal{F}_1 \subset \mathcal{F}$, there is a random variable $\mathbb{E}[X|\mathcal{F}_1]$ satisfying Definition 5.7.7. Moreover, if Y is another random variable satisfying Definition 5.7.7 then $Y = \mathbb{E}[X|\mathcal{F}_1]$ a.s.*

[44]The statement of this theorem is roughly that if two σ-finite measures on the same measure space satisfy "$\nu \ll \mu$" then there is $f \in L^1_\mu$ such that $d\nu = f d\mu$. Here "$\nu \ll \mu$" means that $\nu(B) = 0$ whenever $\mu(B) = 0$.

Existence. Define the measure $\mathbb{Q}_\pm : \mathcal{F}_1 \to \mathbb{R}_+$ by

$$\mathbb{Q}_\pm(A) = \int_A \mathbb{1}_{\pm X>0} X d\mathbb{P}.$$

It is easy to check that $\mathbb{Q}_\pm \ll \mathbb{P}$ so there exist $Y_\pm$, which are $\mathcal{F}_1$ measurable, such that

$$d\mathbb{Q}_\pm = Y_\pm d\mathbb{P}.$$

Let $\mathbb{E}[X|\mathcal{F}_1] = Y_+ - Y_-$. This is clearly $\mathcal{F}_1$ measurable and it satisfies Definition 5.7.7.(iii) by construction. To see that it is L^1, notice that

$$\begin{aligned}\mathbb{E}[|\mathbb{E}[X|\mathcal{F}_1]|] &= \mathbb{E}[\mathbb{1}_{\{\mathbb{E}[X|\mathcal{F}_1]>0\}}\mathbb{E}[X|\mathcal{F}_1] - \mathbb{1}_{\{\mathbb{E}[X|\mathcal{F}_1]<0\}}\mathbb{E}[X|\mathcal{F}_1]] \\ &= \mathbb{E}[\mathbb{1}_{\{\mathbb{E}[X|\mathcal{F}_1]>0\}}] - \mathbb{E}[\mathbb{1}_{\{\mathbb{E}[X|\mathcal{F}_1]<0\}}X] \\ &\leq \mathbb{E}[|X|].\end{aligned}$$

as desired. □

Uniqueness. For each n, let

$$A_n = \{Y - \mathbb{E}[X|\mathcal{F}_1] \geq 1/n\}.$$

Notice that, by Definition 5.7.7,

$$\frac{\mathbb{P}(A_n)}{n} \leq \mathbb{E}[\mathbb{1}_{A_n}(Y - \mathbb{E}[X|\mathcal{F}_1])] = \mathbb{E}[\mathbb{1}_{A_n}(X - X)] = 0.$$

Since this holds for all n, we have

$$\mathbb{P}(Y > \mathbb{E}[X|\mathcal{F}_1]) = \mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = 0.$$

Similarly, one can show that $\mathbb{P}(Y < \mathbb{E}[X|\mathcal{F}_1]) = 0$, which completes the proof. □

Exercise 5.7.4. *Show the following properties of conditional expectation:*

(i) (Linearity) $\mathbb{E}[X + Y|\mathcal{F}_1] = \mathbb{E}[X|\mathcal{F}_1] + \mathbb{E}[Y|\mathcal{F}_1]$*;*

(ii) (Order preserving) $\mathbb{E}[X|\mathcal{F}_1] \leq \mathbb{E}[Y|\mathcal{F}_1]$ *if* $X \leq Y$*;*

(iii) (Monotone convergence) $\mathbb{E}[X_n|\mathcal{F}_1] \to \mathbb{E}[X|\mathcal{F}_1]$ *if* $X_n \nearrow X$*;*

(iv) (Jensen's inequality) $\varphi(\mathbb{E}[X|\mathcal{F}_1]) \leq \mathbb{E}[\varphi(X)|\mathcal{F}_1]$ *if* φ *is convex*[45]*;*

(v) (Tower property) $\mathbb{E}[X|\mathcal{F}_1] = \mathbb{E}[\mathbb{E}[X|\mathcal{F}_2]|\mathcal{F}_1]$ *if* $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \mathcal{F}$*;*

(vi) (Stability) If X *is* $\mathcal{F}_1$*-measurable, then* $\mathbb{E}[X|\mathcal{F}_1] = X$*. Further, for any Borel measurable* g*, we have* $\mathbb{E}[g(X)|\sigma(X)] = g(X)$*;*

(vii) (Removing a "known" factor) If X *is* $\mathcal{F}_1$*-measurable, then* $\mathbb{E}[XY|\mathcal{F}_1] = X\mathbb{E}[Y|\mathcal{F}_1]$.

[45] This gives Markov's inequality for $\mathbb{E}[\cdot|\mathcal{F}_1]$.

Let us make a note about (iii) and (iv). The notation, which involves an $\mathbb{E}$, makes this seems obvious. Keep in mind, however, that notation is just suggestion and that $\mathbb{E}[X|\mathcal{F}_1]$ is a *random variable*, so these are by no means obvious.

Let us also make a note about (v). The intuitive meaning of $\mathbb{E}[X|\mathcal{F}_1]$ is the best guess for X if you are restricted to the information on $\mathcal{F}_1$. Hence $\mathbb{E}[\mathbb{E}[X|\mathcal{F}_2]|\mathcal{F}_1]$ is the best guess for X restricted to the information in $\mathcal{F}_2$ and then further restricting to the information in $\mathcal{F}_1$. Since $\mathcal{F}_1 \subset \mathcal{F}_2$, this should just be the best guess for X restricted to the information in $\mathcal{F}_1$.

Exercise 5.7.5. *Show that $\mathbb{E}[|\mathbb{E}[X|\mathcal{F}_1]|^p] \leq \mathbb{E}[|X|^p]$ for any $p \geq 1$. This shows that conditional expectation is a (weak) contraction.*

There is another perspective of $\mathbb{E}[X|\mathcal{F}_1]$ as the L^2-projection of X onto the linear subspace of $L^2_{\mathbb{P}}$ spanned by random variables of the form $\mathbb{1}_A$ where $A \in \mathcal{F}_1$. One can see why that would be the case from Definition 5.7.7.(iii). We point out that this lends credibility to our intuition about $\mathbb{E}[X|\mathcal{F}_1]$ being the "best guess" for X given the information in $\mathcal{F}_1$. Indeed, recall that the projection of a vector v onto a subspace S is the closest point in S to v.

Taking advantage of this perspective requires X to be L^2. Note that we have not yet proved that projections exist in L^2, but this fact is obvious in finite dimensional inner product spaces so one should not be too surprised it is true in the infinite dimensional case as well.

In this L^2 setting, we get analogous versions of our usual inequalities (just like we did in L^1):

Exercise 5.7.6. *Show that Cauchy-Schwarz holds: $\mathbb{E}[XY|\mathcal{F}_1]^2 \leq \mathbb{E}[X^2|\mathcal{F}_1]\mathbb{E}[Y^2|\mathcal{F}_1]$ for all random variables X and Y.*

5.8. **Information theory.** Using our intuition of σ-algebras as "information," we can think of a scenario where our situation is "updating" step-by-step as one in which we have an infinite increasing sequence of σ-algebras. This leads to the following:

Definition 5.8.1. *Let $(\Omega, \mathcal{F})$ be a measurable space. A filtration on $(\Omega, \mathcal{F})$ is a sequence*

$$\mathcal{F}_1 \subset \mathcal{F}_2 \subset \cdots \subset \mathcal{F}.$$

This is the discrete version, one can also have a filtration indexed by $\mathbb{R}_+$ or $\mathbb{R}$, etc. In this case, $\{\mathcal{F}_t\}$ is a filtration if $\mathcal{F}_t \subset \mathcal{F}_{t+s}$ for all t and any $s \geq 0$.

Example 5.8.2. *Let us consider an infinite sequence of coin flips. Here $\Omega = \{0,1\}^{\mathbb{N}}$ and $\mathcal{F} = \mathcal{P}^{\Omega}$. Let X_i be the random variable that returns the value of the ith flip: $X_i(\omega) = \omega_i$, where $\omega = (\omega_1, \omega_2, \dots)$. We define the filtration $\mathcal{F}_1, \mathcal{F}_2, \dots$ by*

$$\mathcal{F}_i = \sigma(X_1, \dots, X_i).$$

This is a bit abstract, and, luckily, that is usually good enough in practice. That said, what does $\mathcal{F}_i$ look like? Recall that $\sigma(X_1)$ is the smallest σ-algebra on which X_1 is measurable, which requires that contains precisely those sets of the form $X_1^{-1}(B)$ where B is a Borel set (recall Definition 4.4.2).

The smallest σ algebra generated by X_1 is thus $\{\emptyset, \Omega, \{0\} \times \{0,1\}^{\mathbb{N}}, \{1\} \times \{0,1\}^{\mathbb{N}}\}$ since X_1 only specifies the first coordinate. Similarly,

$$\sigma(X_1, X_2) = \sigma\left(\{\emptyset, \Omega, \{(0,1)\} \times \{0,1\}^{\mathbb{N}}, \{(0,1)\} \times \{0,0\}^{\mathbb{N}}, \{(0,1)\} \times \{1,0\}^{\mathbb{N}}, \{(1,1)\} \times \{0,1\}^{\mathbb{N}}\}\right).$$

These are the events that only encode the information in the first two flips.

Let us briefly have a high level discussion of what we expect in quantifying the "amount of information."

(i) The information in each independent coin toss is the same. More generally, if X and Y are independent random variables, the information from knowing both X and Y should be the sum of the informations from each of them separately. In some sense, this is saying we "take products to sums" in our quantification.

(ii) Knowledge that a "rare" event occurred, say winning the lottery, is more informative than knowing that a rare event did not occur. So the information gained from an experiment should depend on the outcome, as well as the *a priori* expectation of the event actually occurring. On the other hand, a probability one event should encode no information.

5.8.1. **Information in a single random variable.** These two principles suggest using logarithms, which take products to sums and (up to sign) take probabilities close to zero to huge values.

Definition 5.8.3. *Suppose that X is a discrete random variable (that is, takes values $\alpha_1, \alpha_2, \ldots$ with $\mathbb{P}(X = \alpha_i) > 0$. Then the* self-information $I_X : \Omega \to \mathbb{R}_+$*, called the "surprisal," is given by*

$$I_X(\omega) = -\log \mathbb{P}(X = X(\omega)).$$

This tells us how "surprised" we should be if ω occurs. In some sense, it tells us how much we can restrict the set of ω that is consistent with the observed outcome. Notice that it is always non-negative.

If we abuse notation briefly, we see why the log function is a good choice. Let A and B be independent events. If both happens, we should deduce the "information" gained from A occurring and the "information" gained from B occurring and this should be a sum (there is no overlap in information because of independence). Sure enough:

$$I_X(A \cap B) = -\log \mathbb{P}(A \cap B) = -\log\left(\mathbb{P}(A)\mathbb{P}(B)\right) = -\log \mathbb{P}(A) - \log \mathbb{P}(B) = I_X(A) + I_X(B).$$

Definition 5.8.4. *The* entropy *of a discrete random variable X taking the values $\alpha_1, \alpha_2, \ldots, \alpha_n$ (where we allow $n = \infty$) is*

$$H(X) = \mathbb{E}[I_X] = \sum_{i=1}^{n} p_i \log(p_i),$$

where $p_i = \mathbb{P}(X = \alpha_i)$. This is sometimes called Shannon entropy*, after Claude Shannon, the founder of Information Theory.*

It is somewhat instructive to pause and discuss the etymology of "entropy" as well as choice of H (instead of a more reasonable E). The term "energy" has, as a root, "erg" which is a unit of work and comes from "ergon" ($\varepsilon\rho\gamma o\nu$). In analogy with this, Rudolph Clausius, coined "entropy" based on the root "tropé" ($\tau\rho o\pi\eta$), meaning "transformation." In gas dynamics, entropy measures how ordered a system is and is a key aspect of its trend towards equilibrium (this is Ludwig Boltzmann's "H Theorem"). In this sense, high entropy means high order and low entropy means low order (everything is "mixed" together).

Entropy was imported into information theory by its founder, Claude Shannon, who took inspiration from the concept in gas dynamics. Roughly, $H(X)$ is meant to represent how much information we

expect to get from an experiment that measures X. The analogy works as follows. In gas dynamics, high entropy means a gas with high order – someone or something arranged them – which means that there is high "information" in experimentally testing the gas. If you check an area of a room and find no gas particles there, you learn that something is affecting the room. On the other hand, low entropy means things are well-mixed. any experimental measurement of a room will essentially reveal the same thing (an essentially constant amount of gas particles in any area), which reveals low information to the experimenter.

Returning to mathematical physics. Ludwig Boltzmann initially used E for entropy; however, over time he and his contemporaries settled on H. Presumably this helps avoid a misunderstanding with "energy," often denoted by E. Why H? It is not totally clear, but the best guess given the limited historical record is that the H is actually a capital η (eta), which corresponds to the short e sound similar to the beginning of the word entropy.

Example 5.8.5. *Let us look at coin flipping. Let $\Omega_n = \{0,1\}^n$, $\mathcal{F} = \mathbb{P}^{\Omega_n}$, and $\mathbb{P}_n(\{\omega\}) = 2^{-n}$ for every n.*

(i) Case $n = 1$: let $X : \Omega_1 \to \mathbb{R}$ be given by $X(\omega) = \omega$. In this case,

$$I_X(\omega) = -\log \mathbb{P}(X = X(\omega)) = -\log(1/2) = \log 2,$$

and

$$H(X) = \sum_{\omega=0,1} \mathbb{P}(X = \omega) I_X(\omega) = \sum_{\omega=0,1} \frac{1}{2} \log 2 = \log 2.$$

Actually, one has a choice in defining "which log*" to use. The standard ones are* $\log_2$, $\log_e = \ln$, *or* $\log_{10}$. *If we use* $\log_2$, *then $I_X(\omega) = 1$ and we call this unit a "shannon." (Think of this as being analogous to a "bit," the unit for the memory on your computer). The unit for* ln *is "nat," and the unit for* $\log_{10}$ *is "Hartley."*

(ii) General case: the information in n coin flips. Let $X : \Omega_n \to \mathbb{R}^n$ be defined by $X(\omega) = \omega$. Then

$$I_X(\omega) = -\log \mathbb{P}(X = \omega) = -\log(1/2^n) = n \log 2.$$

Hence,

$$H(X) = \sum_{\omega \in \Omega} \frac{1}{2^n} I_X(\omega) = \sum_{\omega \in \Omega} \frac{n \log 2}{2^n} = n \log 2.$$

Recall that Ω has 2^n elements in it. This is $n \log 2$ nats or n shannons.

Example 5.8.6. *How much information is in a word? Let $\Omega_{A,n} = \{1, 2, \ldots, N_A\}^n$ and $\mathcal{F} = \mathbb{P}^{\Omega_A}$. Roughly, we think of N_A as being the number of letters in our alphabet. We encode each word with n letters as a sequence of n integers between 1 and N_A. For a first cartoon example, let us assume every letter is equally likely in our alphabet*

$$\mathbb{P}_n(\{\omega\}) = N_a^{-n}.$$

This is, of course, no reasonable: more t's show up in English than q's and the preceding letter skews the probability of the letter (e.g., u is overwhelmingly likely to follow q); however, let us strive for simplicity for the moment.

Exercise 5.8.1. *Argue exactly as above to find*

$$I_X(\omega) = H(X) = n \log N_a.$$

Let L be a random variable that uniformly draws a random letter. Then X is made up by drawing from n i.i.d. random variables distributed as L. It follows from the law of large numbers (Theorem 5.6.7) that

$$\frac{I_X(\omega)}{n} \to H(L).$$

One way to check if a word / sentence / book is "typical" is to check if the above asymptotics hold.

Exercise 5.8.2. *Show that the procedure above provides an upper bound. In other words, giving the ith letter probability p_i of occurring at each location in a word, then $H(X) \leq n \log N_a$.*

5.8.2. **Information in two random variables.**

Definition 5.8.7. *Suppose that X and Y are two, possibly dependent, discrete random variables. We have the following quantities:*

(i) (Joint entropy)

$$H(X,Y) = -\sum_{i,j} p_{ij} \log p_{ij},$$

where i, j range over all possible values of X and Y, respectively. This is obtained via Definition 5.8.4 applied to the random variable (X,Y).

(ii) (Conditional entropy) Using conditional probability, we see that

$$\begin{aligned} H(X|Y=\beta_j) &= \mathbb{E}[I_X|Y=\alpha_j] = -\sum_i \mathbb{P}(X=\alpha_i|Y=\alpha_j) \log \mathbb{P}(X=\alpha_i|Y=\alpha_j) \\ &= -\sum_i \frac{p_{ij}}{p_j} \log \frac{p_{ij}}{p_j}. \end{aligned}$$

Averaging over all β_j, we obtain the conditional entropy:

$$H(X|Y) = -\sum_j p_j \sum_i \frac{p_{ij}}{p_j} \log \frac{p_{ij}}{p_j} = -\sum_{ij} p_{ij} \log \frac{p_{ij}}{p_j}.$$

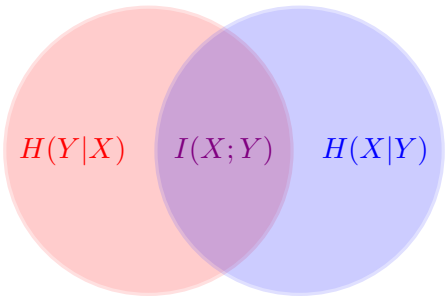
(iii) (Mutual information)

$$I(X;Y) = H(X) - H(X|Y) = H(X) + H(Y) - H(X,Y).$$

The heuristic meaning of (i) is analogous as in Definition 5.8.3 and Definition 5.8.4.

Exercise 5.8.3. *Show that $H(X|Y) = H(X,Y) - H(Y)$. This leads to the interpretation that $H(X|Y)$ is the extra information that X provides, beyond what Y provided.*

Finally, (iii) is simply the total information of shared by both X and Y.



This is represented in the figure above. The blue and violet area is $H(X)$ (the information provided by X), the red and violet area is $H(Y)$ (the information provided by Y), and the union of the blue and red areas is $H(X,Y)$ (the information provided by both X and Y). The blue area is $H(X|Y)$ while the red area is $H(Y|X)$. Finally, the mutual information $I(X;Y)$, the violet area, is the information provided by both X and Y.

5.9. **Markov chains.** Every random variable we have looked at so far is "static," but many of the phenomena in the world that we wish to understand through math are "dynamic." To begin to build up our intuition and theoretical foundations for this, let us consider the simplest possible random process.

Suppose that a very confused person is walking along a long road. At each intersection, they flip a coin, heading right if it comes up heads and left if it comes up tails. This is intuitively a very simple process. How do we make it into rigorous mathematics? We focus on "discrete time" and "discrete space" Markov chains here.

The key features that we want to encode are the fact that process is "homogeneous" and "memoryless" – the next step in the chain depends *only* on the current state of the chain.

Definition 5.9.1. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and let S be a finite "state space." Then an indexed set of random variables $X_n : \Omega \to S$ is a* Markov chain *if, for every $n \in \mathbb{N}$ and $x_0, \ldots, x_{n+1} \in S$, we have*

$$\mathbb{P}(X_{n+1} = x_{n+1} | X_n = x_n, X_{n-1} = x_{n-1}, \ldots, X_0 = x_0) = \mathbb{P}(X_{n+1} = x_{n+1} | X_n = x_n).$$

Let us make a few notes about this before jump into some examples.

1. Since S is countable, we may always consider it to be a finite or infinite subset of $\mathbb{Z}$ by identifying each element of S with an integer through the bijection in the definition of countable Definition 3.2.7.

2. The second condition encodes the "memoryless" quality discussed above. Indeed, it says that knowing the first $n-1$ values of the chain allows you no better guess of X_{n+1} than just knowing the current state X_n.

3. Actually the definition above does not encode the "time homogeneity" mentioned above, but, for the remainder of these notes, let us take that as an assumption. To encode this, we further assume that, for any $i, j \in S$,

   $$\mathbb{P}(X_{n+1} = j | X_n = i) = \mathbb{P}(X_1 = j | X_0 = i).$$

   In other words, the probability to step from i to j does not depend on the "time" n.

The last point above allows us to define, for any $i, j \in S$,

$$p_{ij} = \mathbb{P}(X_1 = i | X_0 = j). \tag{5.9.1}$$

and the $N \times N$ matrix

$$P = (p_{ij})_{ij}, \tag{5.9.2}$$

where N is the number of elements of S. Note that P need not be symmetric and that the order of the indices ij reflects the order of the step from i to j.

Definition 5.9.2. *Let P be the matrix with entries defined by* (5.9.1)-(5.9.2). *We call this the* transition matrix *of the Markov chain X_n.*

In general, an $n \times n$ matrix P is called a stochastic matrix *if $P_{i1} + P_{i2} + \cdots + P_{in} = 1$ for each i.*

We immediately notice that

$$\sum_{i \in S} p_{ij} = 1 \qquad \text{that is,} \quad P\mathbf{1} = \mathbf{1},$$

where $\mathbf{1}$ is the vector with all entries equal to one. Let us consider what happens when we apply P. Suppose that $X_0 = i$. We consider $\bar{e}_i$ (the vector of all zeros except a one in the ith entry) and, for any j, notice that

$$(\bar{e}_i P)_j = P_{ij} = \mathbb{P}(X_1 = j | X_0 = i).$$

Next

$$\begin{aligned}(\bar{e}_i P^2)_j = \sum_{k=1}^{n} P_{ik} P_{kj} &= \sum_{k=1}^{n} \mathbb{P}(X_2 = j | X_1 = k) \mathbb{P}(X_1 = k | X_0 = i) \\ &= \mathbb{P}(X_2 = j | X_0 = i).\end{aligned}$$

By induction, we get that, for every m,

$$(\bar{e}_i P^m)_j = \mathbb{P}(X_m = j | X_0 = i).$$

Exercise 5.9.1. *Show this!*

This is because the matrix multiplication encodes all paths with which the Markov chain $X_\cdot$ can traverse to make it from i to j in m steps.

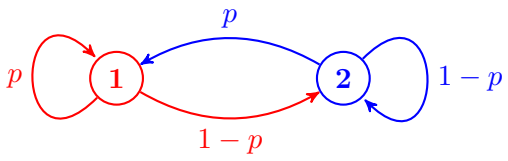
More generally, one can randomly pick a start location according to the distribution $\pi^{(0)}$, a vector where $\pi_i^{(0)} = \mathbb{P}(X_0 = i)$, and

$$\pi^{(k)} := \pi^{(0)} P^k = (\mathbb{P}(X_k = 1), \mathbb{P}(X_k = 2), \ldots, \mathbb{P}(X_k = n)).$$

Example 5.9.3. *(Bernoulli process) Fix $p \in (0,1)$ and $S = \{1, 2\}$. A cartoon picture to get you started is of a javelina that has two trash cans that it likes to eat from. Every day, it wakes up and decides which trash can it wants to eat from: with probability p, it eats from trash can 1 and with probability $1 - p$, it eats from trash can 2. Notice that its past does not influence its decision. Actually, in this silly model, even its current state (location) does not affect its decision!*

Define

$$p_{ij} = \begin{cases} p & \text{if } j = 1, \\ 1 - p & \text{if } j = 2. \end{cases}$$



Let us point out two related things. First, notice that, if we let $\bar{\pi} = (p, 1-p)$*, then*

$$\pi P = [p \cdot p + (1-p) \cdot p, p \cdot (1-p) + (1-p) \cdot (1-p)] = \pi.$$

Additionally, for any distribution $\pi^{(0)}$*,*

$$\lim_{n \to \infty} \pi^{(0)} P^n = \bar{\pi}. \tag{5.9.3}$$

Actually $\pi^{(0)} P = \bar{\pi}$*.*

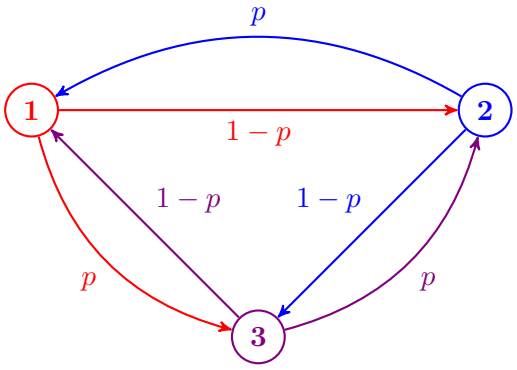
In general, any process where the columns of P *are constant is called a Bernoulli process. This is a special property that makes for a exceptionally simple Markov chain – as we saw above, we reach the stationary distribution in one step! This is because Bernoulli processes do not depend on the current state.*

Definition 5.9.4. *A vector* $\bar{\pi}$ *is an* stationary distribution *if* $\sum_i \bar{\pi}_i = 1$ *and* $\bar{\pi} P = \bar{\pi}$*.*

As we will see, stationary distributions need not be unique, but often we work in settings where they are. When they exist they help us understand the "long-time" dynamics of the Markov chain; that is, what happens to X_n as $n \to \infty$. We call the stationary distribution an *equilibrium distribution* in such a case.

In Example 5.9.3, the interpretation of the stationary distribution and the fact that (5.9.3) is the following. For "long-times," we have probability p of finding the javelina at trash can 1 and we have probability $1-p$ of finding the javelina at trash can 2. We see below that this can get more complicated.

Example 5.9.5. *(i) (Random walk on the 3-cycle) Let us consider* X_n *to be the Markov chain defined on* $S = \{1, 2, 3\}$*. Here, we take steps to the left with probability* $p \in (0, 1)$ *and steps to the right with probability* $1-p$*. Think of* 3 *as being to the "left" of* 1 *so that we have a closed loop.*



In this case, we have the transition matrix

$$P = \begin{pmatrix} 0 & 1-p & p \\ p & 0 & 1-p \\ 1-p & p & 0 \end{pmatrix}.$$

Notice that, letting $\bar{\pi} = [1/3, 1/3, 1/3]$*,*

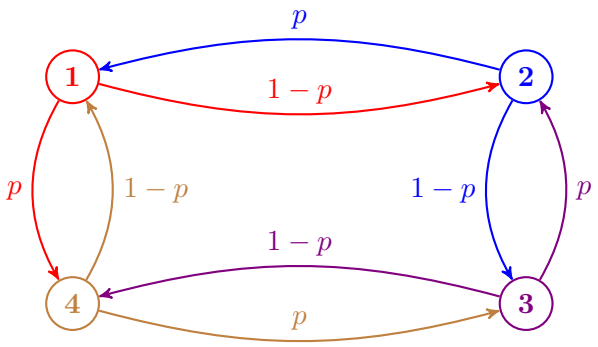
$$\bar{\pi} P = \bar{\pi},$$

that is, $\bar{\pi}$ is an stationary distribution of $X_.$. Moreover, one can show that, for any distribution $\pi^{(0)}$,

$$\lim_{n\to\infty} \pi^{(0)} P^n = \bar{\pi}.$$

In other words, after a large number of steps n, we are (approximately) equally likely to find X_n at any spot $\{1,2,3\}$. We will discuss this at greater length in the future.

(ii) (Random walk on the 4-cycle) Let us consider X_n to be the Markov chain defined on $S = \{1,2,3,4\}$. Here, we take steps to the left with probability $p \in (0,1)$ and steps to the right with probability $1-p$. Think of 4 as being to the "left" of 1 so that we have a closed loop.



In this case, we have the transition matrix

$$P = \begin{pmatrix} 0 & 1-p & 0 & p \\ p & 0 & 1-p & 0 \\ 0 & p & 0 & 1-p \\ 1-p & 0 & p & 0 \end{pmatrix}. \tag{5.9.4}$$

In analogy with the last example, we might guess that $\bar{\pi} = [1/4, 1/4, 1/4, 1/4]$. Indeed, this is an stationary distribution:

$$\bar{\pi} P = \bar{\pi}.$$

Does this mean that after n steps, our process is equally likely to be at any location? Let us consider the specific case where $X_0 = 2$. Then, after one step, we can go "left" to 1 or "right" to 3. Notice that these are both odd so X_1 is odd. Given another step, we find that X_2 is even.

