Applications

- Building blocks of ChatGPT
  - Attention mechanism
  - Cosine Similarity in embedding space (applications to Natural Language Processing)

- Clustering and k-means
  - Topic discovery

Topics

- Inner products (ALA 3.1)
  - Definition; length
  - Inner product spaces
- Angles (ALA 3.2)
  - Cauchy-Schwarz inequality
  - Cosine Similarity / angle
  - Triangle inequality
- Norms (ALA 3.3)
  - Definition and examples

Inner products

In the last two lectures, we generalized how we add and scale vectors in $\mathbb{R}^2$ and $\mathbb{R}^3$ to general Euclidean spaces $\mathbb{R}^n$, and more general vector spaces V. Today, we bring over other key concepts from $\mathbb{R}^2$ & $\mathbb{R}^3$ to vector spaces, namely the ideas of angle, length, and distance.

The notions of angle, length, and distance in general vector spaces play a foundational role in modern applications of engineering, economics, and AI. By the end of the next few lectures, you will be equipped with both conceptual and computational tools that will allow you to solve some really interesting problems with immediate real-world application!

We start with a familiar example of an inner-product for vectors in $\mathbb{R}^n$, the dot product. For two vectors $\underline{v}, \underline{w} \in \mathbb{R}^n$, their dot product $\underline{v} \cdot \underline{w}$ is defined as:

$$\underline{v} \cdot \underline{w} = v_1 w_1 + v_2 w_2 + \cdots + v_n w_n \qquad \text{(DP)}$$

Note that $\underline{v} \cdot \underline{w} = \underline{v}^T \underline{w} = \underline{w}^T \underline{v}$ can be written as a row-vector column-vector product.

A key property of the dot product is that

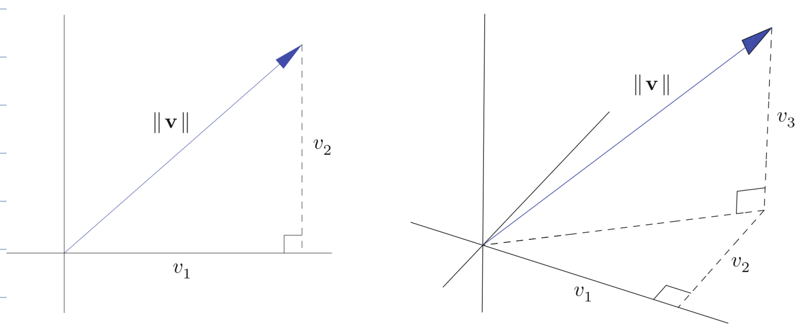
$$\underline{v} \cdot \underline{v} = \underline{v}^T \underline{v} = v_1^2 + v_2^2 + \cdots + v_n^2,$$

i.e., that the dot product of a vector $\underline{v}$ with itself is given by the sum of the square of its entries.

The Pythagorean theorem extends to n-dimensional space, and tells us that $\underline{v} \cdot \underline{v}$ is equal to the square of its length. We use this observation to define the Euclidean norm (or length) $\|\underline{v}\|$ of a vector $\underline{v}$ to be:

$$\|\underline{v}\| = \sqrt{\underline{v} \cdot \underline{v}} = \sqrt{v_1^2 + v_2^2 + \cdots + v_n^2}.$$

This generalizes our idea of length from $\mathbb{R}^2$ and $\mathbb{R}^3$ to $\mathbb{R}^n$:



The Euclidean norm $\|\underline{v}\|$ of a vector $\|\underline{v}\|$ has some intuitive properties. For example, if $\underline{v} \neq 0$, then $\|\underline{v}\| > 0$ (all nonzero vectors have positive length), and $\|\underline{v}\| = 0$ if and only if $\underline{v} = 0$ (only the zero vector has zero length).

These properties, and those of the dot-product, inspire the following abstract definition for more general inner-products:

Definition 3.1. An *inner product* on the real vector space V is a pairing that takes two vectors $\mathbf{v}, \mathbf{w} \in V$ and produces a real number $\langle \mathbf{v}, \mathbf{w} \rangle \in \mathbb{R}$. The inner product is required to satisfy the following three axioms for all $\mathbf{u}, \mathbf{v}, \mathbf{w} \in V$, and scalars $c, d \in \mathbb{R}$.