Iterating this, we see that X_n is odd if n is odd and even if n is even. This means, that after n steps we are not equally likely to be at any step – certain states are impossible depending on the parity of n.

Exercise 5.9.2. *Let $\pi^{(0)} = [1,0,0,0]$. Show that $\pi^{(0)} P^{2n} \to [0, 1/2, 0, 1/2]$ and $\pi^{(0)} P^{2n+1} \to [1/2, 0, 1/2, 0]$.*

On the other hand, if we choose our initial state randomly according to $\bar{\pi}$ we do not have such a problem.

The issues in Example 5.9.5.(ii) are related to periodicity:

Definition 5.9.6. *A state i has* period *d if whenever $(P^n)_{ii} > 0$ then n is divisible by d. A Markov chain is called* aperiodic *if all states are not d-periodic for any $d > 0$.*

Roughly, this means that, after enough steps, there is always a positive probability that the Markov chain has made it back to where it started.

Exercise 5.9.3. *Check that our 4-cycle random walk above is 2-periodic and the 3-cycle random walk is aperiodic.*

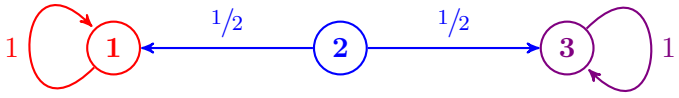
There is a standard way to "fix" periodicity – modifying the Markov chain to give it a $\frac{1}{2}$ probability to stay in its current state. In other words, we consider the related "lazy" Markov chain defined by the transition matrix

$$\tilde{P} = \frac{1}{2}\,\mathrm{id} + \frac{1}{2}P.$$

Example 5.9.7. *(i) (Absorbing boundaries) For any N, let $S = \{1, 2, 3\}$ with the transition matrix*

$$P = \begin{pmatrix} 1 & 0 & 0 \\ 1/2 & 0 & 1/2 \\ 0 & 0 & 1 \end{pmatrix}.$$

In other words, if the process is at one of the endpoints, it is "trapped" there forever. If it is in the middle, it chooses an endpoint at which to be trapped with equal probability.



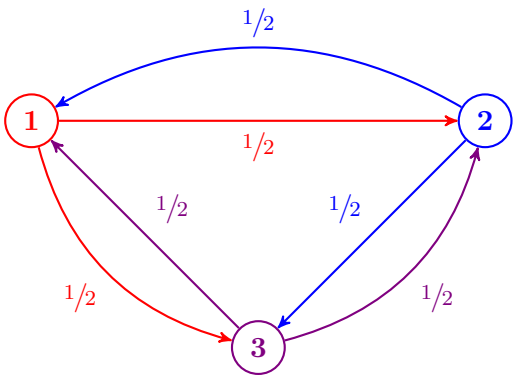
Heuristically, there should be three equilibrium distrbutions – one each for starting from an endpoint and being stuck there forever and one for starting in the middle and ending up stuck at an endpoint with $1/2$ probability each. Indeed, $[1, 0, 0]$, $[1/2, 0, 1/2]$, and $[0, 0, 1]$ are all stationary distributions.

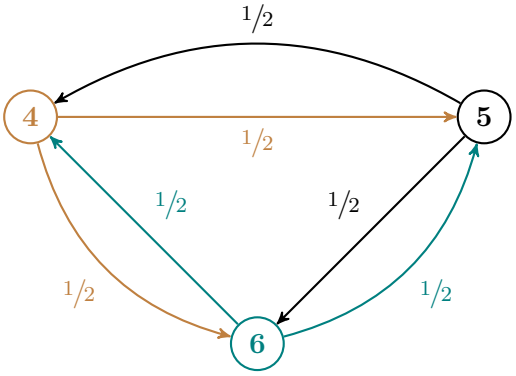
Exercise 5.9.4. *Generalize this to $\{1, 2, \ldots, N\}$ and find all stationary distributions.*

(ii) (Two disjoints 3-cycles) Let $S = \{1, 2, 3, 4, 5, 6\}$ and define

$$P = \begin{pmatrix} Q & 0 \\ 0 & Q \end{pmatrix}$$

where Q is the 3×3 matrix defined in (5.9.4) with $p = 1/2$ and 0 stands for a 3×3 block of zeros. This block diagonal form reflects the fact that our state space is separated: if the random walk starts in $\{1, 2, 3\}$ it can never make it to $\{4, 5, 6\}$ and vice versa.





It should come as no surprise at this point that there are stationary distributions

$$[1/3, 1/3, 1/3, 0, 0, 0], \quad [0, 0, 0, 1/3, 1/3, 1/3], \quad \text{and} \quad [1/6, 1/6, 1/6, 1/6, 1/6, 1/6].$$

The issues with the two above examples are that a process starting at certain states cannot make it to all other states. This motivates the following definition:

Definition 5.9.8. *A Markov chain is* irreducible *if, for every $i, j \in S$, there is n_{ij} such that $P_{ij}^{n_{ij}} > 0$.*

Exercise 5.9.5. *If a Markov chain $X_\cdot$ with finite state space S is aperiodic and irreducible, then there is N such that P^n has all positive entries when $n \geq N$. In other words, $\mathbb{P}(X_n = j | X_0 = i) > 0$ for every $i, j \in S$ and every $n \geq N$.*

The importance of this is that these properties allow us to understand the stationary distribution and the long-time dynamics of $X_\cdot$.

Theorem 5.9.9. *Suppose that $X_\cdot$ is a Markov chain with a finite state space S. Then:*

(i) there exists an stationary distribution $\bar{\pi}$;

(ii) it is unique;

(iii) for any $\pi^{(0)}$, we have $\pi^{(0)} P^n \to \bar{\pi}$ as $n \to \infty$.

Actually, aperiodicity is only need for (iii).

Proof sketch.

Exercise 5.9.6. *Fill in the details here to make a complete proof.*

For any fixed $k \in S$, define

$$V_{j,k} = \#\text{ of time } X_n \text{ visits } j \text{ before returning to } k.$$

One can check that the following is a stationary distribution:

$$\bar{\pi}_j = \frac{1}{\sum_{i=1}^N \mathbb{E}[V_{i,k}]} \mathbb{E}[V_{j,k}],$$

where N is the number of states in S. Note that this yields a nice interpretation of the stationary distribution: $\bar{\pi}_i$ is the fraction of time that the Markov chain spends at the location i.

One can check that, by irreducibility, $\bar{\pi} > 0$ in all entries. Suppose that $\tilde{\pi}$ is another stationary distribution. Let

$$A_0 = \min\{A \in \mathbb{R} : A\bar{\pi} \geq \tilde{\pi}\}.$$

Exercise 5.9.7. *If $A_0 = 1$ then $\bar{\pi} = \tilde{\pi}$.*

Then $A_0\bar{\pi}$ "touches" $\tilde{\pi}$ from above at some ℓ_0; that is $A_0\bar{\pi} \succeq \tilde{\pi}$, and $A_0\bar{\pi}_{\ell_0} = \tilde{\pi}_{\ell_0}$. It follows that

$$((A_0\bar{\pi} - \tilde{\pi})(\mathrm{id} - P))_{\ell_0} = -\sum_j (A_0\bar{\pi}_j - \tilde{\pi}_j)P_{j\ell_0} \leq 0. \tag{5.9.5}$$

The inequality is strict if $A_0\bar{\pi} > \tilde{\pi}$ at any "neighbor" k such that $P_{\ell_0 k} > 0$.

On the other hand,

$$(\mathrm{id} - P)(A_0\bar{\pi} - \tilde{\pi}) = A_0(\mathrm{id} - P)\bar{\pi} - (\mathrm{id} - P)\tilde{\pi} = 0 - 0 = 0$$

since $\bar{\pi}$ and $\tilde{\pi}$ are stationary distributions. Hence, (5.9.5) must be equality, meaning that $A_0\bar{\pi} - \tilde{\pi}$ must be the same at all of ℓ_0's neighbors. One can repeat this argument at each neighbor to see that $A_0\bar{\pi}$ and $\tilde{\pi}$ agree at its subsequent neighbors. By irreducibility, this argument will exhaust all elements of S and yield $A_0\bar{\pi} = \tilde{\pi}$. Since these are probability distributions (and, thus, the entries of $\bar{\pi}$ and $\tilde{\pi} = A_0\bar{\pi}$ must sum to 1), it must be that $A_0 = 1$, showing that $\bar{\pi} = \tilde{\pi}$.

We now show convergence by essentially the same argument. Take any $\pi^{(0)}$ and consider n sufficiently large that $\pi^{(0)}P^n \succ 0$ (by the exercise above). Defining

$$A_1 = \min\{A \in \mathbb{R} : A\pi^{(0)}P^n \succeq \bar{\pi}\}.$$

Then, for every ℓ

$$0 \leq \sum_j (A_1\pi^{(0)}P^n - \bar{\pi})_j P^n_{j\ell} = A_1\pi^{(0)}P^{2n} - \bar{\pi}.$$

On the other hand, P^{2n} has all positive entries so the inequality above is strict unless $\pi^{(0)}P^n = \bar{\pi}$. If this occurs, we are finished. If not, define

$$A_2 = \min\{A \in \mathbb{R} : A\pi^{(0)}P^{2n} \succeq \bar{\pi}\}.$$

It follows that $A_2 < A_1$. Iterating this argument yields $A_1, A_2, \ldots$. It must be that $A_k\pi^{(0)}P^{kn} \to \bar{\pi}$ (otherwise we would have been able to take a "smaller" A at some intermediate step). One can then upgrade this to

$$\lim_{n\to\infty} \pi^{(0)}P^{kn} = \lim_{n\to\infty} A_n\pi^{(0)}P^{kn} = \bar{\pi}.$$

One can repeat this argument for every $\ell \in \{1, \ldots, k-1\}$ to get the convergence of $\pi^{(0)}P^{nk+\ell}$ to $\bar{\pi}$, which completes the proof. □

Remark 5.9.10. *There is a whole field of quantifying the convergence in Theorem 5.9.9.(iii); that is, finding the "mixing time" for a Markov chains. Given the importance of Markov chains and their mixing in applications across virtually all "STEM" fields, this is extremely important work!*

The argument above, with a tiny bit more work, can yield an exponential rate of convergence.

Exercise 5.9.8. *Think about this!*

Example 5.9.11. *An important application of this Google's PageRank algorithm. Here is a cartoon version of it. Suppose there are N websites. Fix website j and let N_j be the number of websites that contain a link to it. Then*

$$p_{ij} = \begin{cases} 0 & \text{if } i \text{ has no link to } j, \\ 1/N_j & \text{if } i \text{ has a link to } j. \end{cases}$$

Then P is a transition matrix. Roughly, this is a Markov chain that randomly traverses the internet without

If we restrict to a slightly smaller subset of websites (ignoring isolated ones), we might suspect that this Markov chain is irreducible and aperiodic[46]*. Then, by Theorem 5.9.9, there is an stationary distribution $\bar{\pi}$.*

The larger $\bar{\pi}_i$ is, the more time our random internet walker spends on website i because more links coming into it, at least in a broad sense[47]*. Hence, we should think of $\bar{\pi}_i$ as giving an ranking of the webpages: the most important website is the website i such that $\bar{\pi}_i$ is greatest and the least important website is the website j such that $\bar{\pi}_j$ is smallest.*

5.9.1. **Duality and the beginnings of connections to partial differential equations.** There is an obvious dual pairing between bounded, continuous functions f and measures μ given by

$$\langle \mu, f \rangle = \int f(x)\, d\mu(x). \tag{5.9.6}$$

It is a fact, that we will not clearly state here or prove, that the dual space of bounded continuous function is the space of "nice" "signed" measures with the dual pairing (5.9.6).

For the Markov chains above, which have a discrete state space, we can take f to be a vector $f = (f_1, \ldots, f_N)$ and we end up with the pairing

$$\langle \pi^{(n)}, f \rangle = \langle \pi^{(0)} P^n, f \rangle = \pi^{(0)} P^n f = \sum_{ij} \pi_i^{(0)} P_{ij}^n f_j.$$

On the other hand, it seems fruitful to think of P as defining an evolution on f

$$f_i^{(n)} := \sum_{j=1}^{N} P_{ij}^n f_j.$$

While $\pi^{(n)}$ is the *forward* evolution of $\pi^{(0)}$, f^n is the *backward* evolution of f. Why? The above becomes

$$\langle \pi^{(n)}, f^{(0)} \rangle = \langle \pi^{(0)}, f^{(n)} \rangle.$$

Hence, evaluating f against a measure after n steps of the Markov chain (forward-in-time) is equivalent to computing $f^{(n)}$ and then evaluating it at the initial time (backward-in-time).

It turns out that this dual pairing is very useful. Let us consider a simple example to illustrate the idea.

[46] We can, of course, just make the Markov chain "lazy" if periodicity becomes a problem.

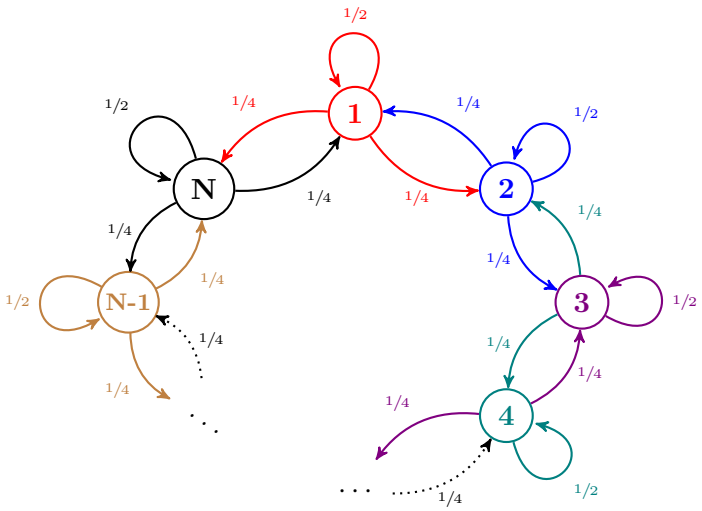
[47] We need to be careful here. There is a subtle point. Suppose there is one "mega-website" k that has a link from every website. But its only link out is to a small website ℓ that has no other incoming links. We might, at first, suspect that website ℓ is not important (it only has *one* incoming link; however, this is incorrect since it will get all of the traffic from mega-website k.

Example 5.9.12. *Consider the "lazy" random walk on an N-cycle with $N \gg 1$: for $i, j \in \{1, 2, \dots, N\}$, let*

$$P_{ij} = \begin{cases} \frac{1}{4} & \text{if } |i-j| = 1, \\ \frac{1}{2} & \text{if } i = j, \\ 0 & \text{otherwise.} \end{cases}$$

Here, we take the convention that $|1 - N| = 1$ with

$$1 - 1 = N \qquad \text{and} \qquad N + 1 = 1. \tag{5.9.7}$$



Fix any continuous function $f : [0, 1] \to \mathbb{R}$ and define

$$f_N = (f(1/N), f(2/N), \dots, f(1)).$$

Then define, for any $i \in \{1, 2, \dots, N\}$,

$$U_N(n, i) = e_i \left(P^n f_N\right).$$

Here we have begun a random walk at i and evolved it for n steps (this is represented by the $e_i P^n$ term) and then evaluated f_N at that point. Let us write this in an equivalent, but perhaps more straightforward, way:

$$U_N(n, i) = \mathbb{E}\left[f(X_n) | X_0 = i\right] = \sum_{j=1}^{N} P_{ij}^n (f_N)_j,$$

where X is the random walk associated to P. Let us note that

$$U_N(0, i) = f\left(i/N\right). \tag{5.9.8}$$

Let us see how u evolves in time. First, notice that

$$\begin{aligned} U_N(n+1, i) &= \sum_{j=1}^{N} \sum_{k=1}^{N} P_{ik} P_{kj}^n (f_N)_j = \sum_{j=1}^{N} \sum_{|k-j| \le 1}^{N} P_{ik} P_{kj}^n (f_N)_j \\ &= \frac{1}{4} \sum_{j=1}^{N} P_{(i+1)j}^n (f_N)_j + \frac{1}{4} \sum_{j=1}^{N} P_{(i-1)j}^n (f_N)_j + \frac{1}{2} \sum_{j=1}^{N} P_{ij}^n (f_N)_j \\ &= \frac{U_N(n, i+1) + U_N(n, i-1) + 2U_N(n, i)}{4}. \end{aligned}$$

Roughly this represents the fact that the chain can arrive at i after $n+1$ steps only if it is at $i+1$, $i-1$, or i after n steps and the occur with probability $1/4$, $1/4$, and $1/2$, respectively. Rewriting this, we see that

$$U_N(n+1,i) - U_N(n,i) = \frac{1}{4}\Delta_{\text{disc}} U_N(n,i),$$

where we have defined the discrete Laplacian (second derivative) as

$$\Delta_{\text{disc}}\varphi(i) = \varphi(i+1) + \varphi(i-1) - 2\varphi(i).$$

Of course, the left hand side looks like a discrete time derivative, so we have arrived at a discrete heat equation.

Let us now shrink this down to a common domain. For any $t \in \{0, 1/N^2, 2/N^2, \dots\}$ and any $x \in \{1/N, 2/N, \dots, 1\}$, we let

$$u_N(t,x) = U_N(tN^2, xN), \tag{5.9.9}$$

where U_N is defined above. Then, from the above, we see that

$$\frac{u_N(t+N^{-2},x) - u_N(t,x)}{N^{-2}} = \frac{1}{4}\frac{u_N(t,x+N) + u_N(t,x-N) - 2u_N(t,x)}{N^{-2}}.$$

At least formally, we see that the limit $N \to \infty$ should lead to a limiting function u that solves

$$\begin{cases} \partial_t u = \frac{1}{4}\Delta u & \text{on } [0,1], \\ u(t,0) = u(t,1), \\ u(0,x) = f(x). \end{cases} \tag{5.9.10}$$

The last equality follows from (5.9.8) *with the change of variables* (5.9.9)*. The second-to-last equality follows from the identification of* 1 *and* N *as neighbors* (5.9.7)*.*

With a bit more work, one can make this limit make sense. Additionally, if one assumes that by scaling the random walk in the way we did above (time like n/N^2 and space like i/N) then we should arrive at a Brownian motion (up to a time scaling factor of $1/2$ due to the "laziness"), one arrives at a basic version of the famous Feynman-Kac formula: the solution of (5.9.10) *is given by*

$$u(t,x) = \mathbb{E}[f(B_{t/2})|B_0 = x],$$

where $B_\cdot$ is a standard Brownian motion. That X_n, scaled as above, is an approximation of Brownian motion is Donsker's theorem*, which we have run out of time to cover.*

6. Convex Analysis

6.1. Some basic objects.

6.1.1. Affine, conic, and convex combinations and sets.

Definition 6.1.1. *Given a vector space V, vectors $v_1, \dots, v_k$, and scalars $\theta_1, \dots, \theta_k \in \mathbb{R}$, a linear combination $\theta_1 v_1 + \cdots \theta_k v_k$ is:*

1. *an* affine combination *if $\theta_1 + \theta_2 + \cdots + \theta_k = 1$;*

2. *a* conic combination *if* $\theta_i \geq 0$*, for all* i*,;*

3. *a* convex combination *if it is an affine combination and a conic combination.*

Exercise 6.1.1. *Take three non-colinear points* v_1, v_2, v_3 *in* $\mathbb{R}^3$*, these define a cone (with a vertex at the origin). Convince yourself that any conic combination* $\theta_1 v_1 + \theta_2 v_2 + \theta_3 v_3$ *lies inside this cone. These three points also define a triangle. Convince yourself that any convex combination lies inside this triangle. How does this latter example generalize to more points?*

Definition 6.1.2. *A set* $C \subset V$ *is:*

1. affine *if it is closed under affine combinations (that is, if* $v_1, \ldots, v_k \in C$*, then so is* $\theta_1 v_1 + \cdots \theta_k v_k$ *if* $\theta_1 + \theta_2 + \cdots + \theta_k = 1$*);*

2. conic *if it is closed under conic combinations (that is, if* $v_1, \ldots, v_k \in C$*, then so is* $\theta_1 v_1 + \cdots \theta_k v_k$ *if* $\theta_i \geq 0$*, for all* i*);*

3. convex *if it is closed under convex combinations (that is, if* $v_1, \ldots, v_k \in C$*, then so is* $\theta_1 v_1 + \cdots \theta_k v_k$ *if* $\theta_1 + \cdots + \theta_k = 1$ *and* $\theta_i \geq 0$ *for all* i*).*

We point out that the arbitrary intersection of affine sets is affine, the arbitrary intersection of conic sets is conic, and the arbitrary intersection of convex sets is convex.

Let's characterize these a bit.

Proposition 6.1.3. *A nonempty set* $A \subset V$ *is affine if and only if* $A = W + x_0$ *for some* $x_0 \in V$ *and some subspace* $W \subset V$*.*

Before proving this, let us note that this allows us to assign dimension so an affine space: the *affine dimension* of an affine set A is the dimension of the subspace W given by Proposition 6.1.3.

Proof. We prove each part one at a time.

Proof of $\Longrightarrow$: Take any point $x_0 \in V$. We need only show that the set

$$W := A - x_0 = \{w \in V : w = a - x_0 \text{ for some } a \in A\}$$

is a subspace. To that end, we first note that $0 \in W$ since $0 = x_0 - x_0 \in A - x_0$. Next, we show that W is closed under addition and scalar multiplication: take $w_1, w_2 \in W$, with $w_i = a_i - x_0$, and $\alpha, \beta \in \mathbb{R}$, then

$$\alpha w_1 + \beta w_2 + x_0 = \alpha(a_1 - x_0) + \beta(a_2 - x_0) + x_0 = \alpha a_1 + \beta a_2 - (\alpha + \beta - 1)x_0.$$

Let $\theta_1 = \alpha$, $\theta_2 = \beta$, and $\theta_3 = 1 - \alpha - \beta$. Then

$$\alpha w_1 + \beta w_2 + x_0 = \theta_1 a_1 + \theta_2 a_2 + \theta_3 x_0 \in A,$$

where the inclusion follows from the fact that the above is an affine combination of vectors in the affine set A. Hence, we have that

$$\alpha w_1 + \beta_2 w \in A - x_0 = W.$$

It follows that W is a subspace.

Proof of $\Longleftarrow$: This is somewhat straightforward. We check it anyways. Fix an affine combination $\theta_1 v_1 + \cdots + \theta_n v_n$ where $v_1, \ldots, v_n \in A = x_0 + W$. For each i, let $w_i = v_i - x_0 \in W$, where the inclusion follows by assumption. Then

$$\begin{aligned}\theta_1 v_1 + \cdots + \theta_n v_n &= \theta_1(x_0 + w_1) + \cdots + \theta_n(x_0 + w_n) \\ &= \underbrace{(\theta_1 + \cdots + \theta_n)}_{=1} x_0 + \underbrace{\theta_1 w_1 + \cdots + \theta_n w_n}_{\in W} \in x_0 + W = A.\end{aligned}$$

□

What does this tell us? Let's work with $A \subset \mathbb{R}^n$, for simplicity. Proposition 6.1.3 shows that an affine space A corresponds to the solution space of a matrix M. Let's work with $A \subset \mathbb{R}^n$, for simplicity. Let $m_1, \ldots, m_k$ be a basis for the space W as in Proposition 6.1.3. Let M be the matrix whose row vectors are $m_1, \ldots, m_k$ and let $b = Mx_0$ (where x_0 is as in Proposition 6.1.3), then we can characterize A as

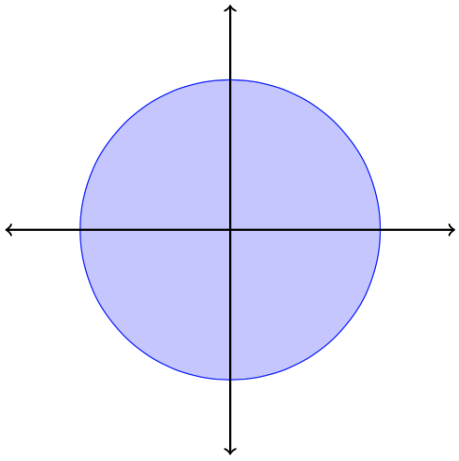
$$A = \{v \in \mathbb{R}^n : Mv = b\}.$$

Useful intuition:

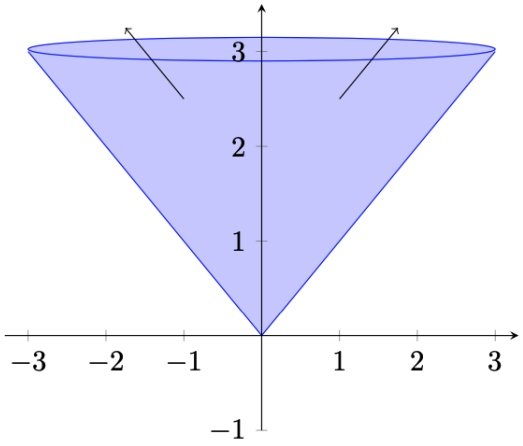
- An affine set A is the translation of a linear subspace. This is the **primal** viewpoint.
- An affine set A is the solution set for a system of linear equations. This is the **dual** viewpoint.

A useful thing to try is to switch between these viewpoints – use the perspective that makes your problem easiest.

Example 6.1.4. *(i) The closed unit ball $B_r = \{x \in C : \|x\| \leq r\}$ of radius $r > 0$ is convex.*



(ii) Take the vector space $V \times \mathbb{R}$ and consider the set $C = \{(v, t) \in V \times \mathbb{R}_+ : \|v\| \leq t\}$ is both conic and convex.



From here one, we specialize to $\mathbb{R}^n$, although it is clear that many of the arguments, ideas, strategies, etc, extend to infinite dimensions.

Definition 6.1.5 (Half-space)**.** *A set of the form* $\{x \in \mathbb{R}^n : x \cdot a \leq b\}$ *for some fixed* $a \in \mathbb{R}^n$ *and* $b \in \mathbb{R}$ *is a* half-space.

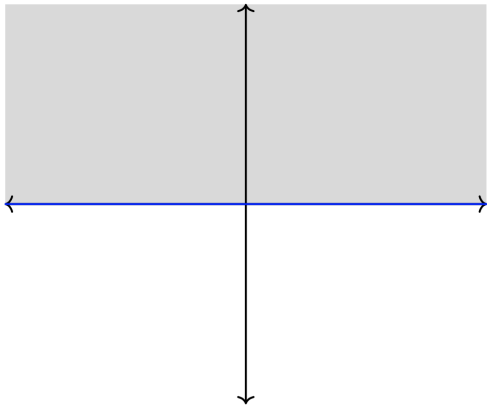
Definition 6.1.6 (Hyperplane)**.** *A set of the form* $H = \{x \in \mathbb{R}^n : x \cdot a = b\}$ *for some fixed* $a \in \mathbb{R}^n$ *and* $b \in \mathbb{R}$ *is a* hyperplane.

We point out two things:

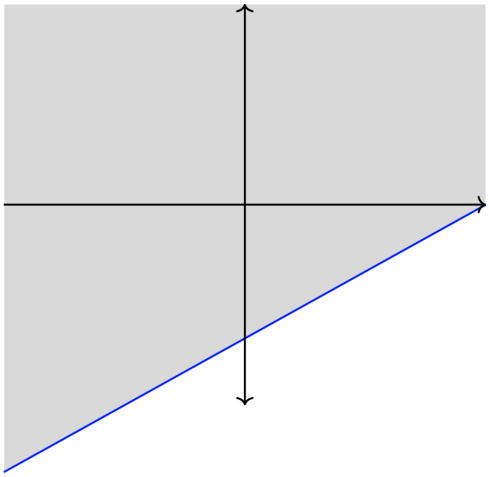
- A hyperplane is an affine set (it is also convex!).
- A half-space is all points to one side of a hyperplane. This why we call it a half-space! The hyperplane cuts $\mathbb{R}^n$ in half, and the half-space is one of the two pieces resulting from this cut.

Let's look at two examples.

Example 6.1.7. *(i) The set* $\mathbb{H} = \{\bar{x} \in \mathbb{R}^n : -x_n \leq 0\}$ *is the traditional half-space. Notice that it sits above the affine space*



(ii) The set $\{(x_1, x_2) \in \mathbb{R}^2 : -2x_1 + 3x_2 \geq 6\} = \{(x_1, x_2) \in \mathbb{R}^2 : 2x_1 - 3x_2 \leq 6\}$ *is a half-space. It is the space of points above the affine space* $(2, -3)^{\perp} + (3, 0)$.



We point out that the half space $\{(x_1, x_2) \in \mathbb{R}^2 : 2x_1 - 3x_2 \geq 6\}$

How does one define these concepts in an arbitrary space? Well, a dot product is the quintessential bounded linear functional (in fact, in 'certain' normed vector spaces, it is the *only* bounded linear functional), so we should use that in place. If V is an infinite dimensional normed linear space, a hyperplane takes the form

$$\{v \in V : \lambda v = 0\}$$

and a half-space takes the form

$$\{v \in V : \lambda v \leq 0\}$$

for some fixed $\lambda \in V^*$.

Returning to the concepts of affine, conic, and convex, we can, given a collection of points, find the "smallest" set of each type that contains all such points. Let us define these here and then check their properties.

Definition 6.1.8 (Affine hull)**.** *Given* $P \subset V$*, let*

$$\operatorname{aff}(P) = \left\{ v \in V : v = \sum_{i=1}^{n} \theta_i p_i \text{ for some } n \in \mathbb{N}, p_1, \ldots, p_n \in P, \theta_1, \ldots, \theta_n \in \mathbb{R} \text{ such that } \sum_{i=1}^{n} \theta_i = 1 \right\}.$$

We point out that, by definition, if A is an affine set and $P \subset A$, then $\operatorname{aff}(P) \subset A$. Hence, our characterization of $\operatorname{aff}(P)$ as the smallest affine set containing P is accurate once we show that $\operatorname{aff}(P)$ is actually affine!

Proposition 6.1.9. *Given any* $P \subset V$*, the set* $\operatorname{aff}(P)$ *is affine. Hence,* $\operatorname{aff}(P)$ *is the smallest affine set containing* P*, and*

$$\operatorname{aff}(P) = \bigcap_{A \supset P, A \text{ is affine}} A. \tag{6.1.1}$$

Proof. We only show that $\operatorname{aff}(P)$ is affine. The equality (6.1.1) follows from this without too much extra work.

Exercise 6.1.2. *Show this!*

Fix any collection of points $v_1, \dots, v_n \in \text{aff}(P)$ and let $\theta_1, \dots, \theta_n \in \mathbb{R}$ be such that $\theta_1 + \cdots + \theta_n = 1$. By definition of $\text{aff}(P)$, there are, for each k, n_k, $p_{k,1}, \dots, p_{k,n_k}$, $\theta_{k,1}, \dots, \theta_{k,n_k}$ such that

$$v_k = \theta_{k,1}p_{k,1} + \cdots + \theta_{k,n_k}p_{k,n_k} \quad \text{and} \quad 1 = \theta_{k,1} + \cdots + \theta_{k,n_k}.$$

Then

$$\theta_1 v_1 + \cdots + \theta_n v_n = \sum_{k=1}^{n} \theta_k \left(\sum_{j=1}^{n_k} \theta_{k,j} p_{k,j} \right) = \sum_{k=1}^{n} \sum_{j=1}^{n_k} \theta_k \theta_{k,j} p_{k,j}.$$

Thus, the right hand side is a linear combination of vectors in P with coefficients $\theta_i \theta_{k,j}$. Hence, we are finished if we show that these coefficients sum to one. We do that here to conclude the proof:

$$\sum_{k=1}^{n} \sum_{j=1}^{n_k} \theta_k \theta_{k,j} = \sum_{k=1}^{n} \theta_k \sum_{j=1}^{n_k} \theta_{k,j} = \sum_{k=1}^{n} \theta_k (\theta_{k,1} + \cdots + \theta_{k,n_k}) = \sum_{k=1}^{n} \theta_k = \theta_1 + \cdots + \theta_n = 1.$$

□

Definition 6.1.10 (Conic hull)**.** *Given $P \subset V$, let*

$$\text{conic}(P) = \Big\{ v \in V : v = \sum_{i=1}^{n} \theta_i p_i \text{ for some } n \in \mathbb{N}, p_1, \dots, p_n \in P, \\ \theta_1, \dots, \theta_n \in \mathbb{R} \text{ such that, for all } i,\ \theta_i \geq 0 \Big\}.$$

We point out that, by definition, if C is a conic set and $P \subset C$, then $\text{conic}(P) \subset C$. Hence, our characterization of $\text{conic}(P)$ as the smallest conic set containing P is accurate once we show that $\text{conic}(P)$ is actually conic!

Proposition 6.1.11. *Given any $P \subset V$, the set* $\text{conic}(P)$ *is conic. Hence,* $\text{conic}(P)$ *is the smallest conic set containing P, and*

$$\text{conic}(P) = \bigcap_{C \supset P, C \text{ is conic}} C.$$

Exercise 6.1.3. *Prove this!*

Example 6.1.12. *Let $P = \{(1,0),(0,1)\}$. Then* $\text{conic}(P) = [0,\infty)^2$.

Definition 6.1.13 (Convex hull)**.** *Given $P \subset V$, let*

$$\text{conv}(P) = \Big\{ v \in V : v = \sum_{i=1}^{n} \theta_i p_i \text{ for some } n \in \mathbb{N}, p_1, \dots, p_n \in P, \\ \theta_1, \dots, \theta_n \in \mathbb{R} \text{ such that } \sum_{i=1}^{n} \theta_i = 1 \text{ and } \theta_i \geq 0 \text{ for all } i \Big\}.$$

We point out that, by definition, if C is a convex set and $P \subset C$, then $\text{conv}(P) \subset C$. Hence, our characterization of $\text{conv}(P)$ as the smallest convex set containing P is accurate once we show that $\text{conv}(P)$ is actually convex!

Proposition 6.1.14. *Given any $P \subset V$, the set* $\text{conv}(P)$ *is convex. Hence,* $\text{conv}(P)$ *is the smallest convex set containing P, and*

$$\text{conv}(P) = \bigcap_{C \supset P, C \text{ is convex}} C.$$

Exercise 6.1.4. *Prove this!*

Example 6.1.15. *Let $P = \{-1,1\}^3$. Then* $\text{conv}(P)$ *is the cube* $[-1,1]^3$.

6.1.2. **Polyhedra.**

Definition 6.1.16. *A set $P \subset V$ is a* polyhedron *if there are vectors $a_1, \dots, a_n$, $c_1, \dots, c_m$, and scalars $b_1, \dots, b_n$, $d_1, \dots, d_m$ such that*

$$P = \{x \in V : a_i \cdot x \leq b_i \ \forall i = 1, \dots, m, \text{ and } c_i \cdot x = d_i \ \forall i = 1, \dots, k\}.$$

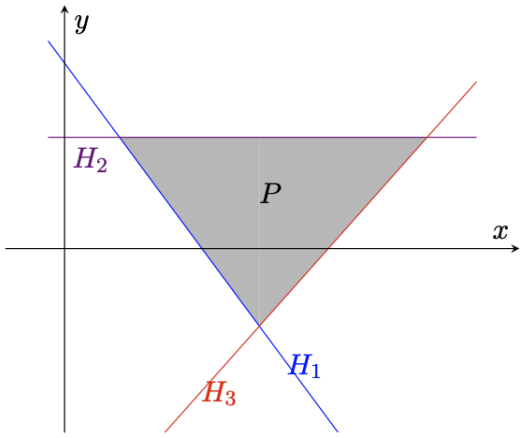
In other words, P is the intersection of a finite number of half-spaces and hyperplanes (or, equivalently, the space between a finite number of hyperplanes intersected with a finite number of hyperplanes).

A more compact way to write this is:

$$P = \{x \in V : Ax \preceq b, Cx = d\},$$

where $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$, $C \in \mathbb{R}^{k \times n}$, $d \in \mathbb{R}^k$, and we define

$$v \preceq w \quad \text{if and only if} \quad v_i \leq w_i \ \text{ for all } i.$$



Exercise 6.1.5. *Any polyhedron is convex.*

One can see that polyhedra are not necessarily bounded. When they are we call them *polytopes*.

Example 6.1.17. *The unit cube $[-1, 1]^3$ is a polytope defined by the choice $n = 6$, $m = 0$, $a_i = e_i$ when $i = 1, 2, 3$, $a_i = -e_{i-3}$ when $i = 4, 5, 6$, and $b_i = 1$ for all i.*

We point out that Definition 6.1.16 is *dual*. What is the primal characterization of a polytope?

Proposition 6.1.18. *A set P is a polytope if and only if there are $n \in \mathbb{N}$ and $v_1, \dots, v_n \in V$ such that*

$$P = \{\bar{v} \in V : v \text{ is a convex combination of } v_1, \dots, v_n\}.$$

Exercise 6.1.6. *Prove this! Heuristically, the vectors $v_1, \dots, v_n$ are the vertices of the polyhedron. The "meat" of the proof is then determining the vertices and making this heuristic rigorous.*

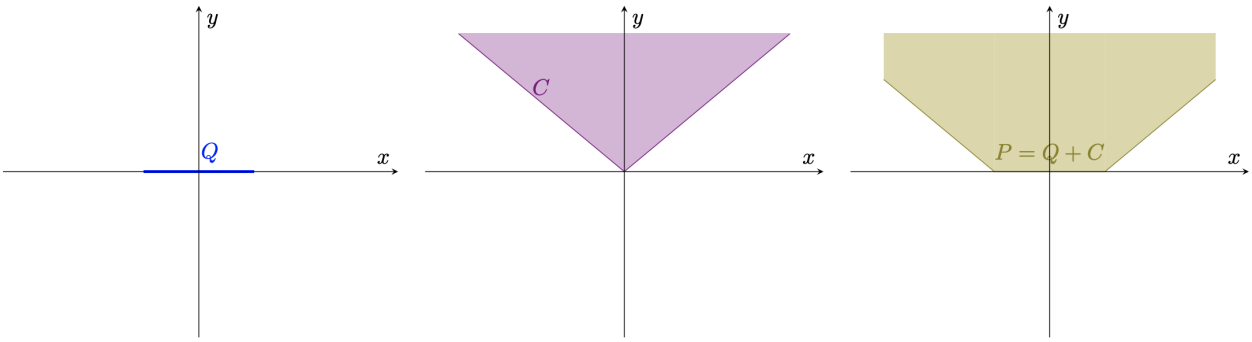
Exercise 6.1.7. *Every polyhedron is of the form*

$$P = Q + C = \{v \in V : v = q + c \text{ for some } q \in Q, c \in C\},$$

where $Q = \text{conv}(q_1, \dots, q_k)$ is a convex set and $C = \text{conic}(c_1, \dots, c_m)$ is a conic set. The polyhedron P is a polytope if $C = \emptyset$.

Example 6.1.19. $P = \{(x, y) \in \mathbb{R}^2 : y \geq 0, y \geq x - 1, y \geq -x - 1\}$

Notice that $P = Q + C = [-1, 1] \times \{0\} + \{|y| \geq x\}$.



6.2. Operations preserving convexity.

Notation 6.2.1. *To be consistent with the book by Boyd and Vandenberghe, we adopt the following notation. Some of it is standard* ($\mathbb{R}$, $\mathbb{R}_+$)*, some of it is nonstandard* ($\mathbb{R}_{++}$)*, and some conflicts*[48] *with standard notation* ($\mathbb{S}^n$, $\mathbb{S}^n_+$)*.*

- $\mathbb{R}$ = *the real numbers;*
- $\mathbb{R}_+ = [0, \infty)$*;*
- $\mathbb{R}_{++} = (0, \infty)$*;*
- $\mathbb{R}^n_+ = [0, \infty)^n = \{\bar{x} \in \mathbb{R}^n : x_i \geq 0$ *for all* $i = 1, \ldots, n\}$*;*
- $\mathbb{R}^n_{++} = (0, \infty)^n = \{\bar{x} \in \mathbb{R}^n : x_i > 0$ *for all* $i = 1, \ldots, n\}$*;*
- $M_{m \times n}$ = *the set of* $m \times n$ *matrices;*
- $\mathbb{S}^n$ = *the set of symmetric* $n \times n$ *matrices;*
- $\mathbb{S}^n_+$ = *the set of symmetric, non-negative definite* $n \times n$ *matrices;*
- $\mathbb{S}^n_{++}$ = *the set of symmetric, positive definite* $n \times n$ *matrices.*

Recall 6.2.2. *A matrix* $M \in \mathbb{S}^n$ *if* $M = M^T$*. It is in* $M \in \mathbb{S}^n_+$ *if, additionally,* $x \cdot Mx \geq 0$ *for all* $x \in \mathbb{R}^n$*. Finally,* $M \in S^n_{++}$ *if, additionally,* $x \cdot Mx > 0$ *for all* $x \in \mathbb{R}^n \setminus \{0\}$*.*

The last inequality can also be written as $x^T Mx$ *if we think of* x *as an* $n \times 1$ *matrix.*

Exercise 6.2.1. *Show that* $\mathbb{S}^n_+$ *is conic.*

What operations preserve convexity?

- **Intersections:** Intuitively this is somewhat clear – if two sets have no "holes" or "indents," then their intersection cannot as well. More precisely, suppose that $\mathcal{I}$ is an indexing set and, for each $\iota \in \mathcal{I}$, the set $A_\iota \in \mathbb{R}^n$ is convex. Then

$$A = \cap_{\iota \in \mathcal{I}} A_\iota$$

  is convex as well.

[48] Usually $\mathbb{S}^n = \{\bar{x} \in \mathbb{R}^{n+1} : \|\bar{x}\| = 1\}$; that is, $\mathbb{S}^n$ is the unit sphere in $n + 1$ dimensions. We use n because it is an n-dimensional manifold. (Think about this carefully, a circle, in two dimensions, is a line "wrapped up," so it is essentially a one dimensional object living in two dimensions. Hence, we call it $\mathbb{S}^1$.)

- **Images of an affine transformation:** A map $f : \mathbb{R}^n \to \mathbb{R}^m$ is *affine* if $f(x) = Ax + b$ where $A \in M_{m\times n}$. Then, if $C \subset \mathbb{R}^n$ is convex, so is

$$f(C) = \{y \in \mathbb{R}^m : y = f(x) \text{ for some } x \in \mathbb{R}^n\}.$$

  Why should this be true? Affine maps preserve linear combinations! Note: affine transformations also take affine sets to affine set.

  Exercise 6.2.2. *Show that this is not true for conic sets. What conditions can you place on f that make the image of a conic set conic?*

- **Inverse image of an affine transformation:** If $C \subset \mathbb{R}^m$ is convex and f is affine, then

$$f^{-1}(C) = \{x \in \mathbb{R}^n : f(x) \in C\}$$

  is convex. Why should this be true? Affine maps preserve linear combinations! Note: affine transformations also take affine sets to affine set.

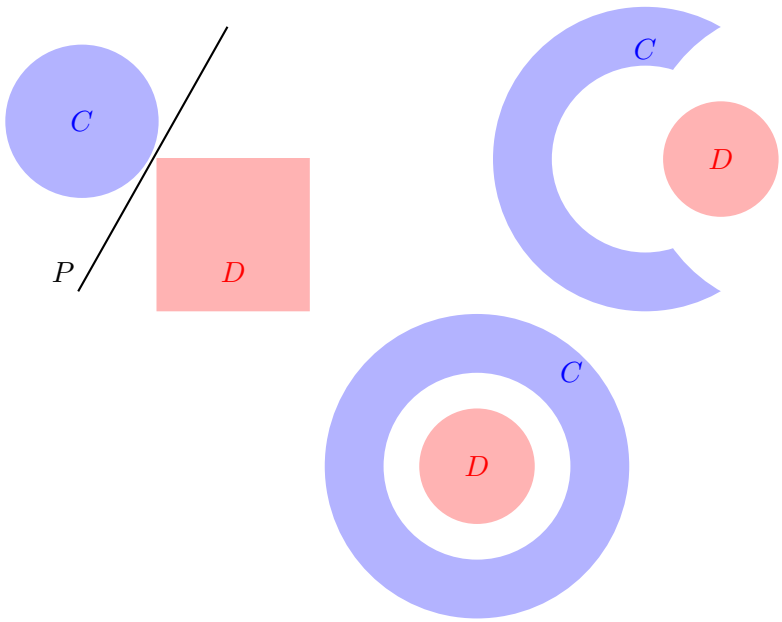
  Exercise 6.2.3. *Show that this is not true for conic sets. What conditions can you place on f that make the image of a conic set conic?*

- **Perspective function:** Let $f : \mathbb{R}^n \times \mathbb{R}_{++} \to \mathbb{R}^n$ be defined by $f(x,t) = x/t$. This function takes line segments to line segments. Roughly, this is what you do when you are trying to draw a three dimensional picture on a two dimensional piece of paper – you scale the far away objects ($t \gg 1$) to be smaller and the close objects ($t \ll 1$) to be larger.

  If $C \subset \mathbb{R}^n \times \mathbb{R}_+ +$ and $K \subset \mathbb{R}^n$ are convex, then so are $f(C)$ and $f^{-1}(K)$.

Exercise 6.2.4. *Prove all of the claims above!*

6.2.1. **Separating hyperplanes.** Often it is useful to separate sets C and D by a hyperplane P. We see in the below examples that this is not always possible.



It turns out that convexity is enough to guarantee this separation.

Proposition 6.2.3. *If $C, D \subset \mathbb{R}^n$ are nonempty disjoint convex sets, then there exists a hyperplane $P = \{x \in \mathbb{R}^n : a \cdot x = b\}$ such that $C \subset \{x \in \mathbb{R}^n : a \cdot x \leq b\}$ and $D \subset \{x \in \mathbb{R}^n : a \cdot x \geq b\}$.*

We call P the separating hyperplane. We note that the inequalities are *non-strict* (they are only strict with stronger assumptions like compactness). We show the proof only in this stronger case, but it is not hard to deduce the full result by a limiting argument (try it!).

Proof. If C, D are compact, then there are $c_0 \in C$ and $d_0 \in D$ such that

$$|c_0 - d_0| = \operatorname{dist}(C, D).$$

(One can see this, e.g., by applying the extreme value theorem applied to $\rho : C \times D \to \mathbb{R}_+$ defined by $\rho(c, d) = |c - d|$.)

We then take $a = d - c$ and $b = (|d|^2 - |c|^2)/2$, so that

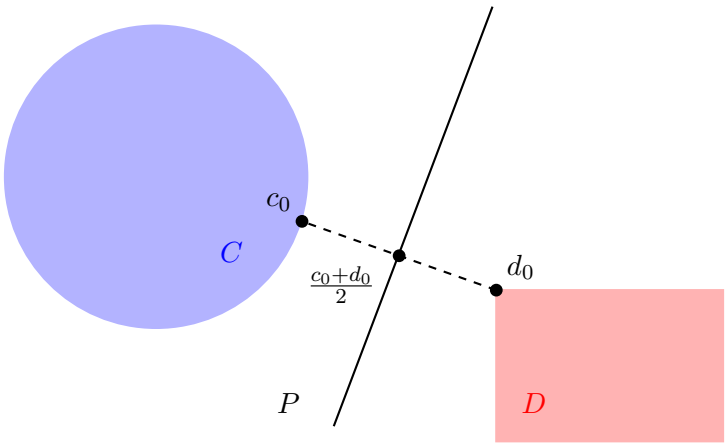
$$P = \left\{x \in \mathbb{R}^n : (d_0 - c_0) \cdot x = \frac{1}{2}(|d_0|^2 - |c_0|^2)\right\}.$$

This is a hyperplane perpendicular to the line connecting c_0 and d_0 that is centered at $\frac{c_0+d_0}{2}$. This can, perhaps, be seen more clearly by rewriting:

$$f(x) := a \cdot x - b = (d_0 - c_0)\Big(x - \frac{1}{2}(d_0 + c_0)\Big).$$

We point out one other useful rewriting:

$$f(x) = a \cdot x - b = (d_0 - c_0) \cdot (x - d_0) + \frac{1}{2}|d_0 - c_0|^2. \tag{6.2.1}$$



We show the inclusion $D \subset \{x \in \mathbb{R}^n : a \cdot x > b\}$ only. The other inclusion is proved similarly since the roles of C and D are essentially symmetric.

Suppose that $\min_D f \leq 0$. By the extreme value theorem applied to $f|_D$, we can find $d_1 \in f$ such that $f(d_1) \leq 0$. Parametrize the line segment connecting d_0 and d_1 by

$$d_t = d_0 + t(d_1 - d_0) = (1 - t)d_0 + td_1 \in D \qquad \text{for } t \in [0, 1].$$

The inclusion in D follows by the convexity of D. Then

$$\frac{d}{dt}|d_t - c_0|^2\Big|_{t=0} = 2(d_0 - c_0)\cdot(d_1 - d_0) = 2f(d_1) - |d_0 - c_0|^2 < 0,$$

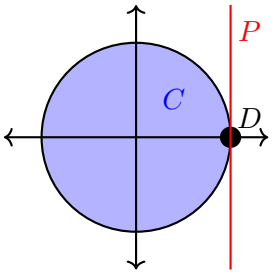
where the last equality uses (6.2.1) and the last inequality follows from the fact that $f(d_1) \leq 0$ and $d_0 \neq c_0$. We conclude that there is some t small but positive such that

$$|d_t - c_0|^2 < |d_0 - c_0|^2 = \operatorname{dist}(C, D)^2.$$

This is a contradiction because $d_t \in D$ and, hence, $|d_t - c_0| \geq \operatorname{dist}(C, D)$. □

Example 6.2.4. *(i)* **(Inequalities do not have to be strict):**

$$C = \{(x, y) : x^2 + y^2 < 1\} \quad \textit{and} \quad D = \{(1, 0)\}.$$



Since there is no "separation" between them, the best we can do is the hyperplane

$$P = \{1\} \times \mathbb{R} = \{(x, y) : x = y\}.$$

Then

$$C \subset \{(x, y) : x \leq 1\} \quad \textit{and} \quad D \subset \{(x, y) : x \geq 1\}.$$

Notice that the $\leq$ and $\geq$ cannot be replaced by $<$ and $>$, respectively. Notice that this also shows that C **and** *D* **do not need to be disjoint to be separated by a hyperplane**.

(ii) **Two sets do not need to be convex to be separated by a hyperplane:** *This is somewhat obvious, because one can easily draw many examples. A simple example is:*

$$C = \{(x, y) : (x - 10)^2 + y^2 = 1\}, \quad D = \{(x, y) : (x + 10)^2 + y^2 = 1\}, \quad \textit{and} \quad P = \{0\} \times \mathbb{R}.$$

Above we see some limitations to an appropriate converse to Proposition 6.2.3. Let us try to find a partial converse.

Definition 6.2.5. *The boundary of a set C in a metric space (X, d) is:*

$$\partial C = \bar{C} \setminus \operatorname{Int} C.$$

Example 6.2.6. *In the metric space $\mathbb{R}$:*

$$\partial(a, b) = \partial[a, b] = \{a, b\}, \quad \partial\{x_1, \ldots, x_n\} = \{x_1, \ldots, x_n\}, \quad \textit{and} \quad \partial\mathbb{Q} = \mathbb{R}.$$

In the metric space $\mathbb{R}^n$

$$\partial B_r(\bar{a}) = \{\bar{x} \in \mathbb{R}^n : |\bar{x} - \bar{a}| = r\}.$$

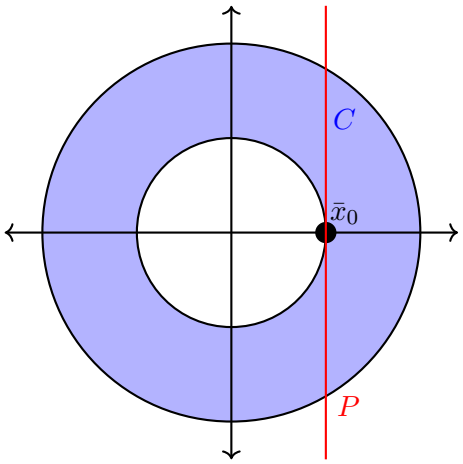
Definition 6.2.7. *Given a set $C \subset \mathbb{R}^n$ and a point $x_0 \in \partial C$, P is a* supporting hyperplane *of C at x_0 if P is a separating hyperplane of C and $\{x_0\}$.*

Example 6.2.8. • *The separating hyperplane P in Example 6.2.4 is a supporting hyperplane of $C = \overline{B_1(0)}$ at $(1, 0)$.*

- **Supporting hyperplanes need not exists:** *When C is not convex, we may not have a supporting hyperplane. Indeed, let*

$$C = \left\{(x, y) \in \mathbb{R}^2 : \frac{1}{2} \leq x^2 + y^2 \leq 1\right\}$$

Then $\bar{x}_0 = (1/2, 0) \in \partial C$, but the hyperplane tangent to $x^2 + y^2 = 1/2$ (the only possibility for a supporting hyperplane) is $P = \{(1/2, y) : y \in \mathbb{R}\}$. Clearly P is not a separating hyperplane!



- **Non-uniqueness of supporting hyperplanes:** *Let*

$$C = [-1, 1] \times [-1, 1]$$

and consider $x_0 = (1, 1) \in \partial C$. Then, for any $r \leq 0$, the hyperplane

$$P = \{(1, 1) + (x, rx) : x \in \mathbb{R}\} = \{(x, y) : rx - y = r - 1\}$$

is a supporting hyperplane at x_0. Hence, there are infinitely many supporting hyperplanes.

Proposition 6.2.9. *If $C \subset \mathbb{R}^n$ is a convex set and $x_0 \in \partial C$, then a supporting hyperplane exists.*

Proof. We assume without loss of generality that $x_0 = 0$. Indeed, were this not the case, we can simply shift all arguments by x_0.

The easy case is when $\operatorname{Int} C \neq \emptyset$. Then the proof is finished by applying Proposition 6.2.3 to $\operatorname{Int} C$ (which is convex) and $\{0\}$.

Now let us consider the case when $\operatorname{Int} C = \emptyset$. First, we claim that, in this case, there is a hyperplane P such that $\overline{C} \subset P$. Notice that P would be a separating hyperplane, so, were we to show this inclusion, our proof would be finished.

We argue by contradiction assuming that C is not contained in any hyperplane. By the contradictory assumption, C is not contained in a proper subspace of $\mathbb{R}^n$. Hence, $\operatorname{Span}(C) = \mathbb{R}^n$; that is, there are n linearly independence vectors $c_1, \ldots, c_n \in C$. We claim that

$$c_{\text{mean}} := \frac{1}{2}0 + \frac{1}{2n}(c_1 + \cdots + c_n) \in \operatorname{Int} C. \tag{6.2.2}$$

Let M be the matrix whose columns are $c_1, \ldots, c_n$. Then M is invertible by the linear independence of its columns. To establish (6.2.2), we show that

$$B_\varepsilon(c_{\text{mean}}) \subset C \quad \text{where } \varepsilon = \frac{1}{100n\|M^{-1}\|}. \tag{6.2.3}$$

for $\varepsilon > 0$ to be chosen. Fix $x \in B_\varepsilon(C_{\text{mean}})$, and note that

$$|M^{-1}(x - c_{\text{mean}})| \leq \|M^{-1}\||x - c_{\text{mean}}| \leq \|M^{-1}\|\varepsilon < \frac{1}{100n}. \tag{6.2.4}$$

Clearly,

$$M^{-1}c_{\text{mean}} = \left(\frac{1}{2n}, \frac{1}{2n}, \ldots, \frac{1}{2n}\right).$$

Hence, by (6.2.4), we have

$$M^{-1}x = (\theta_1, \ldots, \theta_n)$$

where

$$0 < \theta_i < \frac{1}{n}. \tag{6.2.5}$$

Hence, we can write

$$x = (1 - \theta_1 - \cdots - \theta_n)0 + \theta_1 c_1 + \cdots + \theta_n c_n.$$

We deduce from the above (keeping in mind (6.2.5)) that $x \in C$, which implies that (6.2.3), as desired. This concludes the proof. □

6.3. **Convex functions.** The standard second order characterization, that is, that $D^2 f$ is non-negative definite, for convexity is too restrictive, as we see in the plots below. Let us go over two simpler conditions.

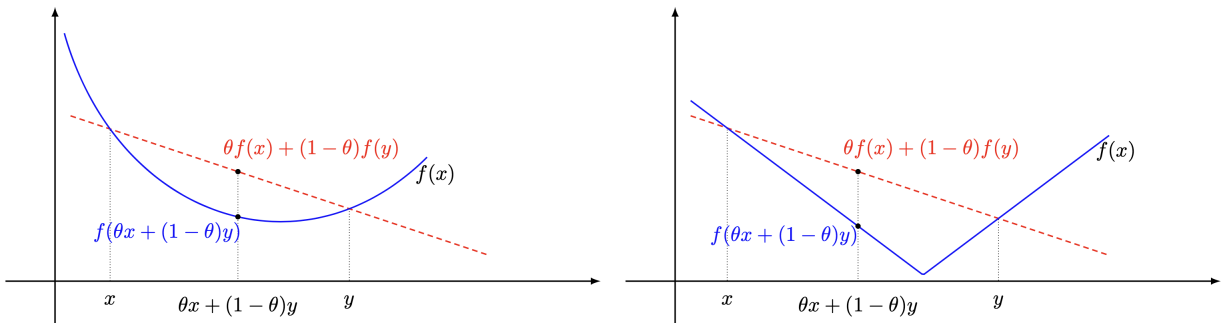
6.3.1. **Zeroth order condition.** Recall that a convex function $f : V \to [-\infty, \infty]$ satisfies

$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta)f(y) \qquad \text{for all } \theta \in (0,1), x \neq y \in \mathbb{R}.$$

We showed this in one dimension on a homework. On the other hand, this is a perfectly good definition of a function being convex, so let us use it as one! It essentially captures the fact that the set of points above the graph of f is a convex set.

We say that f is strictly convex if the above is true with "$<$" in place of "$\leq$."

Let us illustrate this definition in some figures:



An advantage to this definition is that it does not require that f is smooth. Indeed, in the second figure above, we see that f need not be everywhere differentiable to satisfy the definition of convexity.

If the domain of f is $K \subset V$ then we additionally require that K be convex. On the other hand, we say that f is **concave** if $-f$ is convex.

We allow f to take infinite values under the convention that $\theta \cdot (\pm\infty) = \pm\infty$ if $\theta > 0$, $\theta \cdot (\pm\infty) = \mp\infty$ if $\theta < 0$, and $0 \cdot (\pm\infty) = 0$. We have to avoid situations with expressions like $\infty - \infty$. We also consider inequalities of the form $\infty \leq \infty$ to be valid; for example, this allows $\infty \leq (1/2)\infty$, although this is not normally valid.

Example 6.3.1. *(i) Fix a set $C \subset \mathbb{R}^n$. Define the function*

$$I_C(x) = \begin{cases} 0 & \text{if } x \in C, \\ \infty & \text{if } x \notin C. \end{cases} \tag{6.3.1}$$

Exercise 6.3.1. *Show that I_C is convex if and only if C is convex.*

(ii) The function

$$f(x) = \begin{cases} \frac{1}{1-x^2} & \text{if } x \in (-1,1), \\ \infty & \text{if } |x| \geq 1 \end{cases}$$

is convex.

Exercise 6.3.2. *Show this!*

(iii) In general, if f_C is a convex function on a set C, then

$$f(x) = \begin{cases} f_C(x) & \text{if } x \in C, \\ \infty & \text{if } x \notin C. \end{cases}$$

is convex.

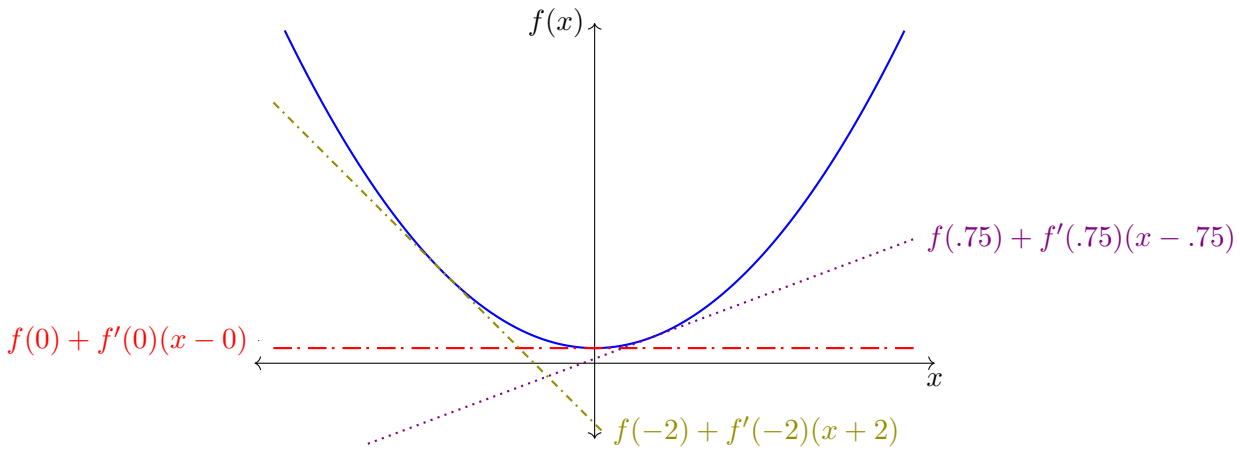
Exercise 6.3.3. *Show this!*

6.3.2. **First order condition.** Another way to characterize a convex function f is that the tangent plane to the graph of f at any point x_0 sits below the graph of f. Mathematically, we write this as: fix $f : K \to \mathbb{R}$, with K open[49] and convex, then f is convex if, for every $x_0 \in K$,

$$f(y) \geq f(x_0) + \nabla f(x_0) \cdot (y - x_0) \qquad \text{for all } y \in K. \tag{6.3.2}$$

Let us illustrate this in a figure, where the dotted, dashed, and dot dashed lines are various tangent lines.

[49] We require open so that we can easily make sense of derivatives.



Let us show that this characterization of f jibes with the zeroth order definition of convex.

Proposition 6.3.2. *Suppose that $f \in C^1$. Then f is a convex function if and only if* (6.3.2) *holds for all $x_0, y \in K$.*

Proof. We first show the forwards direction. Suppose that f is convex. We use a standard trick: let $F : [0, 1] \to \mathbb{R}$ be defined by

$$F(t) = f(x + t(y - x)). \tag{6.3.3}$$

Notice that

$$F'(0) = \nabla f(x) \cdot (y - x).$$

Using the convexity of f, we have, for all $t \in (0, 1)$,

$$\frac{F(t) - F(0)}{t} = \frac{f((1-t)x + ty) - f(x)}{t} \leq \frac{(1-t)f(x) + tf(y) - f(x)}{t} = f(y) - f(x).$$

Taking the limit $t \to 0$ and using (6.3.3), we find

$$\nabla f(x) \cdot (y - x) = F'(0) \leq f(y) - f(x),$$

which completes the proof of the claim.

Now we show the backwards direction. Suppose that f satisfies (6.3.2). Then applying (6.3.2) twice, we find

$$\begin{aligned} f(x) &\geq f(\theta x + (1-\theta)y) + \nabla f(\theta x + (1-\theta)y) \cdot (\theta x + (1-\theta)y - x) \\ &= f(\theta x + (1-\theta)y) + \nabla f(\theta x + (1-\theta)y) \cdot (-(1-\theta)x + (1-\theta)y) \end{aligned}$$

and

$$\begin{aligned} f(y) &\geq f(\theta x + (1-\theta)y) + \nabla f(\theta x + (1-\theta)y) \cdot (\theta x + (1-\theta)y - y) \\ &= f(\theta x + (1-\theta)y) + \nabla f(\theta x + (1-\theta)y) \cdot (\theta x - \theta y). \end{aligned}$$

Multiplying the first inequality by θ and the second inequality by $1 - \theta$ and then adding them, we find

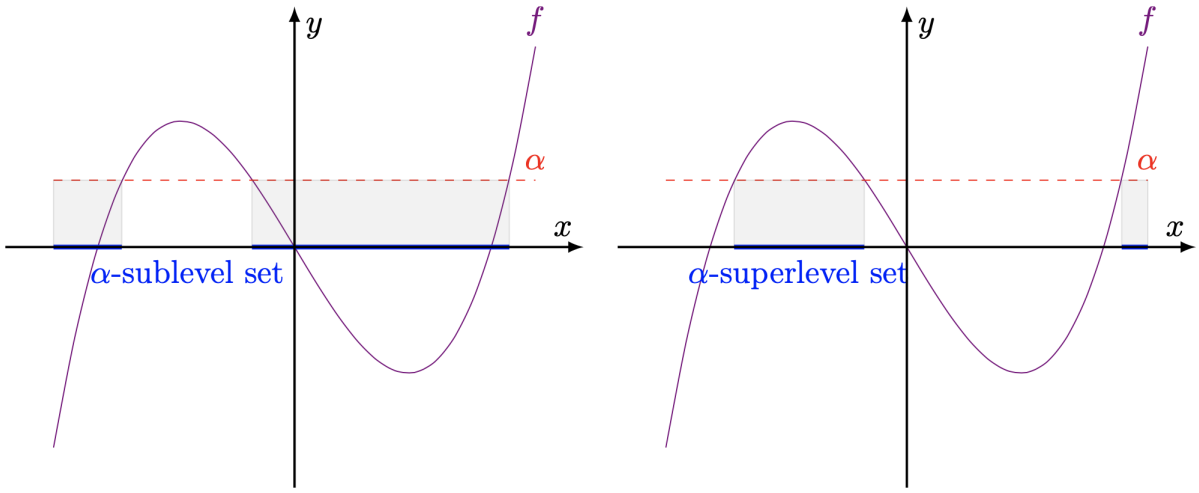
$$\theta f(x) + (1-\theta)f(y) \geq f(\theta x + (1-\theta)y).$$

This concludes the proof. □

6.4. **Epigraphs, hypographs sublevel sets, and superlevel sets.** Recall that the graph of a function $f : \mathbb{R}^n \to \mathbb{R}$ is the set $\{(x, f(x)) : x \in \mathbb{R}^n\}$.

Definition 6.4.1. *Fix $f : \mathbb{R}^n \to \mathbb{R}$ and $\alpha \in \mathbb{R}$.*

- *The α-sublevel set is $\{x \in \mathbb{R}^n : f(x) \leq \alpha\}$.*
- *The α-superlevel set is $\{x \in \mathbb{R}^n : f(x) \geq \alpha\}$.*



Proposition 6.4.2. *If a function $f : \mathbb{R}^n \to \mathbb{R}$ is convex, then its α-sublevel set is convex for every α.*

Proof. This is nearly a tautology when using the zeroth order definition of convexity.

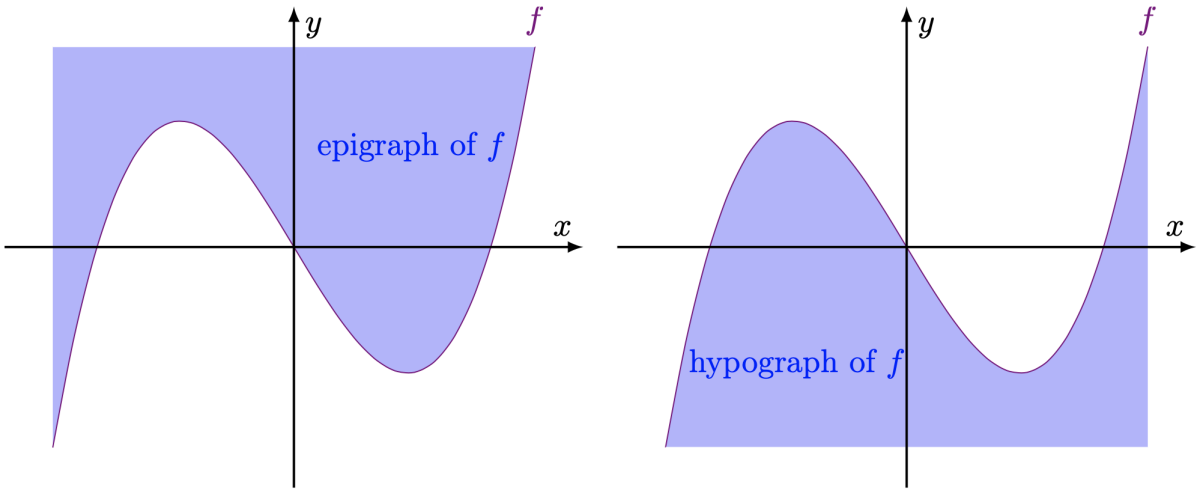
Exercise 6.4.1. *Prove this!*

□

Exercise 6.4.2. *Find an example of f that is* not *convex but whose α-sublevel sets are all convex.*

Definition 6.4.3. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a function.*

- *The* epigraph *of f is the set* $\operatorname{epi} f = \{(x, t) : t \geq f(x)\}$.
- *The* hypograph *of f is the set* $\operatorname{hypo} f = \{(x, t) : t \leq f(x)\}$.



Proposition 6.4.4. *A function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if and only if its epigraph set is convex.*

Proof. This is nearly a tautology when using the zeroth order definition of convexity.

Exercise 6.4.3. *Prove this!*

□

Notice that Proposition 6.4.2 and Proposition 6.4.4 give a nice connection between sets and functions. This way of looking at things lends itself to easy proofs of useful theorems. We cover one (very important consequence!) here:

Proposition 6.4.5. *Suppose that $\mathcal{I}$ is an indexing set and that, for every $\iota \in \mathcal{I}$, the function f_ι is convex. Then the function defined by*

$$f(x) = \sup_{\iota \in \mathcal{I}} f_\iota(x)$$

is convex as well.

Before proving this, let us point out that this is a possibly terrible function! In fact, there is not reason to believe that f has any form of continuity at all!

Proof. We claim that the epigraph of f is the intersection of the epigraphs of the f_ι's:

$$\{(x,t) : t \geq f(x)\} = \bigcap_{\iota \in \mathcal{I}} \{(x,t) : t \geq f_\iota(x)\}. \tag{6.4.1}$$

Note that, were this true, we would be finished because the lefthand side is the intersection of convex sets and is, thus, convex.

Let us now establish (6.4.1). □

6.4.1. **A useful characterization of convex functions.**

Proposition 6.4.6. *Let $f : \mathbb{R}^n \to \mathbb{R}$. The function f is convex if and only if there is an indexing set $\mathcal{I}$ such that, for every x,*

$$f(x) = \sup_{\iota \in \mathcal{I}} (m_\iota \cdot x + b_\iota).$$

Note: $m_\iota \in \mathbb{R}^n$ and $b_\iota \in \mathbb{R}$ for each $\iota \in \mathcal{I}$.

Proof. For simplicity, we prove it only in the case where f does not take infinite values; however, an analogue holds in the case where f takes infinite values.

$\implies$ This direction is straightforward. For each x, let λ_x, b_x be such that $\{(y, m_x \cdot y + b_x) : y \in \mathbb{R}^n\}$ is a supporting hyperplane[50] of the epigraph of f. It follows that, for all y,

$$f(y) \geq m_x \cdot y + b_x.$$

We are then finished after choosing the indexing set $\mathcal{I} = \mathbb{R}^n$. Indeed,

$$f(x) = \lambda_x \cdot x + b_x \quad \text{and} \quad f(x) \geq \lambda_y \cdot x + b_y \ \text{ for all } y.$$

Hence,

$$f(x) = \sup_{y \in \mathcal{I}} (m_y \cdot x + b_y).$$

$\impliedby$ We showed above that the supremum of convex functions remains convex (Proposition 6.4.5). Linear functions are convex. The result follows.

□

Exercise 6.4.4. *Using that, for any symmetric $n \times n$ matrix M, the principal eigenvalue[51] satisfies*

$$\lambda_1(M) = \sup_{x \in \mathbb{R}^n, \|x\|_2 = 1} x \cdot Mx,$$

show that λ_1 is a convex function of M.

Exercise 6.4.5. *Show that $\lambda_2(M)$ is not necessarily convex, and that $\lambda_n(M)$ is concave.*

6.4.2. **Subgradients.** Given a function $f : V \to \mathbb{R}$, we may define a weak notion of gradient even in the cases where f is not differentiable.

Definition 6.4.7. *Fix $x_0 \in V$. We say that $\lambda \in V^*$ is a* subgradient *of f at x_0 if, for all $x \in V$,*

$$\lambda(x - x_0) + f(x_0) \leq f(x). \tag{6.4.2}$$

We denote by $\partial f(x_0)$ the set of all subgradients. This is called the subdifferential.

Proposition 6.4.8. *Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is convex.*

(i) If f is differentiable at x_0 then $\partial f(x_0) = \{\nabla f(x_0)\}$.

[50] Here we are using that f is not infinite. Otherwise, it could be that the hyperplane is vertical and, thus, not of this form.

[51] i.e., the largest eigenvalue

(ii) If $\partial f(x_0)$ has at least two points in it, then it is infinite.

Proof. (i) First we observe that $\nabla f(x_0) \in \partial f(x_0)$ by (6.3.2), the first order condition for convexity. Next, we show that if $a \in \partial f(x_0)$, then

$$a = \nabla f(x_0).$$

We argue by contradiction assuming that $a \neq \nabla f(x_0)$. By definition[52], we know that there is a function $g : \mathbb{R} \to \mathbb{R}_+$ such that $g(x) \to 0$ as $x \to 0$ and

$$f(y) \leq f(x_0) + \nabla f(x_0) \cdot (y - x_0) + g(|x_0 - y|)|x_0 - y|.$$

Then, since $a \in \partial f(x_0)$, we have

$$a \cdot (y - x_0) + f(x_0) \leq f(y) \leq f(x_0) + \nabla f(x_0) \cdot (y - x_0) + g(|x_0 - y|)|x_0 - y|.$$

Rearranging this, we have

$$0 \leq \frac{(\nabla f(x_0) - a) \cdot (y - x_0)}{|y - x_0|} + g(|y - x_0|).$$

Crucially, we may take any choice of y here. For any $\varepsilon > 0$, let us choose $y_\varepsilon = -\varepsilon(\nabla f(x_0) - a)$, and the above becomes

$$0 \leq -|\nabla f(x_0) - a|^2 + g(\varepsilon|\nabla f(x_0) - a|).$$

Taking $\varepsilon \searrow 0$, we find $0 \leq -|\nabla f(x_0) - a|^2$, which is a contradiction. The proof is finished.

(ii) This follows from the homework exercise showing that $\partial f(x_0)$ is a convex set. Hence, if $a_0, a_1 \in \partial f(x_0)$, so is $\theta a_0 + (1 - \theta)a_1$ for every $\theta \in [0, 1]$. This completes the proof.

□

Note that Proposition 6.4.6 shows that, if f is convex, the subdifferential is nonempty.

Example 6.4.9. *Let* $\operatorname{abs} : \mathbb{R}^n \to \mathbb{R}$ *be defined by* $\operatorname{abs}(x) = |x|$. *We claim that, if* $x_0 \neq 0$,

$$\partial \operatorname{abs}(x_0) = \{x_0\}$$

and

$$\partial \operatorname{abs}(0) = \{y \in \mathbb{R}^n : |y| \leq 1\}.$$

We note that the claim about $x_0 \neq 0$ *follows from the fact that* abs *is smooth away from* 0, $\nabla \operatorname{abs}(x_0) = x_0/|x_0|$, *and,* $D^2 \operatorname{abs}$ *is nonnegative definite*[53]. *Indeed, by Taylor's approximation, there is* $\theta \in [0, 1]$ *such that*

$$\begin{aligned} \operatorname{abs}(x) &= \operatorname{abs}(x_0) + D \operatorname{abs}(x_0) \cdot (x - x_0) + \frac{1}{2}(x - x_0) \cdot D^2 \operatorname{abs}(\theta x + (1 - \theta)x_0))(x - x_0) \\ &\geq \operatorname{abs}(x_0) + D \operatorname{abs}(x_0) \cdot (x - x_0). \end{aligned}$$

[52] Recall that $\nabla f(x_0)$ is defined as a vector such that $\lim_{y \to x_0} \frac{|f(y) - f(x_0) - \nabla f(x_0) \cdot (y - x_0)|}{|x_0 - y|} = 0$.

[53] Recall that a symmetric matrix M is nonnegative definite if $x \cdot Mx \geq 0$ for all x. Equivalently, this means that M has all nonnegative eigenvalues.

Hence, (6.4.2) *is satisfied with the choice* $\lambda = D\operatorname{abs}(x_0)\cdot$.

We now consider the case $x_0 = 0$. *First we check that anyy* $\in \mathbb{R}^n$ *is an element of the subdifferential, we see that*

$$\operatorname{abs}(0) + y \cdot (x - 0) = y \cdot x \leq |y|\,|x| \leq \operatorname{abs}(x).$$

Hence, (6.4.2) *holds for* $\lambda = y\cdot$. *Next, we finish by checking that any* $y \in \mathbb{R}^n$ *with* $|y| > 1$ *is not in the subdifferential. Indeed,*

$$\operatorname{abs}(0) + y \cdot (y - 0) = y \cdot y = |y|^2 > |y| = \operatorname{abs}(y).$$

Hence, (6.4.2) *does not hold with* $x = y$, $x_0 = 0$, *and* $\lambda = y\cdot$.

Exercise 6.4.6. *The subdifferential of a convex function is a closed, convex set.*

6.4.3. **Operations preserving convexity.**

Exercise 6.4.7. *Prove the following:*

- **Suprema/Max:** *We have already showed that* $f(x) = \sup_{\iota\in\mathcal{I}} f_\iota(x)$ *is convex if* f_ι *is convex for every* $\iota \in \mathcal{I}$.
- **Convex combinations:** *If* $f_1, \ldots, f_n$ *are convex,* $\theta_1, \ldots, \theta_n \geq 0$, *and* $\theta_1 + \cdots + \theta_n = 1$, *then*

  $$f(x) = \theta_1 f_1(x) + \cdots + \theta_n f_n(x)$$

  is also convex. Why? One can check that $\operatorname{epi} f = \theta_1 \operatorname{epi} f_1 + \cdots + \theta_n \operatorname{epi} f_n$. *The convex combination of combination of convex sets is comvex.*
- **Composition of affine and convex functions:** *Suppose that* $g : W \to V$ *is affine and* $f : V \to \mathbb{R}$ *is convex. Then* $f \circ g : W \to \mathbb{R}$ *is convex. Why? The linearity of* g *preserves sums.*
- **Restriction to lower dimensional convex sets:** *Let* $f : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}$ *be a convex function and* $C \subset \mathbb{R}^m$ *a convex set. Then*

  $$g(x) = \inf_{y\in C} f(y, x)$$

  is a convex set.

Let us prove the last point:

Proof. Fix any $x_1, x_2 \in \mathbb{R}^n$ and $\theta \in [0, 1]$. Fix $\varepsilon > 0$ and find $y_1, y_2 \in C$ such that, for each $i = 1, 2$,

$$f(y_i, x_i) \leq g(x_i) + \frac{\varepsilon}{100}. \tag{6.4.3}$$

Then

$$\begin{aligned} g(\theta x_1 + (1-\theta)x_2) &\leq f(\theta y_1 + (1-\theta)y_2, \theta x_1 + (1-\theta)x_2) = f(\theta(y_1, x_1) + (1-\theta)(y_2, x_2)) \\ &\leq \theta f(y_1, x_1) + (1-\theta) f(y_2, x_2) \leq \theta\left(g(x_1) + \frac{\varepsilon}{100}\right) + (1-\theta)\left(g(x_2) + \frac{\varepsilon}{100}\right) \\ &\leq \theta g(x_1) + (1-\theta) g(x_2) + \varepsilon. \end{aligned}$$

In the first step, we used that g is an infimum over all choices of y (so we may choose any convenient y), while in the following steps we used (6.4.3) and the definition of convexity. Since this holds for every ε, it follows that

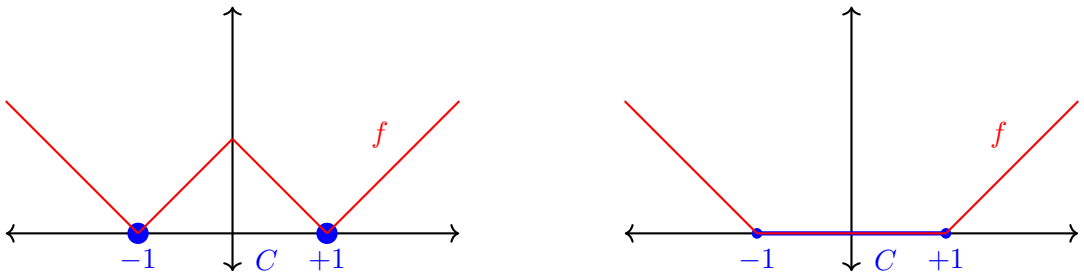
$$g(\theta x_1 + (1-\theta)x_2) \leq \theta g(x_1) + (1-\theta)x_2).$$

□

Example 6.4.10. *If $C \subset \mathbb{R}^n$ and*

$$d_C(x) = \inf_{y \in C} \|x - y\|,$$

then d_C is convex if and only if C is convex. We show two simple examples of this below: the one on the left for the set $C = \{-1, 1\}$ (not convex) and the one on the right for the set $C = [-1, 1]$ (convex).



6.5. **Conjugate functions and duality.**

6.6. **Fenchel conjugate.** Many of the ideas above center on the idea of duality (e.g., Proposition 1.4.7). We can use this idea of optimizing over all 'test' points to define a new functions: given $F : \mathbb{R}^n \to \mathbb{R}$, let

$$F^*(x) = \sup_{y \in \mathbb{R}^n} \left(x \cdot y - F(y) \right). \tag{6.6.1}$$

This is called the **Fenchel conjugate** (or it can be called the convex conjugate or the **Legendre transformation**, although the latter is usually only used when F is convex). The output may be infinite for certain values of x, depending on the shape of F.

Notice that (6.6.1) immediately yields an inequality, the **Fenchel inequality**:

$$x \cdot y \leq F(y) + F^*(x) \quad \text{for all } x, y \in \mathbb{R}^n. \tag{6.6.2}$$

We see below that Young's inequality (Lemma 1.4.5) is a particular case of the Fenchel inequality.

Another consequence of (6.6.2) is a nice geometric interpretation of F^*: we have

$$x \cdot y - F^*(x) \leq F(y) \quad \text{for all } y \in \mathbb{R}$$

and, due to (6.6.1), for any $\varepsilon > 0$, there is y_ε such that

$$x \cdot y_\varepsilon - F^*(x) + \varepsilon > F(y_\varepsilon).$$

In other words, $x \cdot y - F^*(x)$ is the largest plane at angle x that sits below the graph of F. In one dimension, this is perhaps easier to envision: $xy - F^*(x)$ is the largest line with slope x that sits below the graph of F. To really drive this home, let us rephrase it in the following way: start the graph of xy with intercept $-\infty$ and then "slide" it up until it "touches" the graph of F. The intercept of this is $F^*(x)$. Here are some pictures showing this: Let us rephrase this one last time. $F^*(x_0)$ is the smallest vertical shift of a plane defined by x_0 so that the resulting plane (the graph of $x_0 \cdot y - F^*(y)$) sits underneath the graph of F.

Another thing to note is that, when the supremum is a maximum, it is attained at some x^* satisfying

$$0 = \nabla_y \left(x \cdot y - F(y) \right) \Big|_{y = x^*} = x - \nabla F(x^*) \quad \text{–or–} \quad x = \nabla F(x^*).$$

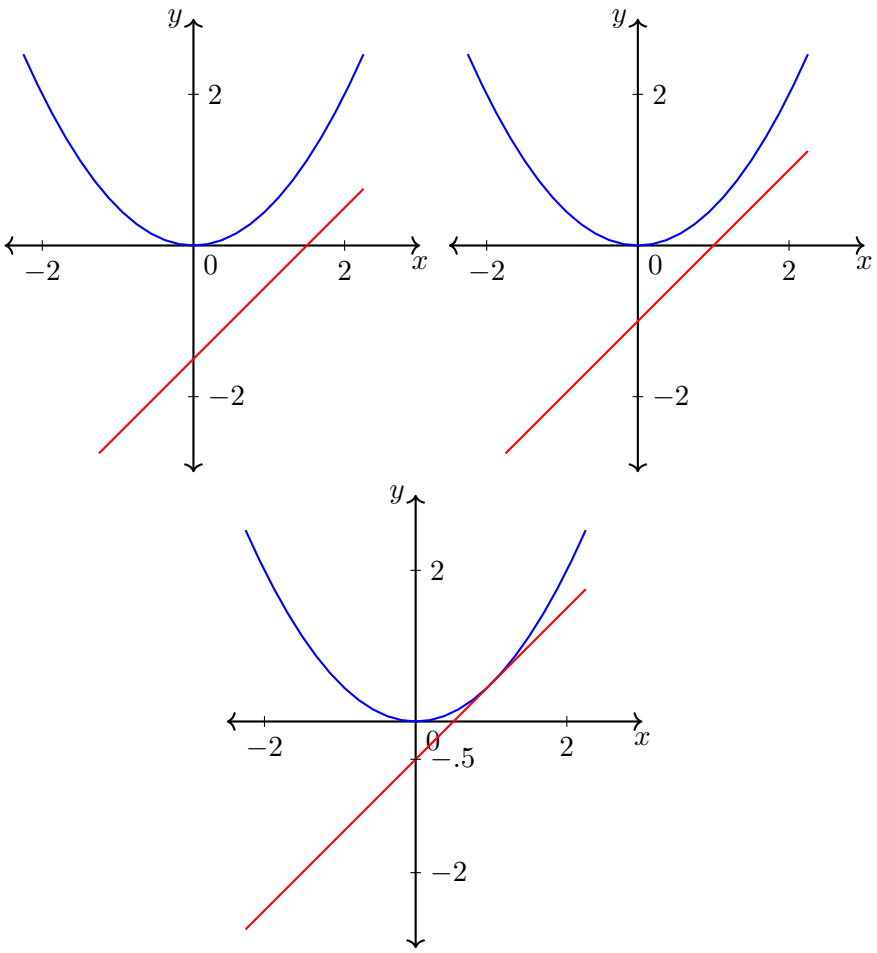


Figure 8: Here we show the geometric meaning of $F^*(1)$ when $F(x) = \frac{1}{2}x^2$. In each frame, we slide a line with slope 1 up until it "touches" the graph of F in the last frame. Since this "touching" line's y-intercept is $-.5$, then $F^*(1) = .5$.

Example 6.6.1. *Let $F : \mathbb{R} \to \mathbb{R}$ and $F(x) = ax$ for some $a > 0$. Then*

$$F^*(x) = \sup_y \left(xy - ay^2\right).$$

Since

$$\lim_{y \to \pm\infty} \left(xy - ay^2\right) = \infty$$

the supremum is actually a maximum. It must, thus, be attained at a critical point:

$$0 = \left(xy - ay^2\right)' = x - 2ay \implies y = \frac{x}{2a}.$$

Thus,

$$F^*(x) = x\frac{x}{2a} - a\left(\frac{x}{2a}\right)^2 = \frac{x^2}{2a} - \frac{x^2}{4a} = \frac{x^2}{4a}.$$

Exercise 6.6.1. *Let p, q be conjugate exponents with $p, q \in (1, \infty)$. Let $F(x) = |x|^p/p$. Show that $F^*(x) = |x|^q/q$. Use this to deduce Young's inequality (Lemma 1.4.5) via Fenchel's inequality.*

Note: the above can also be extended to infinite dimensional spaces using bounded linear functionals in place of $x \cdot y$. *Note that f^* is the supremum of convex (affine) functions so it is itself convex by Proposition 6.4.6.*

When f is continuously differentiable, grows fast enough[54] (as $|x| \to \infty$), and is strictly convex, there is a unique solution $y_x \in \mathbb{R}^n$ to

$$x = \nabla f(y_x)$$

and then

$$f^*(x) = x \cdot y_x - f(y_x).$$

The procedure of finding this y_x and then using it to compute f^* is sometimes called the *Legendre transform* of f[55].

What does one do in the case where f is not be smooth? If f is convex we can replace (6.3.2) with the subdifferential: take any $y_x \in \mathbb{R}^n$ such that

$$x \in \partial f(y_x). \tag{6.6.3}$$

Are we sure this works? What if there are many different elements in the subdifferential? Suppose we have two elements y_1 and y_2 satisfying (6.6.3). Can we compare the two potential values of $f^*(x)$:

$$x \cdot y_1 - f(y_1) \quad \text{and} \quad x \cdot y_2 - f(y_2)?$$

Using the definition of subdifferential (6.4.7), we find

$$x \cdot (y_2 - y_1) + f(y_1) \leq f(y_2) \quad \text{and} \quad x \cdot (y_1 - y_2) + f(y_2) \leq f(y_1).$$

Rearranging these, we find

$$x \cdot y_2 - f(y_2) \leq x \cdot y_1 - f(y_1) \leq x \cdot y_2 - f(y_2).$$

In other words,

$$x \cdot y_2 - f(y_2) = x \cdot y_1 - f(y_1),$$

so the choice does not matter!

Exercise 6.6.2. *Work this out in the case of* $\text{abs} : \mathbb{R}^n \to \mathbb{R}$.

Warning: there may not be a subgradient satisfying (6.6.3). In this case, one has to approach the problem in an *ad-hoc* manner, but usually it is not too difficult to compute in these cases.

Exercise 6.6.3. *(i) Let $C \subset \mathbb{R}^n$ be a convex set. Define I_C as in* (6.3.1)*. Show that I_C^* is the support function*[56] *of C. Is there a subgradient for every $x \in \mathbb{R}^n$?*

(ii) Suppose that f is given by the sum of independent functions:

$$f(\bar{x}) = f(x_1, \ldots, x_n) = f_1(x_1) + \cdots + f_n(x_n).$$

Then $f^* = f_1^* + \cdots + f_n^*$.

[54] That is, that $\liminf_{|x| \to \infty} \frac{f(x)}{|x|} = \infty$.

[55] Actually, in my experience, people often refer to the Fenchel conjugate as the Legendre transform, but this is probably field specific.

[56] see HW1, Spring 2023

(iii) Take the soft-max function

$$f(\bar{x}) = \log\left(e^{x_1} + \cdots + e^{x_n}\right).$$

Let $p \in \Delta_n$, which is the probability[57] *simplex*

$$\Delta_n = \{\bar{p} \in [0,1]^n : p_1 + \cdots + p_n = 1\}.$$

Show that its conjugate is the negative entropy function

$$f^*(\bar{p}) = p_1 \log p_1 + \cdots + p_n \log p_n,$$

where we use the convention that $0 \log 0 = 0$. What happens if $\bar{p} \notin \Delta_n$?

Proposition 6.6.2. *Let $f : \Omega \to \mathbb{R}$ and $x \in \Omega$. Then:*

(i) $f^{**}(x) \leq f(x)$;

(ii) $f^{**}(x) = f(x)$ if f is convex.

Proof.

Exercise 6.6.4. *Prove (i).*

If f is convex, take a supporting hyperplane of epi f; that is, find $w \in \mathbb{R}^n$ such that

$$w \cdot (y - x) + f(x) \leq f(y) \qquad \text{for all } y \in \mathbb{R}^n.$$

Rearranging this, we see that

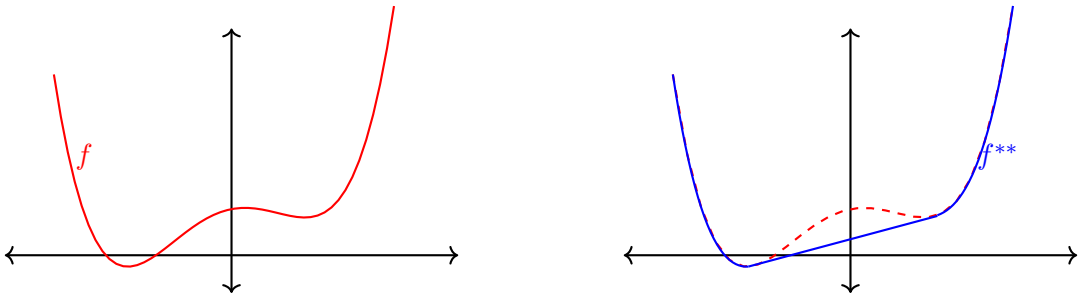
$$f^*(w) = \sup_{y \in \mathbb{R}^n} (w \cdot y - f(y)) = w \cdot x - f(x).$$

Hence,

$$f^{**}(x) = \sup_{u \in \mathbb{R}^n} (x \cdot u - f^*(u)) \geq (x \cdot w - f^*(w)) = f(x).$$

Since $f^{**}(x) \leq f(x)$ by part (i), it follows that $f^{**}(x) = f(x)$. □

Note that, in general, f^{**} will be the "largest" convex function smaller than f. This is illustrated in the images below.



[57] So named because p_i can be the probability of i different events, one of which has to happen (e.g., p_1 = the probability of heads on a coin flip and p_2 = the probability of tails on that coin flip.

6.7. **Jensen's inequality.** A fundamental inequality that is especially useful in probability is Jensen's inequality. We begin with the discrete version of the inequality

Proposition 6.7.1. *Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is a convex function. Suppose that $\theta_1, \ldots, \theta_k \in [0,1]$ with $\theta_1 + \cdots + \theta_k = 1$. Then*

$$f\Big(\sum_{i=1}^{k} \theta_i x_i\Big) \leq \sum_{i=1}^{k} \theta_i f(x_i). \tag{6.7.1}$$

Before proving this, let us make a quick note above how this relates to probability. Let us say that a random event with k different possible outcomes $x_1, \ldots, x_n$ occurs and each has probability $\theta_1, \ldots, \theta_n$. Suppose that f is some "payout" function. Then

$$\sum_{i=1}^{k} \theta_i x_i$$

is the mean outcome. Thus, the left hand side of (6.7.1) is the payout function evaluated at the mean outcome. On the other hand, the right hand side of (6.7.1) is the expected value of the payout. In other words, the payout of the mean value is less than the expected payout.

Proof. If $k = 1$, then $\theta_1 = 1$ and the result is obvious. If $k = 2$, the inequality (6.7.1) is exactly the one appearing in the definition of convexity and are finished. Let us now prove it for $k > 2$.

By definition, $(x_i, f(x_i)) \in \operatorname{epi} f$ for each i. Since f is convex, so is $\operatorname{epi} f$. It follows that

$$\left(\sum_{i=1}^{k} \theta_i x_i, \sum_{i=1}^{k} \theta_i f(x_i)\right) = \sum_{i=1}^{k} \theta_i (x_i, f(x_i)) \in \operatorname{epi} f.$$

This is equivalent to (6.7.1), which finishes the proof. □

The continuous version of the inequality is the one perhaps used more often in practice.

Proposition 6.7.2. *Let $f : K \to \mathbb{R}$ be a convex function on a convex set K, and let $g : K \to \mathbb{R}_+$ be a Riemann integrable function such that*

$$\int_K g(x)dx = 1. \tag{6.7.2}$$

Then

$$f\Big(\int_K x g(x)dx\Big) \leq \int_K f(x)g(x)dx.$$

Two comments are in order before we begin our proof. First, $g(x)$ here plays the role of θ_i in Proposition 6.7.1. Second, we can integrate against arbitrary "measures" instead of just measures of the form $g(x)dx$. This will make sense later in the notes, but its importance is related to integrating against an arbitrary probability measure dP. In this case, we obtain the following version of Jensen's inequality:

$$f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)]$$

when $\mathbb{E}$ is an expectation, X is a random variable (e.g., the value of a roll of dice), and f is a convex function. In this sense, we find

$$\mathbb{E}[X]^2 \leq \mathbb{E}[X^2],$$

by using $f(x) = x^2$, which yields

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

Proof. For notational simplicity, let

$$\bar{x} = \int_K x g(x) dx. \tag{6.7.3}$$

Exercise 6.7.1. *Show that* $\bar{x} \in K$.

Then, by convexity of f, there is a subgradient $a \in \partial f(\bar{x})$. Hence, we have, for every $x \in K$,

$$a \cdot (x - \bar{x}) + f(\bar{x}) \leq f(x). \tag{6.7.4}$$

Notice that

$$\int_K a \cdot (x - \bar{x}) g(x) dx = a \cdot \Big(\int_K x g(x) dx - \int_K \bar{x} g(x) dx \Big) = a \cdot \Big(\bar{x} - \bar{x} \Big) = 0.$$

Above, we used (6.7.3) and that

$$\int_K \bar{x} g(x) dx = \bar{x}$$

due to (6.7.2). Similarly,

$$\int_K f(\bar{x}) g(x) dx = f(\bar{x})$$

Hence, integrating (6.7.4) with respect to $g(x)dx$, we find

$$f(\bar{x}) = \int_K \left(a \cdot (\bar{x} - \bar{x}) + f(\bar{x}) \right) g(x) dx \leq \int_K f(x) g(x) dx,$$

which is exactly the desired conclusion. □

One application of this is the Kullback-Liebler Divergence

$$D_{\mathrm{KL}}(p \| q) := - \int \log \left(\frac{q}{p} \right) p \, dx,$$

that is a kind of relative entropy that describes the distance between two probability measures[58]. Since $-\log$ is a convex function, we have

$$D_{\mathrm{KL}}(p \| q) \geq - \log \left(\int \left(\frac{q}{p} \right) p \, dx \right) = - \log \left(\int q(x) dx \right) = - \log(1) = 0.$$

Moreover, it is easy to check that log is strictly convex and this, along with the proof of Proposition 6.7.2 indicates that

$$D_{\mathrm{KL}}(p \| q) = 0 \quad \Longleftrightarrow \quad p = q.$$

Hence, D_{KL} is a reasonable measure of the alikeness of two measures.

[58] That is, $p, q \geq 0$ and $\int p(x) dx = \int q(x) dx = 1$.

7. Optimization

7.1. **Linear Programming.** A *linear program* is a convex optimization problem in which all functions are affine:

maximize	$c \cdot x + d$
subject to	$Mx \preceq t$
	$Ax = b.$

where we say that

$$v \preceq w \quad \text{if } v_i \leq w_i \text{ for all } i.$$

Here $x, c \in \mathbb{R}^n$, $t \in \mathbb{R}$, $M \in \mathbb{R}^{m \times n}$, $t \in \mathbb{R}^m$, $A \in \mathbb{R}^{k \times n}$, and $b \in \mathbb{R}^k$.

Clearly the choice of d does not affect the identification of the minimizing x so we may omit it if our goal is only to find the location of the maximum. We opt to do this in these notes. We call those x satisfying $Mx \preceq t$ and $Ax = b$ the *feasible set*. We note that the feasible set is a polyhedron.

A linear program is *bounded* if the objective function $c \cdot x + d$ is bounded on the feasible set. The value of an unbounded linear program is $+\infty$.

We note that there is no algorithm to exactly solve a linear program, although the field is mature. On the other hand, there are a number of good strategies to try and algorithms that give approximate solutions.

7.1.1. **Standard and inequality forms.** We say that a linear program is in *standard form* if it is of the form:

maximize	$c \cdot x$
subject to	$x \succeq 0$
	$Ax = b.$

In other words, the componentwise inequality involves only nonnegativity constraints $x \succeq 0$.

We can convert a linear program

maximize	$c \cdot x$
subject to	$Mx \preceq t$
	$Ax = b$

to standard form in the following way. Let us introduce the "slack variable" $s \succeq 0$ such that $Mx + s = t$. Additionally, let us write

$$x = x^+ - x^-$$

for two new vectors $x^+, x^- \succeq 0$. Then, the above becomes

maximize	$c \cdot x^+ - c \cdot x^-$
subject to	$Mx^+ - Mx^- + s = t$
	$Ax^+ - Ax^- = b$
	$x^+ \succeq 0,\ x^- \succeq 0,\ s \succeq 0.$

Thought of properly, this is a linear program in standard form. Indeed, let

$$\tilde{x} = (x_1^+, \dots, x_n^+, x_1^-, \dots, x_n^-, s_1, \dots, s_n)$$
$$\tilde{c} = (c_1, \dots, c_n, -c_1, \dots, -c_n, 0, \dots, 0), \qquad \tilde{b} = (t_1, \dots, t_m, b_1, \dots, b_k),$$
$$\text{and } \tilde{A} = \begin{bmatrix} M & -M & \mathrm{id} \\ A & -A & 0 \end{bmatrix},$$

then

maximize	$\tilde{c} \cdot \tilde{x}$
subject to	$\tilde{x} \succeq 0$
	$\tilde{A}\tilde{x} = \tilde{b}.$

This last step is really quite awkward, so it is perfectly fine to use the more reasonable form

maximize	$c \cdot x^+ - c \cdot x^-$
subject to	$x^+ \succeq 0,\ x^- \succeq 0,\ s \succeq 0$
	$Mx^+ - Mx^- + s = t$
	$Ax^+ - Ax^- = b.$

On the other hand, it is in *inequality form* if it is of the form:

maximize	$c \cdot x$
subject to	$Mx \preceq t.$

A linear program inequality form can be solved in $O(n^2 m)$ steps to a given accuracy (recall that $M \in \mathbb{R}^{m \times n}$).

7.1.2. **The dual problem: an extended example.** Let us consider the following linear program:

maximize	$2x_1 + 3x_2$
subject to	$4x_1 + 8x_2 \leq 12$
	$2x_1 + x_2 \leq 3$
	$3x_1 + 2x_2 \leq 4$
	$(x_1, x_2) \succeq 0.$

We refer to this as the *primal* problem, for reasons that become clear below.

The first question that one might ask is if the problem is *feasible*; that is, are there any x that satisfy the constraints? Here it is easy to see that $x = (0, 0)$ is in the feasible set.

Next, we check that the maximum is bounded:

$$2x_1 + 3x_2 \leq 4x_1 + 8x_2 \leq 12.$$

A slightly better argument is

$$2x_1 + 3x_2 \leq \frac{1}{2}(4x_1 + 8x_2) \leq 6.$$

Refining this further, we find

$$2x_1 + 3x_2 = \frac{1}{3}(4x_1 + 8x_2 + 2x_1 + x_2) \leq \frac{1}{3}(12 + 3) = 5.$$

What are we doing in each case? What is this strategy? We are choosing $\lambda \succeq 0$ and writing

$$2x_1 + 3x_2 \leq \lambda_1(4x_1 + 8x_2) + \lambda_2(2x_1 + x_2) + \lambda_3(3x_1 + 2x_2) \leq \lambda \cdot [12\ 3\ 4]. \tag{7.1.1}$$

The first inequality only works if

$$\lambda \begin{bmatrix} 4 & 8 \\ 2 & 1 \\ 3 & 2 \end{bmatrix} \succeq [2\ 3]. \tag{7.1.2}$$

To get the "best" inequality in (7.1.1) (that is, the smallest upper bound), we should choose λ that minimizes $\lambda \cdot [12\ 3\ 4]$ subject to the constraint (7.1.2).

This leads to the *dual problem*

minimize	$12\lambda_1 + 3\lambda_2 + 4\lambda_3$
subject to	$4\lambda_1 + 2\lambda_2 + 3\lambda_3 \geq 2$
	$8\lambda_1 + \lambda_2 + 2\lambda_3 \geq 3$
	$\lambda \succeq 0.$

It perhaps "feels" like this should lead us to the solution of the *primal* problem, even if that is not mathematically clear yet. What we do have is *weak duality*: the value of the primal problem is less than or equal to that of the dual problem. Let us call P the value of the primal problem and D the value of the dual problem. The above is summarized as

$$P \leq D. \tag{7.1.3}$$

This leads to the following strategy: if we can find candidate feasible points $x = (x_1, x_2)$ to the primal problem and $\lambda = (\lambda_1, \lambda_2, \lambda_3)$ to the dual problem such that the primal problem at x is larger than the value of the dual problem at λ, then we have solved both the dual and the primal problems!

Here, let us try $x = (1/2, 5/4)$ and $\lambda = (5/16, 0, 1/4)$.

- **Primal problem:** We check that x is feasible:

$$4x_1 + 8x_2 = 2 + 10 \leq 12, \quad 2x_1 + x_2 = 1 + 5/4 \leq 3,$$
$$3x_1 + 2x_2 = 3/2 + 5/2 \leq 4, \qquad x_1, x_2 \geq 0.$$

  We deduce that the value of the primal problem satisfies:

$$P \geq 2x_1 + 3x_2 = 1 + 15/4 = 19/4. \tag{7.1.4}$$

- **Dual problem:** We check that λ is feasible:

$$4\lambda_1 + 2\lambda_2 + 3\lambda_3 = 5/4 + 0 + 3/4 \geq 2, \quad 8\lambda_1 + \lambda_2 + 2\lambda_3 = 5/2 + 0 + 1/2 \geq 3,$$
$$\lambda_1, \lambda_2, \lambda_3 \geq 0.$$

  We deduce that the value of the dual problem satisfies

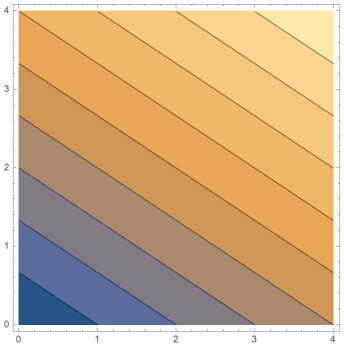
$$D \leq 12\lambda_1 + 3\lambda_2 + 4\lambda_3 = 15/4 + 0 + 1 = 19/4. \tag{7.1.5}$$

Putting together (7.1.3), (7.1.4), and (7.1.5), we find

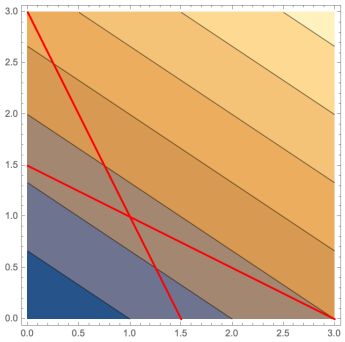
$$19/4 \leq P \leq D \leq 19/4;$$

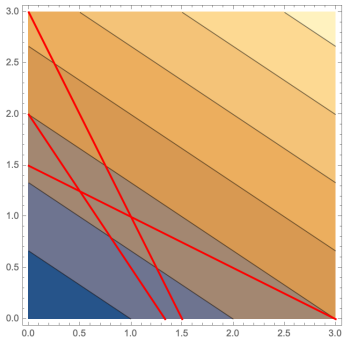
that is, $P = D = 19/4$.

Let us look further into the problem of finding these "good" points. Here is a contour plot of $2x_1 + 3x_2$:



The feasible sets imposed by $2x_1 + x_2 \leq 3$ and $4x_1 + 8x_2 \leq 12$ lead to the figure on the left below, while adding the constraint $3x_1 + 2x_2 \leq 4$ leads to the the right on the right below. The feasible sets are those "below" the red lines.





We immediately notice that the second condition ($2x_1 + x_2 \leq 3$) plays no role. Instead, it is heuristically clear that the "best" point is where the lines $3x_1 + 2x_2 = 4$ and $4x_1 + 8x_2 = 12$ meet. This is exactly how we have found the point $(1/2, 5/4)$.

What about the dual problem? Since the inequality associated to λ_2 is "inactive" (cf. (7.1.1) and recall that $2x_1 + x_2 \leq 3$ plays no role in determining the feasible set in the primal problem), we set $\lambda_2 = 0$. Then we find λ_1 and λ_3 by solving the constraint equations with the choice $\lambda_2 = 0$:

$$4\lambda_1 + 3\lambda_3 = 2 \qquad \text{and} \qquad 8\lambda_1 + 2\lambda_3 = 3.$$

This yields exactly the choice $\lambda = (5/16, 0, 1/4)$ above.

7.1.3. **Finding the dual problem: another example.** Consider the following primal problem:

maximize	$2x_1 + 3x_2 - x_3$
subject to	$2x_1 + x_2 \leq 3$
	$x_1 + x_2 + x_3 \geq 4$
	$x_1 - x_3 = 1$
	$(x_1, x_2) \succeq 0,\ x_3 \in \mathbb{R}.$

Notice that the second inequality is a bit weird: it is a $\geq$ not a $\leq$! So let us be very careful with this. As we did before, we use $\lambda = (\lambda_1, \lambda_2, \lambda_3)$ to find a linear combination that may be useful:

$$\begin{aligned} 2x_1 + 3x_2 - x_3 &\leq \lambda_1(2x_1 + x_2) + \lambda_2(x_1 + x_2 + x_3) + \lambda_3(x_1 - x_3) \\ &\leq 3\lambda_1 + 4\lambda_2 + \lambda_3. \end{aligned} \tag{7.1.6}$$

The first inequality can only hold if

$$\begin{aligned} 2 &\leq 2\lambda_1 + \lambda_2 + \lambda_3 \\ 3 &\leq \lambda_1 + \lambda_2, \end{aligned}$$

because $x_1, x_2 \geq 0$. On the other hand, x_3 is not necessarily positive! So we require $-x_3 = \lambda_2 x_3 - \lambda_3 x_3$; that is,

$$-1 = \lambda_2 - \lambda_3.$$

On the other hand, the second inequality in (7.1.6) holds only if

$$\lambda_1 \geq 0, \qquad \lambda_2 \leq 0, \qquad \text{and} \qquad \lambda_3 \in \mathbb{R}.$$

Why? Because $x_1 + x_2 + x_3 \geq 4$, so $\lambda_2(x_1 + x_2 + x_3) \leq 4\lambda_2$ holds only if $\lambda_2 \leq 0$. Additionally, $x_1 - x_3 = 1$, so $\lambda_3(x_1 - x_3) \leq \lambda_3$ for *any* $\lambda_3 \in \mathbb{R}$. As before, the "best" choice of λ will minimize the right hand side in (7.1.6). We end up with the dual problem

minimize	$3\lambda_1 + 4\lambda_2 + \lambda_3$
subject to	$2\lambda_1 + \lambda_2 + \lambda_3 \geq 2$
	$\lambda_1 + \lambda_2 \geq 3$
	$\lambda_2 - \lambda_3 = -1$
	$\lambda_1 \geq 0, \lambda_2 \leq 0, \lambda_3 \in \mathbb{R}.$

Pulling back for a moment notice that the coefficients in the objective function for the dual problem $(3, 4, 1)$ are precisely the right hand side of the constraints in the primal problem and vice versa. Additionally, the matrix making up the constraints in the primal problem and dual problem are, respectively,

$$\begin{bmatrix} 2 & 1 & 0 \\ 1 & 1 & 1 \\ 1 & 0 & -1 \end{bmatrix} \qquad \text{and} \qquad \begin{bmatrix} 2 & 1 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & -1 \end{bmatrix}$$

Notice that these are the transpose of each other! Additionally, every constraint involving a $\leq$ in the primal problem became a $\geq$ in the dual problem (similarly with $\geq$ and $\leq$). Finally, an equality in the primal problem, lead to any real-value in the dual problem.

7.1.4. **The duality theorem.** In general, we get the following summary for finding the dual problem: Again we stress that the value of the primal problem is no more than the value of the dual problem. When are they equal? We have a handy theorem for this:

Theorem 7.1.1 (Duality Theorem)**.** *Let $c \in \mathbb{R}^n$, $A \in M_{m\times n}$, and $b \in \mathbb{R}^m$. For the linear programs*

$$\textit{maximize } c \cdot x \textit{ subject to } Ax \preceq b \textit{ and } x \succeq 0 \tag{P}$$

and

$$\textit{minimize } b \cdot y \textit{ subject to } A^T y \succeq c \textit{ and } y \succeq 0, \tag{D}$$

there are the following alternatives, exactly one of which occurs:

Dualization Recipe

	Primal linear program	Dual linear program
Variables	$x_1, x_2, \ldots, x_n$	$y_1, y_2, \ldots, y_m$
Matrix	A	A^T
Right-hand side	$\mathbf{b}$	$\mathbf{c}$
Objective function	$\max \mathbf{c}^T\mathbf{x}$	$\min \mathbf{b}^T\mathbf{y}$
Constraints	ith constraint has $\leq$	$y_i \geq 0$
	$\geq$	$y_i \leq 0$
	$=$	$y_i \in \mathbb{R}$
	$x_j \geq 0$	jth constraint has $\geq$
	$x_j \leq 0$	$\leq$
	$x_j \in \mathbb{R}$	$=$

Figure 9: Credit to Matoušek-Gärtner, 2007

(i) Neither (P) nor (D) is feasible;

(ii) (P) is unbounded and (D) is not feasible;

(iii) (P) is not feasible and (D) is unbounded;

(iv) Both (P) and (D) are feasible, have an optimal solution x^* and y^* (respectively), and

$$c \cdot x^* = b \cdot y^*. \tag{7.1.7}$$

We note, critically, that if (iv) occurs, (7.1.10) guarantees that the value of (P) is equal to the value of (D).

The key step in our proof is Farkas's Lemma:

Lemma 7.1.2 (Farkas's Lemma)**.** *Let $A \in M_{m\times n}$ and $b \in \mathbb{R}^m$. Then exactly one of the two alternatives holds:*

(i) $Ax = b$ has a solution $x \in \mathbb{R}^n$ with $x \succeq 0$;

(ii) there is $\lambda \in \mathbb{R}^m$ such that $A^T\lambda \succeq 0$ and $b \cdot \lambda < 0$.

Before proving this lemma, we show it in two examples Let us look at an example before we try to prove Lemma 7.1.2.

Example 7.1.3. *(i) The conditions $x_1 + x_2 = 1$, $x_1 \geq 0$, and $x_2 \geq 0$. This is feasible since $x_1 = x_2 = 1/2$ satisfies these conditions. This corresponds to Lemma 7.1.2.(i) holding with*

$$A = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad \text{and} \quad b = 1. \tag{7.1.8}$$

What does Lemma 7.1.2.(ii) correspond to? We would need $[\lambda, \lambda]^T = A^T\lambda \succeq 0$ and $\lambda = b\lambda < 0$. These are clearly not consistent! Hence, (ii) cannot hold as Farkas's lemma predicts.

(ii) Let us look at a similar problem. Let A be as in (7.1.8) *and $b = -1$. Notice that, for $\lambda = \pi^e$, we get $A^T\lambda \succeq 0$ and $b\lambda = -3 < 0$. Hence, (i) cannot hold. In other words, because we could find* a single solution *to (ii), we conclude that*

$$\text{if } x \succeq 0,\ Ax \neq b,$$

which is a statement for all $x \succeq 0$!

Proof. While at first the reason such an alternative exists is mysterious, it is somewhat "obvious" once the claim is translated into geometric objects. Let $a_1, \dots, a_n \in \mathbb{R}^m$ be the column vectors of A:

$$A = \begin{bmatrix} | & | & & | \\ a_1 & a_2 & \cdots & a_n \\ | & | & & | \end{bmatrix}.$$

Let

$$\begin{aligned} C &= \text{conic}\{a_1, \dots a_n\} \\ &= \{y \in \mathbb{R}^m : y = x_1a_1 + \cdots + x_na_n, \text{ such that } x_i \geq 0 \text{ for all } i = 1, \dots, n\} \\ &= \{Ax : x \succeq 0\}. \end{aligned}$$

From this, we see that the case (i) is exactly when $b \in C$.

Let us then consider the case $b \notin C$. We first give a "fake" proof that is simple and contains the major points. Take H to be a separating hyperplane of b and C. It is "obvious" that, since b is a point and C is a cone, we can take H to go through the origin and not contain b. In this case, let λ be the normla vector defining H; that is, $y \in H$ if and only if $y \cdot \lambda = 0$. By definition of separating hyperplane, $\lambda \cdot a_i \geq 0$ for all i (which implies that $Ay \succeq 0$) and $\lambda \cdot b < 0$ (indeed, $b \cdot \lambda \geq 0$, but $b \notin H$ implies that $b \cdot \lambda > 0$). This completes the proof that we are in case (ii).

The proof above does not work because the "obvious" part is actually nontrivial to prove. Let us show the "real" proof. Since C is closed, then $\mathbb{R}^m \setminus C$ is open and contains b, there is $\varepsilon > 0$ such that[59]

$$B_\varepsilon(b) \cap C = \emptyset.$$

Let H be the separating hyperplane of C and $B_\varepsilon(b)$. There there is $\lambda \in \mathbb{R}^m$ and $r \in \mathbb{R}$ such that

$$H = \{y \in \mathbb{R}^m : \lambda \cdot y = r\},$$

with

$$C \subset \{y \in \mathbb{R}^m : \lambda \cdot y \geq r\},$$

and $z \cdot \lambda \leq r$ for all $z \in B_\varepsilon(b)$. We claim that we may assume without loss of generality[60] that $r = 0$. We justify this in the final paragraph of the proof. Then,

$$H = \{y \in \mathbb{R}^m : \lambda \cdot y = 0\},$$

[59] The use of ε and $B_\varepsilon(b)$ is purely technical. We need to do it to get the strict inequality in $b \cdot \lambda < 0$. On a first read, it might be useful to think about just showing $b \cdot \lambda \leq 0$. In this case, one can take the hyperplane separating b and C, which is much simpler.

[60] Intuitively, one just takes a hyperplane the hits the "point" of the cone tangently and has a "good" angle to not hit b. While this is heuristically clear, it is annoying to show. I suggest ignoring the technical point about r on a first read and just take for granted that we can take $r = 0$.

and, by definition of separating hyperplane,

$$C \subset \{y \in \mathbb{R}^m : \lambda \cdot y \geq 0\}.$$

Then this means precisely that $a_i \cdot \lambda \geq 0$ for each i; that is, $A^T\lambda \succeq 0$. On the other hand, since H is a separating hyperplane of C and $B_\varepsilon(b)$. In particular, we have

$$0 \geq \lambda \cdot \left(b + \varepsilon \frac{\lambda}{2|\lambda|}\right) = \lambda \cdot b + \frac{\varepsilon}{2}.$$

Thus, $b \cdot \lambda < 0$. This completes the proof under the assumption that $r = 0$.

All that remains to complete the proof is to justify that we may take $r = 0$ without loss of generality. This is a purely technical First, we note that $0 \in C$ so that $0 = 0 \cdot \lambda \geq r$. Hence,

$$B_\varepsilon(b) \subset \{y \in \mathbb{R}^m : y \cdot \lambda \leq 0\}. \tag{7.1.9}$$

Now, consider any $c \in C$. Then, $\alpha c \in C$ for any $\alpha \geq 0$ by the definition of a cone. We deduce that

$$\frac{r}{\alpha} \leq \frac{1}{\alpha}(\alpha c) \cdot \lambda = c \cdot \lambda.$$

Taking $\alpha \to \infty$, we deduce that $c \cdot \lambda \geq 0$. We conclude that

$$C \subset \{y \in \mathbb{R}^m : \lambda \cdot y \geq 0\},$$

which, along with (7.1.9), implies that

$$\tilde{H} := \{y \in \mathbb{R}^m : \lambda \cdot y = 0\}$$

is a separating hyperplane of $B_\varepsilon(b)$ and C. □

Using slack variables as we did in Section 7.1.1, we can translate this to two equivalent forms of Farkas's lemma:

Lemma 7.1.4 (Farkas's Lemma II)**.** *Let $A \in M_{m\times n}$ and $b \in \mathbb{R}^m$. Then exactly one of the two alternatives holds:*

(i) $Ax \preceq b$ has a solution $x \succeq 0$;

(ii) there is $\lambda \succeq 0$ such that $A^T\lambda \succeq 0$ and $b \cdot \lambda < 0$.

Lemma 7.1.5 (Farkas's Lemma III)**.** *Let $A \in M_{m\times n}$ and $b \in \mathbb{R}^m$. Then exactly one of the two alternatives holds:*

(i) $Ax \preceq b$ has a solution $x \in \mathbb{R}^n$;

(ii) there is $\lambda \succeq 0$ such that $A^T\lambda = 0$ and $b \cdot \lambda < 0$.

We are now in a position to prove Theorem 7.1.1. Before doing so, let us "translate" it to standard form, where Farkas's lemma is more useful. The following is equivalent to Theorem 7.1.1.

Theorem 7.1.6 (Duality Theorem)**.** *Let $c \in \mathbb{R}^n$, $A \in M_{m\times n}$, and $b \in \mathbb{R}^m$. For the linear programs*

$$\text{maximize } c \cdot x \text{ subject to } Ax = b \text{ and } x \succeq 0 \tag{P}$$

and

$$\text{minimize } b \cdot y \text{ subject to } A^T y \succeq c \text{ and } y \in \mathbb{R}^m, \tag{D}$$

there are the following alternatives, exactly one of which occurs:

(i) Neither (P) nor (D) is feasible;

(ii) (P) is unbounded and (D) is not feasible;

(iii) (P) is not feasible and (D) is unbounded;

(iv) Both (P) and (D) are feasible, have an optimal solution x^* and y^* (respectively), and

$$c \cdot x^* = b \cdot y^*. \tag{7.1.10}$$

Proof of Theorem 7.1.1. If neither (P) nor (D) is feasible, we are finished because clearly (ii), (iii), and (iv) cannot occur.

Let us first assume that (P) is feasible and (D) is not. If (P) is unbounded, then we are done. If it is bounded, then take x^* in the feasible set such that $P = c \cdot x^*$. Since x^* is the optimizer, we find that, for any $\varepsilon > 0$,

$$Ax \leq b \qquad c \cdot x \geq c \cdot x^* + \varepsilon, \qquad x \succeq 0$$

is not feasible. Let, for any $\varepsilon \geq 0$,

$$\tilde{A} = \begin{bmatrix} A \\ -c^T \end{bmatrix} \quad \text{and} \quad \tilde{b}_\varepsilon = \begin{bmatrix} b \\ -c \cdot x^* - \varepsilon \end{bmatrix}.$$

Then, when $\varepsilon > 0$,

$$\tilde{A}x \leq \tilde{b}_\varepsilon \quad \text{with } x \succeq 0, \tag{7.1.11}$$

is infeasible. Hence, by Lemma 7.1.4, there is $\tilde{\lambda} \in \mathbb{R}^{m+1}$ with $\tilde{A}^T\tilde{\lambda} \succeq 0$ and $\tilde{b}_\varepsilon \cdot \lambda < 0$.

When $\varepsilon = 0$, (7.1.11) is feasible. Hence, Lemma 7.1.4 implies that $b_0 \cdot \tilde{\lambda} \geq 0$ (since $\tilde{\lambda} \succeq 0$ and $\tilde{A}^T\lambda \succeq 0$, by above).

Writing $\tilde{\lambda} = (y, z)$ with $z \in \mathbb{R}$, the above means that

$$A^T y \succeq zc, \qquad y \succeq 0, \qquad z \geq 0, \quad \text{and} \quad zc \cdot x^* \leq b \cdot y < z(c \cdot x^* + \varepsilon). \tag{7.1.12}$$

The last two inequalities give us that $z > 0$. Let

$$\bar{y} = \frac{1}{z} y. \tag{7.1.13}$$

It follows that $\bar{y}$ is in the feasible set of (D), which contradicts that assumption that (D) is not feasible. This is exactly case (ii).

Next let us assume that (P) and (D) are feasible. Then, we argue exactly as above to obtain x^* and $\bar{y}$. Then (7.1.12) and (7.1.13) and

$$b \cdot \bar{y} \leq c \cdot x^* + \varepsilon.$$

This yields $D \leq c \cdot x^* + \varepsilon$. As this is true for every ε, we conclude that $D \leq c \cdot x^*$. In other words

$$P = D,$$

which concludes the proof of case (iv).

The proof of case (iii) is exactly that the same as case (ii), so we omit it. □

7.2. **Fourier-Motzkin elimination.** Similar in spirit to Gaussian elimination, Fourier-Motzkin elimination allows us to rewrite our constraints in a way that removes a variable. We see this through a few examples.

Consider the trivial problem

maximize	0
subject to	$2x_1 + x_2 \leq 3$
	$4x_1 + 8x_2 \leq 12$
	$x_1, x_2 \geq 0.$

Our goal is to check if this is feasible.

Let us "solve" these inequalities for x_2:

$$x_2 \leq 3 - 2x_1 \qquad \text{and} \qquad x_2 \leq \frac{3}{2} - \frac{x_1}{2}.$$

Since $x_2 \geq 0$, we can summarize this as

$$0 \leq 3 - 2x_1 \qquad \text{and} \qquad 0 \leq \frac{3}{2} - \frac{x_1}{2}.$$

In other words,

$$x_1 \leq \frac{3}{2} \qquad \text{and} \qquad x_1 \leq 3.$$

Since $x_1 \geq 0$, we see that the allowed range of x_1 is

$$0 \leq x_1 \leq \frac{3}{2}. \tag{7.2.1}$$

For each x_1, we can also find the allowed range of x_2's. Regardless, since (7.2.1) is feasible, then so is the original program!

Let us consider another trivial problem:

maximize	0
subject to	$2x_1 + x_2 \leq 3$
	$x_1 - x_2 \geq 3$
	$x_1, x_2 \geq 0.$

We again "solve" for x_2:

$$x_2 \leq 3 - 2x_1 \qquad \text{and} \qquad x_2 \leq x_1 - 3,$$

which, along with $x_2 \geq 0$, yields

$$0 \leq 3 - 2x_1 \qquad \text{and} \qquad 0 \leq x_1 - 3,$$

which yields

$$3 \leq x_1 \leq \frac{3}{2}.$$

This is clearly not feasible! Hence, the original program is not feasible.

The general procedure: Suppose that we have r variables $x_1, \ldots, x_r$ and n constraints. Let m of our constraints are upper bounds on x_r and $n-m$ are lower bounds on x_r. In other words, we have n affine maps $U_1, \ldots, U_m, L_1, \ldots, L_{n-m}$ such that

$$x_r \leq U_i(x_1, \ldots, x_{r-1}) \quad \text{and} \quad x_r \geq L_j(x_1, \ldots, x_{r-1}) \qquad \text{for all } i = 1, \ldots, m \text{ and } j = 1, \ldots, n-m.$$

For each i and j, we get a new constraint on just $x_1, \ldots, x_{n-1}$ given by:

$$L_j(x_1, \ldots, x_{r-1}) \leq U_i(x_1, \ldots, x_{r-1}).$$

Notice that we now have $m(n-m)$ constraints on $r-1$ variables.

A word of warning: in the two simple examples above, we ended up with essentially the same number of constraints as we had initially. This is atypical. In general, if you begin with n constraints, you can end up with as many as $n^2/4$ constraints after eliminating a variable! Why is this? Suppose that half of our constraints are upper bounds on x_n and half are lower bounds on x_n (i.e., $m = n/2$) . This leads to exactly $(n/2)(n/2) = n^2/4$ constraints on $x_1, \ldots, x_{r-1}$. The upside, though, is we have reduce the number of variables by one.

7.3. **Constrained Optimization.** A more general optimization problem is one of the form

minimize	$f_0(x)$	
subject to	$f_i(x) \leq 0$	$i = 1, \ldots, m$
	$h_j(x) = 0$	$j = 1, \ldots, p.$

This is a *convex problem* if each $f_0, f_1, \ldots, f_m$ are convex and $h_1, \ldots, h_p$ are affine. Of course, this is a special setting and need not be true in general.

Letting

$$D = \left(\bigcap_{i=0}^{m} \operatorname{dom} f_i \right) \cap \left(\bigcap_{j=1}^{p} \operatorname{dom} h_j \right).$$

The primal feasible set is

$$\mathcal{F} = D \cap \left(\bigcap_{i=1}^{m} \{x \in \operatorname{dom} f_i : f_i(x) \leq 0\} \right) \cap \left(\bigcap_{j=1}^{p} \{x \in \operatorname{dom} h_j : h_j(x) = 0\} \right).$$

Letting

$$p^* = \inf_{x \in \mathcal{F}} f_0(x),$$

we call $x^* \in \mathcal{F}$ an *optimal point* if $f_0(x^*) = p^*$, although this is not guaranteed to exist.

To access the dual problem, first we define the *Lagrangian*

$$\mathcal{L} : D \times \mathbb{R}^m_+ \times \mathbb{R}^p \to \mathbb{R}$$

for this optimization problem by

$$\mathcal{L}(x, \lambda, \mu) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{j=1}^{p} \mu_j h_j(x).$$

The dual variables are (λ, μ) with the constraint $\lambda \succeq 0$. The dual function is

$$g(\lambda, \mu) = \inf_{x \in D} \mathcal{L}(x, \lambda, \mu).$$

Since g is the infimum of a family of linear functions then it is always *concave* (which is good for finding a maximum!).

Notice that we still have weak duality:

$$g(\lambda, \mu) \leq p^* \tag{7.3.1}$$

for every $\lambda \succeq 0$ and $\mu \in \mathbb{R}^p$, and, hence,

$$d^* = \sup_{\lambda \succeq 0, \mu \in \mathbb{R}^p} g(\lambda, \mu) \leq p^*.$$

We check (7.3.1) now. Take any feasible point $x \in \mathcal{F}$ and $\lambda \succeq 0$. Then

$$g(\lambda, \mu) \leq \mathcal{L}(x, \lambda, \mu) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{j=1}^{p} \mu_j h_j(x) \leq f_0(x), \tag{7.3.2}$$

where the inequality follows from the fact that $f_i(x) \leq 0$, $\lambda_i \geq 0$, and $h_j(x) = 0$. Taking the infimum of both sides in (7.3.2), we find

$$g(\lambda, \mu) \leq \inf_{x \in \mathcal{F}} f_0(x) = p^*.$$

In general, unfortunately, it will not be that $d^* = p^*$. We call $p^* - d^*$ the *duality gap*.

Example 7.3.1. *(i)* **Linear least squares***:*

minimize	$\|x\|^2$
subject to	$Ax = b.$

We begin by writing the Lagrangian:

$$\mathcal{L}(x, \mu) = |x|^2 + \mu \cdot (Ax - b).$$

To compute the dual function, we notice that

$$\nabla_x \mathcal{L}(x, \mu) = 2x + A^T \mu,$$

so that

$$\begin{aligned} g(\mu) &= \frac{1}{4}|A^T\mu|^2 + \mu \cdot \left(A\left(-\frac{1}{2} A^T \mu \right) - b \right) = \frac{1}{4}|A^T\mu|^2 - \frac{1}{2}|A^T\mu|^2 - \mu \cdot b \\ &= -\frac{1}{4}|A^T\mu|^2 - \mu \cdot b \end{aligned}$$

We will see, later, that the duality gap for this problem is zero; that is, $d^* = p^*$. The reason for this is because the problem is convex and has no non-affine constraints. In this case, it can be easily checked. In fact, this is exactly the strategy of finding an extremum via Lagrange multipliers.

Exercise 7.3.1. *Think about this!*

(ii) **Linear programming:**

minimize	$c \cdot x$
subject to	$Ax = b,\ x \succeq 0.$

To put this into the form of the general constrained optimization problem above, we "switch" the inequality $x \succeq 0$:

minimize	$c \cdot x$
subject to	$Ax = b,\ -x \preceq 0.$

We begin by writing the Lagrangian:

$$\mathcal{L}(x, \lambda, \mu) = c \cdot x - \lambda \cdot x + \mu \cdot (Ax - b).$$

To compute the dual function, we notice that, if $c - \lambda + A^T\mu = 0$, then

$$\mathcal{L}(x, \lambda, \mu) = -\mu \cdot b,$$

while, if $c - \lambda + A^T\mu \neq 0$, we can take $x = -r(c - \lambda + A^T\mu)$ with $r \to \infty$ to find

$$\inf_x \mathcal{L}(x, \lambda, \mu) \leq \liminf_{r\to\infty} \left(-r|c - \lambda + \mu^T A|^2 - \mu \cdot b\right) = -\infty.$$

Hence

$$g(\lambda, \mu) = \begin{cases} -\mu \cdot b & \text{if } c - \lambda + A^T\mu = 0, \\ -\infty & \text{if } c - \lambda + A^T\mu \neq 0. \end{cases} \tag{7.3.3}$$

Notice that, $c - \lambda + A^T\mu = 0$ is equivalent to $A^T\mu = -c + \lambda$, which is equivalent to

$$A^T\mu \preceq -c.$$

in view of the restriction $\lambda \succeq 0$. Since μ can be any vector, we may swap $\mu \mapsto -\mu$ to find

$$A^T\mu \succeq c.$$

Hence, maximizing (7.3.3) *amounts to solving the problem*

maximize	$b \cdot y$
subject to	$A^T y \succeq c.$

which is exactly the dual problem we would have found by our old duality method.

(iii) **Linear constraints:**

minimize	$f_0(x)$
subject to	$Ax \leq b$
	$Cx = d.$

We begin by writing the Lagrangian:

$$\mathcal{L}(x, \lambda, \mu) = f_0(x) + \lambda \cdot (Ax - b) + \mu \cdot (Cx - d).$$

Let us compute the dual function:

$$\begin{aligned} g(\lambda, \mu) &= \inf_x \mathcal{L}(x, \lambda, \mu) = \inf_x \left(f_0(x) + \lambda \cdot (Ax - b) + \mu \cdot (Cx - d)\right) \\ &= -\sup_x \left((-\lambda^T A - \mu^T C) \cdot x - f_0(x)\right) - \lambda \cdot b - \mu \cdot d \\ &= -f_0^*(-\lambda^T A - \mu^T C) - \lambda \cdot b - \mu \cdot d. \end{aligned}$$

The dual problem is thus to maximize $-f_0^*(-\lambda^T A - \mu^T C) - \lambda \cdot b - \mu \cdot d$, or, equivalently, minimize $f_0^*(-\lambda^T A - \mu^T C) + \lambda \cdot b + \mu \cdot d$. Hence, the dual problem is exactly a convex optimization problem! (Even if f_0 is not convex.)

7.4. Conditions for optimality: constraint qualifications.

7.4.1. **Karush–Kuhn–Tucker (KKT) condition.** Suppose that f_i, h_j are differentiable, and that the duality gap is zero. Let x^* is a primal optimal point and (λ^*, μ^*) is a dual optimal point. Then, we have

$$\begin{aligned} f_0(x^*) = g(\lambda^*, \mu^*) &= \inf_x \left(f_0(x) + \lambda^* \cdot F(x) + \mu^* \cdot H(x)\right) \\ &\leq f_0(x^*) + \lambda^* \cdot F(x^*) + \mu^* \cdot H(x^*) \leq f_0(x^*). \end{aligned}$$

where we let F and H be the vector of f_i's and h_j's, respectively. Hence, all above inequalities must actually be equalities. This tell us a few things:

- (Complementary slackness): $\lambda_i^* f_i(x^*) = 0$ – note the above gives $\lambda \cdot F(x^*) = 0$, but recall that all terms in the two vectors are nonnegative, so the stronger claim holds;
- (Stationarity in x): x^* is the location of a minimum of $\mathcal{L}(x, \lambda^*, \mu^*)$. This yields

$$0 = \nabla_x \mathcal{L}(x^*, \lambda^*, \mu^*) = \nabla_x f_0(x^*) + \sum_{i=1}^m \lambda_i^* \cdot \nabla_x f_i(x^*) + \sum_{j=1}^p \mu_j^* \nabla_x h_j(x^*).$$

This leads us to the *KKT conditions*: any primal optimal point x^* and dual optimal point (λ^*, μ^*) must satisfy complementary slackness, stationarity in x, and

- (primal feasibility): $f_i(x^*) \leq 0$ and $h_j(x^*) = 0$ for all i, j;
- (dual feasibility): $\lambda^* \succeq 0$;

This of this as a "first derivative test" for linear programs: if you are searching for an optimal point, one way is to find all solutions of the four conditions above and then check them for optimality.

We note two things. First, when there are no inequality constraints $f_i \leq 0$, this is simply the method of Lagrange multipliers. Second, the point (x^*, λ^*, μ^*) is almost never an extremum of $\mathcal{L}$, it is usually a saddle point. This is hinted at by the fact that (λ^*, μ^*) is found by taking an infimum (over x) and then a supremum (over λ, μ). Numerically, attempting to minimize $\mathcal{L}$ is *not* a useful approach.

It turns out that, when the problem is convex, the conditions above are sufficient for (x^*, λ^*, μ^*) to be optimal points.

Theorem 7.4.1. *Consider the primal problem above. If x^* is a primal optimal point and (λ^*, μ^*) is a dual optimal point, then they satisfy the KKT conditions (primal feasible, dual feasible, complementary slackness, and stationarity in x).*

If the primal problem is convex, then any (x^*, λ^*, μ^*) satisfying the KKT conditions is optimal (primal and dual) and there is zero duality gap.

Proof. The first statement is essentially proven above, so we omit it here. Let us consider the second statement. Since $\mathcal{L}(\cdot, \lambda^*, \mu^*)$ is convex in the x argument – this uses the convexity of the problem and the nonnegativity of λ^* (dual feasibility), any critical point must be the location of a minimum. Hence,

$$\mathcal{L}(x^*, \lambda^*, \mu^*) = \inf_x \mathcal{L}(x, \lambda^*, \mu^*) = g(\lambda^*, \mu^*). \tag{7.4.1}$$

On the other hand, using complementary slackness and the primal feasibility of the problem, we find

$$\mathcal{L}(x^*, \lambda^*, \mu^*) = f_0(x^*) + \lambda^* \cdot F(x^*) + \mu^* \cdot H(x^*) = f_0(x^*). \tag{7.4.2}$$

We conclude from (7.4.1)-(7.4.2) and weak duality that

$$d^* \geq g(\lambda^*, \mu^*) = f_0(x^*) \geq p^* \geq d^*.$$

Hence all above inequalities are equalities, which completes the proof. □

Exercise 7.4.1. *Given a $m \times n$ matrix A and considering it as an operator with the ℓ^2-norm, the induced norm of A is*

$$\|A\| = \sup_{\|x\|=1} \|Ax\|.$$

Interpret this as a constrained optimization problem and apply the Lagrange multiplier / KKT approach to find potential optimal points. Is the critical point of the Lagrangian a local extremum?

Example 7.4.2 (Water filling)**.**

minimize	$-\sum_{i=1}^{n} \log(\alpha_i + x_i)$
subject to	$x \succeq 0$, $\sum_{i=1}^{n} x_i = 1.$

Here $\alpha \succeq 0$ is a fixed vector. This problem arises in information theory – x_i is the power transmitted to the nth channel and $\log(\alpha_i + x_i)$ is the capacity (communication rate) of each channel. Clearly we wish to maximize the communication rate.

Notice that this is a convex problem. Hence, if we find a point satisfying the KKT conditions, we have an optimal point. In this case,

$$f_i(x) = -x_i, \quad h_j(x) = h(x) = \sum_{i=1}^{n} x_i$$

so the conditions are:

- *(Primal feasibility): $f_i(x^*) = -x_i^* \leq 0$ for $i = 1, \ldots, n$, and $h(x^*) = \sum_{i=1}^{n} x_i^* - 1 = 0$;*
- *(dual feasibility): $\lambda_j^* \geq 0$ for each $j = 1, \ldots, n$;*

- *(Complementary slackness):* $\lambda_i^* x_i^* = 0$ *for all* $i = 1, \dots, n$;
- *(Stationarity in* x*):* $0 = -\frac{1}{\alpha_i + x_i^*} - \lambda_i^* + \mu^*$.

Let us consider first the case where $\mu^* < 1/\alpha_i$. *Then the stationarity-in-*x *condition can only hold if* $x_i^* > 0$. *Complementary slackness thus implies that* $\lambda_i^* = 0$ *and, hence,*

$$\frac{1}{\alpha_i + x_i^*} = \mu^*, \qquad \text{that is,} \qquad x_i^* = \frac{1}{\mu^*} - \alpha_i.$$

On the other hand, if $\mu^* > 1/\alpha_i$, *then*

$$\mu^* > \frac{1}{\alpha_i} \geq \frac{1}{\alpha_i + x_i^*} = \mu^* - \lambda_i^*.$$

It follows that $\lambda_i^* > 0$, *whence* $x_i^* = 0$ *by the complementary slackness condition. Finally, one can easily extend the above reasoning to find that* $\mu^* = 1/\alpha_i$ *implies that* $x_i^* = 0$. *We compile the above in the succinct formula*

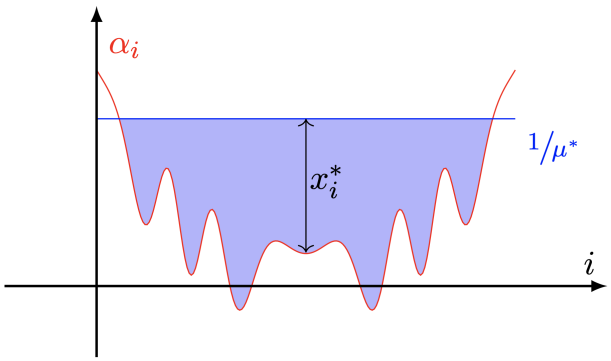
$$x_i^* = \max\{0, 1/\mu^* - \alpha_i\}.$$

Plugging this into the condition for primal feasibility, we arrive at

$$\sum_{i=1}^{n} \max\{0, 1/\mu^* - \alpha_i\} = 1.$$

There is clearly a unique solution $\mu^* > 0$ *to this.*

Why is this called "water filling"? Imagine that α_i *is the elevation of land in some patch* $i \in \{1, \dots, n\}$. *Then we have volume 1 of water that we pour over all patches. The fluid as a whole will settle at the same height everywhere (which is height* $1/\mu^*$*), thus filling the deepest holes first.*



7.4.2. **Slater's condition.** Slater's condition is a sufficient condition for strong duality to hold (that is, there is zero duality gap). Let us take a convex optimization problem

minimize	$f_0(x)$
subject to	$f_i(x) \leq 0 \quad i = 1, \dots, m$ $Ax = b.$

Slater's condition: there is a *strictly feasible* $x \in \text{relint}(D)$, where we define strictly feasible to mean that

$$Ax = b \qquad \text{and} \qquad f_i(x) < 0 \quad \text{for all } i = 1, \dots, m,$$

and the relative interior[61] is defined as follows: $x \in \text{relint}(C)$ if there is $r > 0$ such that

$$B_r(x) \cap \text{aff}(C) \subset C.$$

Actually, one does not need strict feasibility to hold for all f_i, only the ones that are not affine.

Theorem 7.4.3 (Strong duality with Slater's condition)**.** *If Slater's condition holds at some point x_{sl}, then the duality gap is zero: $p^* = d^*$ and there is a dual optimal point (λ^*, μ^*).*

Proof. We begin with a few simplifications. First, we assume that p^* is finite. By the existence of x_{sl}, it can only be $p^* = -\infty$ or a finite value. If it is $-\infty$, then weak duality implies that $d^* = -\infty$ and we are finished.

Second, we assume that D has nonempty interior. This assumption on D is not much of a restriction – we are essentially saying that each f_i and h_j are defined on "most" of $\mathbb{R}^n$, which is a reasonable assumption. It is made for simplicity.

Finally, we remove the condition $Ax = b$. Indeed, up to translating by x_{sl} and applying a bijective linear transformation $T : \mathbb{R}^d \to \text{Ker}(A)$, where d is the rank of A, our problem becomes

minimize	$f_0(Ty + x_{\text{sl}})$
subject to	$f_i(Ty + x_{\text{sl}}) \leq 0 \quad i = 1, \dots, m$ $y \in \mathbb{R}^d$.

This remains a convex problem satisfying the assumptions of the problem.

The basis for our proof is the separating hyperplane theorem (Proposition 6.2.3). Let us define our two convex sets on which we apply the theorem. Our first set is

$$\mathcal{A} = \{(F, t) \in \mathbb{R}^m \times \mathbb{R}^p \times \mathbb{R} : \text{for all } x \in D, F \succeq (f_1(x), \dots, f_m(x))t \geq f_0(x)\}.$$

It is easy to check the convexity of $\mathcal{A}$ using the convexity of the problem.

Exercise 7.4.2. *Do this!*

The second set is

$$\mathcal{B} = \{(0, t) \in \mathbb{R}^m \times \mathbb{R} : t < p^*\}.$$

This is also clearly convex as it is a half-line.

We check that $\mathcal{A} \cap \mathcal{B} = \emptyset$ so that we can separate them by a hyperplane. If $(F, t) \in \mathcal{A} \cap \mathcal{B}$, then the definition of $\mathcal{B}$ and $\mathcal{A}$ imply that

$$t < p^* \leq f_0(x_{\text{sl}}) \leq t,$$

which is a contradiction.

[61] This is, roughly, the same as saying the following. Let $T : \mathbb{R}^n \to \mathbb{R}^p$ be the affine function whose restriction $T : \text{aff}(C) \to \mathbb{R}^p$ is bijective. Then $x \in \text{relint}(C)$ if and only if $T(x) \in \text{Int}(T(C))$. A careful definition is required because, if $\text{aff}(C)$ is a lower-dimensional space, it cannot have an interior even if we, heuristically, identify certain points as being in the interior.

Hence, we can apply Proposition 6.2.3 to obtain λ, s, and α such that

$$\mathcal{A} \subset \{(u, v, t) : u \cdot \lambda + ts \geq \alpha\} \tag{7.4.3}$$

and

$$\mathcal{B} \subset \{(u, t) : u \cdot \lambda + ts \leq \alpha\}, \tag{7.4.4}$$

where

$$(\lambda, m) \neq 0. \tag{7.4.5}$$

Since (7.4.5) is a crucial point in the proof it is worth clarifying why this is true. Suppose $\alpha \geq 0$ and (7.4.5) does not hold, i.e., $\lambda = 0$ and $m = 0$. It is clear, then, that

$$\mathcal{A}, \mathcal{B} \subset \{(u, t) : u \cdot 0 + t0 \leq \alpha\},$$

which contradicts the definition of separating hyperplane. The case when $\alpha \leq 0$ is similar.

From (7.4.3), we see that $s \geq 0$. Indeed, if $s < 0$, we can take $t \to \infty$ and $st \to -\infty$, which would contradict the inequality in (7.4.3). A similar argument shows that $\lambda \succeq 0$. Hence,

$$\lambda \succeq 0 \quad \text{and} \quad s \geq 0.$$

On the other hand, the condition in (7.4.4) implies that $ts \leq \alpha$ for all $t < p^*$. It follows that

$$p^* m \leq \alpha.$$

Using the definition of $\mathcal{A}$ with this, we find, for any $x \in D$,

$$\sum_{i=1}^{m} \lambda_i f_i(x) + s f_0(x) \geq \alpha \geq s p^*. \tag{7.4.6}$$

We now break into cases depending on whether $s = 0$ or not.

Case one: $s > 0$**.** From (7.4.6) and the definition of $\mathcal{L}$, we find, for any $x \in D$,

$$\mathcal{L}(x, \lambda^*) \geq p^*, \qquad \text{where } \lambda^* = \frac{\lambda}{s}.$$

This, along with weak duality, implies that

$$d^* \leq p^* \leq \mathcal{L}(x, \lambda^*).$$

Taking the infimum of both sides in x, we find

$$d^* \leq p^* \leq g(\lambda^*) \leq d^*,$$

where the last inequality holds because d^* is the supremum of g. Since the right and left hand sides of the above are the same, all inequalities must be equalities, whence $d^* = p^*$ and the proof is finished with λ^* as a dual optimal point.

Case two: $s = 0$**.** Then we have, by (7.4.6),

$$\sum_{i=1}^{m} \lambda_i f_i(x_{\text{sl}}) \geq 0.$$

By the feasibility of x_{sl} and the nonnegativity of λ, each term in the sum above is nonpositive. Hence, they must all be zero: $\lambda_i f_i(x_{\text{sl}}) = 0$ for all $i = 1, \ldots, m$. By the strict feasibility of x_{sl}, it follows that $\lambda_i = 0$ as $f_i(x_{\text{sl}}) < 0$ for each i. We have now reached a contradiction to (7.4.5) since we have shown that $\lambda = 0$ and $s = 0$. This concludes the proof. □

7.5. **Relaxation.** As we have seen with the KKT conditions and Slater's condition, convexity is very nice to have. Unfortunately, not all constrained optimization problems are convex. However, let us notice something. Take a primal problem with value p^*

$$\text{minimize } f_0(x) \quad \text{subject to } f_i(x) \leq 0,\ h_j(x) = 0 \quad \text{for all } i = 1, \dots, m,\ j = 1, \dots, p.$$

Let its dual problem with value d^* be given by

$$\text{maximize } g(\lambda, \mu) \quad \text{subject to } \lambda \succeq 0.$$

Consider the new primal problem, with value $\tilde{p}^*$, given by

$$\text{minimize } -g(\lambda, \mu) \quad \text{subject to } -\lambda \preceq 0.$$

Clearly $-d^*$ is its value. Let us compute its dual problem. The Lagrangian is

$$\mathcal{L}(\lambda, \mu, z) = -g(\lambda, \mu) - \lambda \cdot z.$$

Taking its infimum over all λ, μ, we find

$$F(z) = \inf_{\lambda,\mu} \left(-g(\lambda, \mu) - \lambda \cdot z\right) = -\sup_{\lambda,\mu} \left((z, 0) \cdot (\lambda, \mu) - (-g)(\lambda, \mu)\right) = -(-g)^*(z, 0).$$

Hence, the dual problem is

$$\text{maximize } F(z) \quad \text{subject to } z \succeq 0,$$

which, up to a minus sign, is equivalent to

$$\text{minimize } (-g)^*(z, 0) \quad \text{subject to } -z \preceq 0.$$

This a convex problem! Let $\tilde{p}^*$ be its optimal value. Then

$$\tilde{p}^* = \inf_z (-g)^*(z, 0) \leq (-g)^*(0, 0) = \left(\sup_{\lambda,\mu} \left(g(\lambda, \mu) + \lambda \cdot 0\right)\right) = d^* \leq p^*.$$

The last inequality holds by weak duality.

In other words, by taking the "double dual," we end up with a problem that is convex and provides a lower bound for our primal problem!

Exercise 7.5.1. *Perform this procedure for the constrained optimization problem: minimize $f(x) = (1 - x^2)^2$ subject to $x \in \mathbb{R}$. What is the objective function for the resulting problem and what is its relationship to f?*

A. The Cantor-Schöder-Bernstein theorem

Here we provide a proof of Theorem 3.2.8. Let us restate this for convenience.

Theorem A.1. *Suppose that X and Y are sets such that there exist injections*

$$f : X \to Y \qquad \text{and} \qquad g : Y \to X.$$

Then $\operatorname{card}(X) = \operatorname{card}(Y)$.

This proof is kind of a behemoth. Let me stress, however, that one uses the same "trick" over and over again. Hence, most of the length below is redoing the same work in slightly different contexts.

Proof. We need to use f and g to construct a bijection $\varphi : X \to Y$. Let us define equivalence classes as follows. For each $x \in X$, we set

$$x \sim_X \tilde{x} \text{ if and only if there is } n \in \mathbb{N} \cup \{0\} \text{ such that } (g \circ f)^n(x) = \tilde{x} \text{ or } (g \circ f)^n(\tilde{x}) = x.$$

This is clearly an equivalence relation (it is symmetric, reflexive, and associative – recall Definition 3.3.5 and check this on your own!). We define φ on each equivalence class

$$[x]_X = \{\tilde{x} \in X : x \sim_X \tilde{x}\}.$$

We define also an equivalence relation on Y:

$$y \sim_Y \tilde{y} \text{ if and only if there is } n \in \mathbb{N} \cup \{0\} \text{ such that } (f \circ g)^n(y) = \tilde{y} \text{ or } (f \circ g)^n(\tilde{y}) = y.$$

Let $\tilde{X}$ and $\tilde{Y}$ be sets containing one representative from each equivalence class. Then

$$X = \bigcup_{\tilde{x} \in \tilde{X}} [\tilde{x}]_X \qquad \text{and} \qquad Y = \bigcup_{\tilde{y} \in \tilde{Y}} [\tilde{y}]_Y$$

and both unions are unions of *disjoint* sets.

Notice that there is a correspondence between equivalence classes. Indeed, we may associate $[x]_X$ with $[f(x)]_X$ and $[g(y)]_X$ with $[y]_Y$. Hence, if we can construct φ such that, for every x,

$$\varphi|_{[x]_X}[x]_X \to [f(x)]_X$$

is a bijection, then $\varphi : X \to Y$ is a bijection.

Case one: $[x]_X$ is finite.. Roughly, this is the case where applying f and g iteratively yields a finite cycle. We let $\varphi(x) = f(x)$. Let us check that $\varphi|_{[x]_X} : [x]_X \to [f(x)]_Y$ is a bijection. It is clearly injective (this is inherited from f).

On the other hand, notice that $(g \circ f)^n(x)$ cannot be infinite, so there is n, N such that $(g \circ f)^{n+N}(x) = (g \circ f)^n(x)$. Because g and f are injective, so is $(g \circ f)^n$, which implies that

$$(g \circ f)^N(x) = x.$$

Further, by iterating this, we see that, for all $k \in \mathbb{N}$,

$$(g \circ f)^{kN}(x) = x. \tag{A.1}$$

Fix any $y \in [f(x)]_Y$. if $y = f(x)$, then it is clear that $\varphi(x) = y$ and we are finished. Otherwise, there is $n \in \mathbb{N}$ such that either

$$(f \circ g)^n(y) = f(x) \text{ or } (f \circ g)^n(f(x)) = y. \tag{A.2}$$

Consider the case where the first equality holds. Then, by the injectivity of f, we have

$$x = g \circ (f \circ g)^{n-1}(y) = (g \circ f)^{n-1}(g(y)).$$

Fix any k such that $kN > n+1$. Then, by (A.1),

$$(g \circ f)^{kN-(n-1)}(x) = (g \circ f)^{kN-(n-1)}(g \circ f)^{n-1}(g(y)) = g(y).$$

By the injectivity of g, we see that

$$y = (f \circ g)^{kN-(n-1)-1}(f(x)) = (f \circ g)^{kN-n}(f(x)) = f\left((g \circ f)^{kN-n-1}(x)\right).$$

Since $(g \circ f)^{kN-n-1}(x) \in [x]_X$, this shows that $y = \varphi|_{[x]_X}(\tilde{x})$ for some $\tilde{x} \in [x]_X$.

The case where the second equality in (A.2) holds can be handled similarly. We omit the proof.

Case two: there exists $x_0 \in [x]_X$ such that $x_0 \notin g(Y)$. Roughly, the equivalence class has a first element that generates the whole equivalence class after applying f and g iteratively.

For every n, let $x_n = (g \circ f)^n(x_0)$. Let us first show that $x_n \neq x_m$ whenever $n \neq m$. Assume, without loss of generality, that $n < m$. Then

$$(g \circ f)^n(x_0) = x_n = x_m = (g \circ f)^m(x_0)$$

and, by injectivity,

$$x_0 = (g \circ f)^{m-n}(x_0) = g\left((f \circ g)^{m-n-1}(x_0)\right).$$

This is a contradicts the fact that $x_0 \notin g(Y)$. Hence, $x_n \neq x_m$.

Next, we claim that $[x]_X = \{x_0, x_1, x_2, \dots\}$. Clearly "$\supset$" holds, so we show that "$\subset$" holds as well. Fix any $\tilde{x} \in [x]_X$. Since $\sim_X$ is an equivalence relation, there is n such that

$$(g \circ f)^n(x_0) = \tilde{x} \qquad \text{or} \qquad (g \circ f)^n(\tilde{x}) = x_0.$$

The latter cannot hold because $x_0 \notin g(Y)$. Hence, the former holds, which implies that $\tilde{x} = x_n$. This concludes the proof that $[x]_X = \{x_0, x_1, x_2, \dots\}$.

We now define

$$\varphi(x_n) = f(x_n).$$

Let us check that $\varphi|_{[x]_X} : [x]_X \to [f(x)]_Y$ is a bijection. It is, again, clearly an injection because f is. We check that $\varphi|_{[x]_X}$ is surjective. Fix $y \in [f(x_0)]_Y$. If $y = f(x_0)$, then we are finished. Otherwise, there is $n \in \mathbb{N}$ such that either

$$(f \circ g)^n(y) = f(x_0) \qquad \text{or} \qquad (f \circ g)^n(f(x_0)) = y.$$

The former cannot hold because, by the injectivity of f, it follows that $g \circ (f \circ g)^{n-1}(y) = x_0$, which is a contradiction to the assumption that $x_0 \notin g(Y)$. Hence the latter holds, which implies that

$$y = (f \circ g)^n(f(x_0)) = f\left((g \circ f)^{n-1}(x_0)\right) = f(x_{n-1}).$$

This concludes the proof of surjectivity.

Case three: there exists $x_0 \in [x]_X$ such that $x_0 \in g(Y) \setminus g(f(X))$. This, again, is roughly the case where the equivalence class has a first element that generates the whole equivalence class after applying f and g iteratively. Note, though, that φ is defined differently here than in the previous case.

Let $y_0 \notin f(X)$ be such that $g(y_0) = x_0$. For every n, let

$$x_n = (g \circ f)^n(x_0).$$

First we claim that $x_n \neq x_m$ whenever $n \neq m$. Assume, without loss of generality, that $n < m$. Then

$$(g \circ f)^n(x_0) = x_n = x_m = (g \circ f)^m(x_0)$$

and, by injectivity,

$$x_0 = (g \circ f)^{m-n}(x_0) = g\left((f \circ g)^{m-n-1}(x_0)\right).$$

By injectivity, $(f \circ g)^{m-n-1}(x_0) = y_0$; however, this contradicts the fact that $y_0 \notin f(Y)$.

Next, we claim that $[x]_X = \{x_0, x_1, x_2, \dots\}$. Clearly "$\supset$" holds, so we show that "$\subset$" holds as well. Fix any $\tilde{x} \in [x]_X$. Since $\sim_X$ is an equivalence relation, there is n such that

$$(g \circ f)^n(x_0) = \tilde{x} \qquad \text{or} \qquad (g \circ f)^n(\tilde{x}) = x_0.$$

If $n = 0$, then $\tilde{x} = x_0$ and we are finished. If $n \geq 1$, the second equality above cannot hold because $x_0 \notin g(f(X))$. Hence, the former holds, which implies that $\tilde{x} = x_n$. This concludes the proof that $[x]_X = \{x_0, x_1, x_2, \dots\}$.

We now define

$$\varphi(x_n) = \begin{cases} f(x_{n-1}) & \text{if } n > 1, \\ y_0 & \text{if } n = 0. \end{cases}$$

Let us check that $\varphi|_{[x]_X} : [x]_X \to [f(x)]_Y$ is a bijection.

First we show that it is an injection. If $\varphi(x_n) = \varphi(x_m)$ and $n, m \neq 0$, we have

$$f(x_{n-1}) = f(x_{m-1}),$$

which, by injectivity of f, implies that $x_{n-1} = x_{m-1}$. By the work above, this can only occur if $n = m$, finishing the proof. If $n = 0 = m$, then we are finished. Hence, assume that, without loss of generality $n = 0$ and $m > 0$. Then $y_0 = f(x_m)$, which contradicts the fact that $y_0 \notin f(X)$. It follows that this case cannot occur. This concludes the proof of the injectivity of φ.

We check that $\varphi|_{[x]_X}$ is surjective. Fix $y \in [f(x_0)]_Y$. There is $n \in \mathbb{N} \cup \{0\}$ such that either

$$(f \circ g)^n(y) = f(x_0) \qquad \text{or} \qquad (f \circ g)^n(f(x_0)) = y. \tag{A.3}$$

If $n = 0$, then $y = f(x_0)$. We are then finished because $\varphi(x_1) = f(x_0)$.

Suppose that the first equality in (A.3) holds. By the injectivity of f, it follows that

$$g \circ (f \circ g)^{n-1}(y) = x_0.$$

If $n = 1$, we get that $g(y) = x_0 = g(y_0)$, which implies that $y = y_0$. By construction, $\varphi(x_0) = y_0$, and we are finished. If $n > 1$, which is a contradiction to the assumption that $x_0 \notin g(f(X))$. This completes the proof in the case when the first equality in (A.3) holds.

Suppose that the second equality in (A.3) holds for $n \geq 1$. Then

$$y = f\left((g \circ f)^n(x_0)\right) = f(x_n) = \varphi(x_{n+1}),$$

which completes the proof that $\varphi|_{[x]_X}$ is surjective.

Case four: $[x]_X$ is infinite, is a subset of $g(f(X))$, and there exists n, m and x_0 such that $(g \circ f)^{n+m}(x_0) = (g \circ f)^n(x_0)$. Roughly, this equivalence class has a "last element" x_0 but extends infinitely "backwards" by iteratively applying g^{-1} and f^{-1}.

First, we show that $(g \circ f)(x_0) = x_0$. Arguing as in case 1, we see that $(g \circ f)^m(x_0) = x_0$. It follows that the set

$$\{x_0, (g \circ f)(x_0), (g \circ f)^2(x_0), \dots\} \tag{A.4}$$

has at most m elements and that, for any $k \in \mathbb{N}$,

$$(g \circ f)^{km}(x_0) = x_0.$$

Let $\tilde{x} \in [x]_X$. Then there is n such that either

$$(g \circ f)^n(x_0) = \tilde{x} \qquad \text{or} \qquad (g \circ f)^n(\tilde{x}) = x_0.$$

In the latter case, take k such that $km > n$ and

$$(g \circ f)^n(\tilde{x}) = (g \circ f)^{km}(x_0),$$

which, by injectivity, implies that

$$\tilde{x} = (g \circ f)^{km-n}(x_0).$$

Hence, in both cases, $\tilde{x}$ is in the set defined in (A.4). This implies that $[x]_X$ has at most m elements, which is a contradiction. We deduce that $(g \circ f)(x_0) = x_0$.

Next, we show that there are $x_0, x_1, x_2, \dots$ such that

$$(g \circ f)(x_k) = x_{k-1} \qquad \text{for all } k \in \mathbb{N}.$$

We do this inductively. First, take any $\tilde{x} \in [x]_X \setminus \{x_0\}$, which must exist because $[x]_X$ is infinite. As usual, there exists n such that

$$(g \circ f)^n(x_0) = \tilde{x} \qquad \text{or} \qquad (g \circ f)^n(\tilde{x}) = x_0.$$

The former cannot hold because $(g \circ f)^n(x_0) = x_0$ for all n. Then, let

$$x_1 = (g \circ f)^{n-1}(\tilde{x}).$$

This has the required properties.

Now, assume that we already have defined $x_0, x_1, \dots, x_k$. Take any $\tilde{x} \in [x]_X \setminus \{x_0, x_1, \dots, x_k\}$. Again, there must be m such that

$$(g \circ f)^n(x_0) = \tilde{x} \qquad \text{or} \qquad (g \circ f)^n(\tilde{x}) = x_0.$$

As before, the former cannot hold, so we latter must. First, suppose that $n \leq k$. Then, $(g \circ f)^n(\tilde{x}) = (g \circ f)^n(x_n)$, which implies that $\tilde{x} = x_n$. This is a contradiction. Hence, $n > k$. Let

$$x_{k+1} = (g \circ f)^{n-k-1}(\tilde{x}).$$

Clearly,

$$(g \circ f)^{k+1}(x_{k+1}) = (g \circ f)^n(\tilde{x}) = x_0 = (g \circ f)^k(x_k).$$

By injectivity, we have that $(g \circ f)^k(x_{k+1}) = x_k$.

Finally, we show that $[x]_X = \{x_0, x_1, x_2, \dots\}$. Here, we argue exactly as above to see that, for any $\tilde{x} \in X$, there is n such that $(g \circ f)^n(\tilde{x}) = x_0$, which, by injectivity, implies that $\tilde{x} = x_n$.

We now define

$$\varphi(x_n) = f(x_{n+1}).$$

Let us check that $\varphi|_{[x]_X} : [x]_X \to [f(x)]_Y$ is a bijection.

It is an injection because, if $\varphi(x_n) = \varphi(x_m)$ for some n and m, then $g(x_{n+1}) = g(x_{m+1})$. This implies that $x_{n+1} = x_{m+1}$, which, by above, can only happen if $n = m$. We deduce that $x_n = x_m$.

Now we check that $\varphi|_{[x]_X}$ is surjective. Fix $y \in [f(x_0)]_Y$. If $y = f(x_0)$, then we are clearly finished since

$$g(y) = g(f(x_0)) = x_0 = (g \circ f)(x_1)$$

and, thus, $y = f(x_1)$, by the injectivity of g. If $y \neq f(x_0)$, then there is $n \in \mathbb{N}$ such that

$$(f \circ g)^n(f(x_0)) = y \qquad \text{or} \qquad (f \circ g)^n(y) = f(x_0).$$

In the former case, since $(g \circ f)^n(x_0) = x_0$, we have

$$f(x_0) = y,$$

which is a contradiction. In the latter case, we have

$$\begin{aligned}(f \circ g)^{n+1}(y) &= f\left((g \circ f)^n(g(y))\right) = f(x_0) \\ &= f\left((g \circ f)^n\right)(x_n) = (f \circ g)^n\left(f(x_n)\right),\end{aligned}$$

it follows, by injectivity, that $f(x_n) = y$. Thus $\varphi|_{[x]_X}$ is surjective.

Case five: $[x]_X$ is infinite, is a subset of $g(f(X))$, and $(g \circ f)^{n+m}(\tilde{x}) \neq (g \circ f)^n(\tilde{x})$ for every $\tilde{x} \in [x]_X$ and $n, m \in \mathbb{N}$.

In this case, we once again define $\varphi(\tilde{x}) = f(\tilde{x})$ for every $\tilde{x} \in [x]_X$. As usual, this is clearly injective.

Before we show that it is a surjection, we claim that

$$[x]_X = \{\dots, x_{-2}, x_{-1}, x_0, x_1, x_2, \dots\},$$

where $(g \circ f)(x_k) = x_{k+1}$. Let $x_0 = x$ and inductively define $x_n = (g \circ f)(x_{n-1})$ for each $n \in \mathbb{N}$. Since $x \in g(f(X))$, there must be $\tilde{x} \in X$ such that $(g \circ f)(\tilde{x}) = x$. Note that, if $\tilde{x} = x_k$ for some k, we have

$$x = (g \circ f)(\tilde{x}) = (g \circ f)(x_k) = (g \circ f)\left((g \circ f)^k(x)\right) = (g \circ f)^{k+1}(x),$$

which contradicts our hypothesis about $[x]_X$. Hence, $\tilde{x} \in [x]_X \setminus \{x_0, x_1, x_2, \dots\}$. Let $x_{-1} = \tilde{x}$. Repeating this argument yields x_{-2}, x_{-3}, etc.

We now have

$$[x]_X \supset \{\dots, x_{-2}, x_{-1}, x_0, x_1, x_2, \dots\}.$$

Let us show the other inclusion. Fix any $\tilde{x} \in [x]_X$. Then there is n such that

$$(g \circ f)^n(x_0) = \tilde{x} \qquad \text{or} \qquad (g \circ f)^n(\tilde{x}) = x_0.$$

In the former case, we see that $\tilde{x} = x_n$, and in the latter case we see that $\tilde{x} = x_{-n}$ (this uses the injectivity of $(g \circ f)^n$).

Now we show that φ is a surjection. Fix any $y \in [f(x_0)]_X$. If $y = f(x_0)$, we are finished. Otherwise, there is n such that either

$$(f \circ g)^n(y) = f(x_0) \qquad \text{or} \qquad (f \circ g)^n(f(x_0)) = y.$$

In the latter case, we have

$$f(x_n) = f\left((g \circ f)^n(x)\right) = y,$$

and we are finished. In the former case, we have

$$(f \circ g)^n(y) = f(x_0) = f\left((g \circ f)^n(x_{-n})\right) = (f \circ g)^n\left(f(x_{-n})\right).$$

By injectivity, we find $y = f(x_{-n})$, which concludes the proof in case five.

We have now exhausted all cases, which completes the proof of Theorem 3.2.8. □

B. Constructing the real numbers

We follow the outline for the general construction given in Section 3.3.1. Let

$$\mathbb{Q}_{\text{Cauchy}} = \left\{\bar{x} \in \mathbb{Q}^{\mathbb{N}} : \bar{x} \text{ is Cauchy}\right\}.$$

To be very careful, we should redefine Cauchy sequences using only $\varepsilon \in \mathbb{Q}$ (cf. Definition 3.3.1). Define an equivalence relation $\sim$ on $\mathbb{Q}_{\text{Cauchy}}$ by

$$\bar{x} \sim \bar{y} \quad \text{if and only if} \quad \lim_{n\to\infty}(x_n - y_n) = 0.$$

This is clearly an equivalence class (see Exercise 3.3.5).

Then we define the real numbers to be set of equivalence classes with respect this equivalence relation:

$$\mathbb{R} = \{[\bar{x}] : \bar{x} \in \mathbb{Q}_{\text{Cauchy}}\}.$$

A useful fact is that we may always take a representative $\bar{y}$ of $[\bar{x}]$ such that $\bar{y}$ is increasing. Similarly, we may find a decreasing representative. We prove this now. Fix $\bar{x}$. Let $N_0 = 0$ and, for each k, choose $N_k > N_{k-1}$ such that, for $n, m \geq N_k$, we have

$$|x_n - x_m| < \frac{1}{k}.$$

If $k = 1$, let $y_1 = x_{N_1} - 1$. Otherwise, let

$$y_k = \max\left\{y_{k-1}, x_{N_k} - \frac{1}{k}\right\}.$$

We easily observe that $y_k \geq y_{k-1}$. It follows that $\bar{y}$ is increasing. Additionally, one sees that

$$x_n - \frac{2}{k} \leq y_k \leq x_n \qquad \text{for all } n \geq N_k.$$

After a small amount of work (using the fact that $\bar{x}$ is Cauchy), it follows that $\bar{y} \sim \bar{x}$.

We note that "our" set $\mathbb{R}$ inherits all of the operations that we expect from real numbers:

(i) **(Addition)** We define $[\bar{x}] + [\bar{y}] = [\bar{x} + \bar{y}]$. Let us check that this is well-defined; indeed, we might worry that if we choose different representatives of $[\bar{x}]$ and $[\bar{y}]$, their sum might give a different equivalence class.

Let $\bar{a} \in [\bar{x}]$ and $\bar{b} \in [\bar{y}]$. Then

$$\lim_{k\to\infty}(a_k - x_k) = 0 \quad \text{and} \quad \lim_{k\to\infty}(b_k - y_k) = 0.$$

We must show that $[\bar{a} + \bar{b}] = [\bar{x} + \bar{y}]$.

We first claim that $[\bar{a} + \bar{b}] \subset [\bar{x} + \bar{y}]$. Indeed, letting $\bar{z} \in [\bar{a} + \bar{b}]$, we see that

$$\begin{aligned} 0 &= \lim_{k\to\infty}(z_k - (a_k + b_k)) = \lim_{k\to\infty}(z_k - (x_k + y_k)) + \lim_{k\to\infty}((x_k + y_k) - (a_k + b_k)) \\ &= \lim_{k\to\infty}(z_k - (x_k + y_k)) + \lim_{k\to\infty}(x_k - a_k) + \lim_{k\to\infty}(y_k - b_k) \\ &= \lim_{k\to\infty}(z_k - (x_k + y_k)). \end{aligned}$$

Thus $\bar{z} \in [\bar{x} + \bar{y}]$.

The argument that $[\bar{a} + \bar{b}] \supset [\bar{x} + \bar{y}]$ is similar, so we omit it.

Let us note that subtraction can be defined similarly, and the proof of well-definedness is clear. Finally, we make a note about notation. If $r \in \mathbb{Q}$ and $[\bar{x}] \in \mathbb{R}$, we use the shorthand

$$[\bar{x}] + r := [\bar{x}] + [(r, r, \dots)].$$

(ii) **(Multiplication)** We define $[\bar{x}] \cdot [\bar{y}] = [(x_1y_1, x_2y_2, \dots)]$. Let us check that this is well-defined. First we show that $(x_1y_1, x_2y_2, \dots)$ is Cauchy. By Exercise 3.3.4, there exists a rational number $R > 0$ such that

$$|x_n|, |y_n| \leq R \qquad \text{for all } n.$$

Fix any $\varepsilon \in \mathbb{Q}$ such that $\varepsilon > 0$. Take N sufficiently large so that

$$|x_n - x_m|, |y_n - y_m| \leq \frac{\varepsilon}{100(R+1)} \qquad \text{for all } n, m \geq N.$$

This exists because $\bar{x}$ and $\bar{y}$ are Cauchy. Then, we see that

$$\begin{aligned} |x_ny_n - x_my_m| &= |x_ny_n - x_ny_m + x_ny_m - x_my_m| \\ &\leq |x_n||y_n - y_m| + |y_m||x_n - x_m| \leq R\frac{\varepsilon}{100(R+1)} + R\frac{\varepsilon}{100(R+1)} \\ &= \frac{\varepsilon R}{50(R+1)} < \varepsilon. \end{aligned}$$

Hence, $(x_1y_1, x_2y_2, \dots)$ is Cauchy.

Exercise B.0.1. *Let $\bar{a} \in [\bar{x}]$ and $\bar{b} \in [\bar{y}]$. Show that $[\bar{a}+\bar{b}] = [\bar{x}+\bar{y}]$. The ideas in the previous paragraph will be useful here.*

Let us note that division can be defined similarly, and the proof of well-definedness is clear.

(iii) **(Ordering)** We define $[\bar{x}] < [\bar{y}]$ if and only if there exists $0 < \varepsilon \in \mathbb{Q}$ and N such that $x_n + \varepsilon < y_n$ for all $n \geq N$.

Exercise B.0.2 (Well-defined)**.** *Suppose that $[\bar{x}] < [\bar{y}]$. Let $\bar{a} \in [\bar{x}]$ and $\bar{b} \in [\bar{y}]$. Show that there exists N such that $a_n < b_n$ for all $n \geq N$.*

Note that we may define $[\bar{x}] \leq [\bar{y}]$ to mean that $[\bar{x}] < [\bar{y}]$ or $[\bar{x}] = [\bar{y}]$. We use this notation in what follows.

(iv) **(Well-ordering)** Either $[\bar{x}] < [\bar{y}]$, $[\bar{x}] > [\bar{y}]$ or $[\bar{x}] = [\bar{y}]$. Let us prove this. Suppose that $[\bar{x}] \neq [\bar{y}]$. It follows that the limit of $x_n - y_n$ is not zero (it also may not exist).

This implies that there is $\varepsilon > 0$ and a subsequence $n_1, n_2, n_3, \ldots$ such that $|x_{n_i} - y_{n_i}| > \varepsilon$ for all i. Thus, either $x_{n_i} - y_{n_i} > \varepsilon$ for infinitely many i or $x_{n_i} - y_{n_i} < \varepsilon$ for infinitely many i. Without loss of generality, we assume the former. Additionally, up to relabeling (that is, passing to a further subsequence), we may assume that

$$x_{n_i} - y_{n_i} > \varepsilon \qquad \text{for all } i.$$

Since $\bar{x}$ and $\bar{y}$ are Cauchy, we take N sufficiently large that

$$|x_n - x_m|, |y_n - y_m| < \frac{\varepsilon}{3} \qquad \text{for all } n, m \geq N.$$

Fix $i, k \geq N$. Notice that $n_i \geq i \geq N$, by definition of a subsequence. We, thus, have that

$$\begin{aligned} x_k &= (x_k - x_{n_i}) + (x_{n_i} - y_{n_i}) + (y_{n_i} - y_k) + y_k \\ &> -\frac{\varepsilon}{3} + \varepsilon - \frac{\varepsilon}{3} + y_k > y_k, \end{aligned}$$

which implies that $[\bar{x}] > [\bar{y}]$. This concludes the proof.

(v) **(Least upper bound)** We claim that any nonempty set $S \subset \mathbb{R}$ that is bounded above, has a supremum. Here, we say that S is bounded above if there is $[\bar{x}] \in \mathbb{R}$ such that $[\bar{S}] < [\bar{x}]$ for every $[\bar{s}] \in S$. Typically, this is denoted

$$\sup S := [\bar{u}].$$

Let us also note that once the supremum has been established, the infimum follows: given a set $\tilde{S}$ that is bounded from below, we let

$$\inf \tilde{S} := -\sup -\tilde{S} = -\sup\{s : -s \in \tilde{S}\}.$$

This is well-defined because $-\tilde{S}$ will be bounded from above and, thus, have a least upper bound (supremum).

The proof is somewhat long and complicated, so we only sketch the main points. Let

$$\mathcal{U}_S = \{[\bar{x}] \in \mathbb{R} : [\bar{s}] \leq [\bar{x}] \text{ for all } [\bar{s}] \in S\}.$$

This is the set of upper bounds, and it is nonempty by assumption.

First, we show that, for any $0 < \varepsilon \in \mathbb{Q}$, there is $[\bar{u}] \in \mathcal{U}_S$ and $[\bar{s}] \in S$ such that

$$[\bar{u}] - \varepsilon \leq [\bar{s}] \leq [\bar{u}]. \tag{B.1}$$

Fix any element $[\bar{\mu}] \in \mathcal{U}_S$, which exists by assumption. Let

$$n = \max\{m \in \mathbb{N} \cup \{0\} : [\bar{\mu}] - m\varepsilon \in \mathcal{U}_S\}.$$

To show that n exists, we need to show that the set on the right hand side is nonempty. Fix any $[\bar{s}] \in S$. Since $\bar{s}$ and $\bar{\mu}$ are Cauchy, there is $N > 0$ such that if $n \geq N$, we have

$$-N\varepsilon \leq s_n \qquad \text{and} \qquad \mu_n \leq N\varepsilon.$$

It follows that $[\bar{\mu}] - (2N+1)\varepsilon \leq [\bar{s}]$. Hence n is well-defined.

Let $[\bar{u}] = [\bar{\mu}] - n\varepsilon$. Since $[\bar{\mu}] - (n+1)\varepsilon \notin \mathcal{U}_S$, there must be $[\bar{s}]$ such that

$$[\bar{u}] - \varepsilon = [\bar{\mu}] - (n+1)\varepsilon < [\bar{s}].$$

On the other hand, $[\bar{u}] \in \mathcal{U}_S$ by construction. Hence, $[\bar{u}] \geq [\bar{s}]$. This completes the proof of (B.1).

Next, we use (B.1) to take a sequence that approximates the supremum. Indeed, for any n, we find sequences $[\bar{s}_1], [\bar{s}_2], \cdots \in S$ and $[\bar{u}_1], [\bar{u}_2], \cdots \in \mathcal{U}_S$ such that

$$[\bar{u}_n] - \frac{1}{2^n} \leq [\bar{s}_n] \leq [\bar{u}_n] \tag{B.2}$$

and, for all n, $[\bar{u}_{n+1}] \leq [\bar{u}_n]$. It follows from (B.2) that, for any k and n,

$$[\bar{u}_n] \leq [\bar{u}_k] + \frac{1}{2^n}. \tag{B.3}$$

Our goal is to create a "best" lower bound by taking a suitable "diagonal" sequence.

Up to removing finitely many elements of the sequence, we may assume that, for all $k, \ell \geq n$,

$$|u_{n,k} - u_{n,\ell}| \leq \frac{1}{2^n}, \qquad \text{and} \qquad |s_{n,k} - s_{n,\ell}| \leq \frac{1}{2^n} \tag{B.4}$$

where $u_{n,k}$ is the k^{th} element of the $\bar{u}_n$ (similarly for $s_{n,k}$). It is easy to check that

$$u_{n,n} - \frac{3}{2^n} \leq s_{n,n} \leq u_{n,n} + \frac{3}{2^n}.$$

We define a sequence

$$\bar{u} = \Big(u_{1,1}, u_{2,2}, \ldots, u_{n,n}, \ldots\Big).$$

We expect that $[\bar{u}]$ is our least upper bound. To show this, we must check that $[\bar{u}] \in \mathcal{U}_S$ and that $[\bar{u}] \leq [\bar{u}']$ for all $[\bar{u}'] \in \mathcal{U}_S$.

We check that $[\bar{u}] \in \mathcal{U}_S$. If not, then there is $[\bar{s}] \in S$ such that

$$[\bar{s}] > [\bar{u}].$$

By definition, there is N_1 and $\varepsilon > 0$ such that $s_n > u_n + \varepsilon$ for all $n \geq N_1$. We deduce that

$$s_n > \varepsilon + u_{n,n}.$$

Up to increasing N_1 so that $2^{-n} < \varepsilon/2$, we have that

$$s_n > \frac{\varepsilon}{2} + u_{n,n} + \frac{1}{2^n}.$$

There is N_2 such that, for all $i, j \geq N_2$.

$$|s_i - s_j| < \frac{\varepsilon}{4}.$$

Let $N = \max\{N_1, N_2, n\}$. For any $k \geq N$, it follows that

$$s_k > s_n - \frac{\varepsilon}{4} > \frac{\varepsilon}{4} + u_{n,n} + \frac{1}{2^n}.$$

From this and (B.4), it follows that, for all $k \geq n$,

$$s_k > \frac{\varepsilon}{4} + u_{n,k}.$$

We deduce that $[\bar{s}] > [\bar{u}_n]$, which is a contradiction. It follows that $[\bar{u}] \in \mathcal{U}_S$.

We now fix any $[\bar{u}'] \in \mathcal{U}_S$. Assume, by contradiction, that $[\bar{u}'] < [\bar{u}]$. By definition, there is N_1 and $\varepsilon > 0$ such that $u_n > u'_n + \varepsilon$ for all $n \geq N_1$. We deduce that

$$u'_n + \varepsilon < u_n = u_{n,n} < s_{n,n} + \frac{3}{2^n}.$$

Here we used (B.4) and (B.3) to get the last inequality above. Next, using (B.4), we find, for all $k \geq n$,

$$u'_n + \varepsilon < s_{n,k} + \frac{5}{2^n}.$$

We take N_2 such that, for $n \geq N_2$,

$$\frac{5}{2^n} < \frac{\varepsilon}{2}$$

Finally, we take N_3 such that, for all $i, j \geq N_3$,

$$|u'_i - u'_j| < \frac{\varepsilon}{4}.$$

Fix $n = \max\{N_1, N_2, N_3\}$. If $k \geq n$, we have

$$u'_k + \frac{\varepsilon}{4} < u'_n + \frac{\varepsilon}{2} < s_{n,k} + \frac{5}{2^n} - \frac{\varepsilon}{2} < s_{n,k}.$$

It follows that $[\bar{u}'] < [\bar{s}_n]$, which contradicts the fact that $[\bar{u}'] \in \mathcal{U}_S$. This completes the proof that $[\bar{u}]$ is the least upper bound.

(vi) **(Completeness)** All of the hard work for this happened in the last step (least upper bound property). Indeed, suppose that $[\bar{u}_n] \in \mathbb{R}$ is a Cauchy sequence. By standard arguments from undergraduate analysis[62], every sequence has a monotone subsequence $[\bar{u}_{n_1}], [\bar{u}_{n_2}], \ldots$.

If this subsequence is increasing, we let

$$[\bar{u}] = \sup\{[\bar{u}_{n_k}] : k = 1, 2, \ldots\}.$$

Here the sup is the least upper bound of the set.

Exercise B.0.3. *Show that* $[\bar{u}_{n_k}] \to [\bar{u}]$ *as* $k \to \infty$. *The use Exercise 3.3.1 to conclude that* $[\bar{u}_n] \to [\bar{u}]$ *as* $n \to \infty$.

If this sequence is decreasing, we let

$$[\bar{u}] = -\sup\{-[\bar{u}_{n_k}] : k = 1, 2, \ldots\}.$$

[62]Perhaps this will be added one day...

C. Identifying a Stieltjes measure

Suppose that $F : \mathbb{R} \to \mathbb{R}$ is a non-decreasing, càdlàg function associated to a measure space $(\mathbb{R}, \mathcal{F}, \mu_F)$. If $(\mathbb{R}, \mathcal{G}, \mu)$ is another measure space, how can we tell it is the same as $(\mathbb{R}, \mathcal{F}, \mu_F)$? We note that this comes up often in Section 5.1.2; see, e.g., Example 5.1.3. The interested reader may wish to explore further topics such as Carathéodory's extension theorem and Dynkin's π-λ theorem.

The first thing that we need notice is the following:

Lemma C.1. *The measure space $(\mathbb{R}, \mathcal{F}, \mu_F)$ defined via (5.1.3)-(5.1.4) is complete.*

We omit the proof as it is exactly the same as for the Lebesgue measure; see Example 4.2.7.((i)).

Proposition C.2. *Suppose that $(\mathbb{R}, \mathcal{G}, \mu)$ is a complete measure space, $\mathcal{B} \subset \mathcal{G}$, and*

$$\mu_F((a,b]) = \mu((a,b])$$

for every $a < b$. Then $\mathcal{F} \subset \mathcal{G}$ and $\mu_F(A) = \mu(A)$ for all $A \in \mathcal{F}$.

Proof. First we show that $\mathcal{F} \subset \mathcal{G}$. Let us argue by contradiction, assuming that there exists $A \in \mathcal{F} \setminus \mathcal{G}$. Since

$$A = \bigcup_{n=1}^{\infty} A \cap [-n, n],$$

it must be that

$$A_N := A \cap [-N, N] \in \mathcal{F} \setminus \mathcal{G} \qquad \text{for some } N.$$

Let $C_N = [-N, N] \cap A_N^c$, and it follows that

$$C_N \in \mathcal{F} \setminus \mathcal{G}$$

as well.

By definition, for all $n \in \mathbb{N}$, there are

$$A_N \subset I_n = \bigcup_{i=1}^{\infty} (a_{n,i}, b_{n,i}] \qquad \text{and} \qquad C_N \subset J_n = \bigcup_{i=1}^{\infty} (c_{n,i}, d_{n,i}]$$

so that

$$\mu_F(A_N) \geq \sum_{i=1}^{\infty} (F(b_{n,i}) - F(a_{n,i})) - \frac{1}{n} \quad \text{and} \quad \mu_F(C_N) \geq \sum_{i=1}^{\infty} (F(d_{n,i}) - F(c_{n,i})) - \frac{1}{n}.$$

Let us point out that, by assumption,

$$\mu_F(I_n) = \mu(I_n) \quad \text{and} \quad \mu_F(J_n) = \mu(J_n) \quad \text{for all } n \in \mathbb{N}. \tag{C.1}$$

Clearly, we can choose these intervals such that, for all i and n,

$$-N - \frac{1}{n} \leq a_{n,i}, c_{n,i} \quad \text{and} \quad b_{n,i}, d_{n,i} \leq N. \tag{C.2}$$

Then we take the set

$$I := \bigcap_{n=1}^{\infty} I_n \quad \text{and} \quad J := \bigcap_{n=1}^{\infty} J_n.$$

By (C.1), Lemma 4.5.2, and Exercise 4.5.1, we have

$$\mu_F([-N,N]) = \mu([-N,N]), \quad \mu_F(I) = \mu(I) \quad \text{and} \quad \mu_F(J) = \mu(J).$$

Clearly $I, J \in \mathcal{B} \subset \mathcal{G}$ and, by (C.2),

$$I, J \subset [-N,N]. \tag{C.3}$$

It follows that $I \setminus A_N, J \setminus C_N \in \mathcal{F} \setminus \mathcal{G}$. Additionally, we observe that

$$A_N \subset I \quad \text{and} \quad C_N \subset J$$

and that

$$\mu_F(I) = \mu_F(A_N) \quad \text{and} \quad \mu_F(J) = \mu_F(C_N). \tag{C.4}$$

We claim that

$$\mu(I \cap J) = 0 \tag{C.5}$$

and

$$I \setminus A_N \subset I \cap J. \tag{C.6}$$

This is a contradiction to completeness because then $I \setminus A_N$ is a non-$\mathcal{G}$-measurable subset of the set $I \cap J$ of measure zero.

We begin with (C.6) as it is more straightforward. Indeed,

$$I \setminus A_N = I \cap A_N^c = I \cap ([-N,N]^c \cup C_N) = I \cap C_N \subset I \cap J.$$

In the third equality, we used (C.3). For later, it will be convenient to notice that, similarly,

$$J \cap C_N^c = J \setminus C_N = J \cap A_N. \tag{C.7}$$

This is precisely (C.6).

We now show (C.5). Notice that $I \cap J \in \mathcal{B}$, so $\mu_F(I \cap J) = \mu(I \cap J)$ by assumption. Hence, we actually show that

$$\mu_F(I \cap J) = 0. \tag{C.8}$$

Next, we notice that, by the $\mathcal{F}$-measurability of $J \cap A_N$,

$$\mu_F(J \cap A_N) = \mu_F(J) - \mu_F(J \cap A_N^c) = \mu_F(J) - \mu_F(J \cap C_N) = \mu_F(J) - \mu_F(C_N) = 0.$$

We used (C.7) in the second equality and (C.4) in the last equality. Similarly, we have

$$\mu_F(I \cap C_N) = 0.$$

Thus,

$$\mu_F(I \cap J) = \mu_F(I \cap J \cap A_N) + \mu_F(I \cap J \cap A_N^c) \leq \mu_F(J \cap A_N) + \mu_F(I \cap C_N) = 0.$$

Thus, (C.8) is established and completes the proof that $\mathcal{F} \subset \mathcal{G}$.

Now we show that $\mu_F(A) = \mu(A)$ for all $A \in \mathcal{F}$. Fix any $A \in \mathcal{F}$. By Lemma 4.5.2, if

$$\mu_F(A_N)\mu(A_N), \qquad \text{where } A_N = A \cap [-N,N],$$

for every $N \in \mathbb{N}$, then $\mu_F(A_N) = \mu(A_N)$. Hence, we may assume that $A \subset [-N, N]$ for some N. Arguing as in the first part of this proof, we may find Borel sets I, J such that

$$\begin{aligned} &A \subset I \subset [-N, N], \quad C := A^c \cap [-N, N] \subset J \subset [-N, N], \\ &\mu_F(A) = \mu_F(I) = \mu(I), \quad \text{and} \quad \mu_F(C) = \mu_F(J) = \mu(J). \end{aligned} \tag{C.9}$$

By measurability, we have

$$\mu(I) = \mu(A) + \mu(I \setminus A).$$

By (C.1) and then (C.9), we have

$$\mu(I) = \mu_F(I) = \mu_F(A).$$

Thus, the proof is finished if we show that

$$\mu(I \setminus A) = 0.$$

Notice that, by (C.9), $I \setminus A = I \cap C$. Hence,

$$\begin{aligned} \mu(I \setminus A) &= \mu(I \cap C) \leq \mu(I \cap J) = \mu(I) + \mu(J) - \mu(I \cup J) \\ &= \mu_F(I) + \mu_F(J) - \mu_F(I \cup J) = \mu_F(A) + \mu_F(C) - \mu_F([-N, N]) \\ &= \mu_F(A \cup C) - \mu_F([-N, N]) = \mu_F([-N, N]) - \mu_F([-N, N]) = 0. \end{aligned}$$

All steps above either use the additivity of measures on disjoint unions, (C.1), or (C.9). Thus, the proof is complete. □

D. Proof of the law of the unconscious statistician

Proof of (5.2.5). Let us first reduce this question to a simpler one. First, it is enough to prove this only for nonnegative functions g; the general case can be handled by writing $g = g_+ - g_-$. Additionally, it is enough to show it simply only for indicator functions $g = \mathbb{1}_A$ for $A \in \mathcal{B}$. Indeed, the result then follows for all simple functions by linearity, and then the case of a general Borel function can be handled by approximation (see, e.g., the approximation procedure in the proof of Lemma 4.4.17).

We now proceed with the proof that, for all $A \in \mathcal{B}$,

$$\mathbb{E}[\mathbb{1}_A(X)] = \int \mathbb{1}_A(x)\, d\mu_X. \tag{D.1}$$

Define the Borel measure

$$\mu(A) = \mathbb{E}[\mathbb{1}_A(X)] = \mathbb{P}(X \in A).$$

Then (D.1) follows by showing that

$$\mu(A) = \mu_{F_X}(A), \tag{D.2}$$

for all $A \in \mathcal{B}$.

Actually, we claim that we need only show that

$$\mu(A) \leq \mu_F(A) \qquad \text{for all } A \in \mathcal{B}. \tag{D.3}$$

Indeed, then we would have, by applying (D.3) to A^c,

$$1 - \mu(A) = \mu(A^c) \leq \mu_F(A^c) = 1 - \mu(A),$$

which clearly yields that $\mu(A) \geq \mu_F(A)$. As a consequence, we deduce (D.2) and, thus, (D.1).

We now prove (D.3). First note that it holds for sets $A = (a, b]$. Now fix any $A \in \mathcal{B} \subset \mathcal{F}_X$. Then, by definition of outer measure, there are sets V_n of the form

$$V_n = \bigcup_{i=1}^{\infty} (a_{n,i}, b_{n,i}]$$

with $(a_{n,i}, b_{n,i}] \cap (a_{n,j}, b_{n,j}] = \emptyset$ if $i \neq j$, such that

$$A \subset V_n \qquad \text{and} \qquad \mu_{F_X}(V_n) \geq \mu_{F_X}(A) + \frac{1}{n}.$$

Up to taking finite intersections, we may assume that $V_1 \supset V_2 \supset \cdots$. Notice that

$$\mu(V_n) = \sum_{i=1}^{\infty} \mu((a_{n,i}, b_{n,i}]) = \sum_{i=1}^{\infty} \mu_{F_X}((a_{n,i}, b_{n,i}]) = \mu_{F_X}(V_n). \tag{D.4}$$

Let

$$V = \bigcap_{n=1}^{\infty} V_n.$$

Since $A \subset V$ and μ is a finite measure, we deduce from Exercise 4.5.1, that

$$\mu(A) \leq \mu(V) = \lim_{n \to \infty} \mu(V_n) = \lim_{n \to \infty} \mu_{F_X}(V_n) \leq \mu_{F_X}(A).$$

We used (D.4) in the second equality above. This is precisely (D.3). Hence, we deduce that (D.2) holds as a result. □