(*i*) *Bilinearity*:
$$\begin{aligned} \langle c\,\mathbf{u} + d\,\mathbf{v}, \mathbf{w} \rangle &= c \langle \mathbf{u}, \mathbf{w} \rangle + d \langle \mathbf{v}, \mathbf{w} \rangle, \\ \langle \mathbf{u}, c\,\mathbf{v} + d\,\mathbf{w} \rangle &= c \langle \mathbf{u}, \mathbf{v} \rangle + d \langle \mathbf{u}, \mathbf{w} \rangle. \end{aligned} \tag{3.4}$$

(*ii*) *Symmetry*:
$$\langle \mathbf{v}, \mathbf{w} \rangle = \langle \mathbf{w}, \mathbf{v} \rangle. \tag{3.5}$$

(*iii*) *Positivity*:
$$\langle \mathbf{v}, \mathbf{v} \rangle > 0 \quad \text{whenever} \quad \mathbf{v} \neq \mathbf{0}, \quad \text{while} \quad \langle \mathbf{0}, \mathbf{0} \rangle = 0. \tag{3.6}$$

As we will see soon, an inner product allows us to define notions of angle, length, and distance in a vector space. This added structure is very useful, so when a vector space is equipped with an inner product, we call it an inner product space.

WARNING: A vector space can admit many different inner products. It is therefore necessary (and polite) to specify which inner product is being used when defining an inner product space.

You can (and should!) check that the dot product satisfies Def'n 3.1.

Now, just as the dot product could be used to define the Euclidean norm of a vector, we can also define a norm induced by a general inner-product:

$$\|\underline{v}\| = \sqrt{\langle \underline{v}, \underline{v} \rangle}.$$

The positivity axiom ensures that $\|\underline{v}\| \geq 0$ for all $\underline{v} \in V$, and that $\|\underline{v}\| = 0$ if and only if $\underline{v} = 0$.

WARNING: We are using the same norm symbol $\|\cdot\|$ for many different norms. If we do not specify which norm/inner-product is being used, you should interpret this as a "generic" norm induced by a "generic" inner product satisfying Def'n 3.1. The good news is that both behave in ways similar to the familiar dot product and Euclidean norm.

Example: Let's define a different inner product on $\mathbb{R}^2$. Instead of the typical dot product, let's consider the weighted dot product:

$$\langle \underline{v}, \underline{w} \rangle = 2v_1w_1 + 5v_2w_2$$

for $\underline{v} = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$, $\underline{w} = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$.

In the online notes (and Example 3.2), we show you that it is not too hard to verify that the weighted dot product satisfies Def'n 3.1. The weighted norm it induces on $\mathbb{R}^2$ is then

$$\|\underline{v}\| = \sqrt{\langle \underline{v}, \underline{v} \rangle} = \sqrt{2v_1^2 + 5v_2^2}$$

which assigns more weight to the second coordinate relative to the first.

We can generalize this example to $\mathbb{R}^n$ and arbitrary positive weights. Let $c_1, \ldots, c_n > 0$ be positive numbers. Then the corresponding weighted inner product and weighted norm on $\mathbb{R}^n$ are defined to be

$$\langle \underline{v}, \underline{w} \rangle = c_1v_1w_1 + c_2v_2w_2 + \cdots + c_nv_nw_n$$

$$\|\underline{v}\| = \sqrt{\langle \underline{v}, \underline{v} \rangle} = \sqrt{c_1v_1^2 + c_2v_2^2 + \cdots + c_nv_n^2}$$

The numbers $c_i > 0$ are called the weights. Weighted norms play a very important role in statistics and data fitting, where one picks large/small c_i to emphasize/de-emphasize the importance of certain measurements. We'll revisit weighted norms in the context of least squares data fitting in a few classes.

Example: [OPTIONAL] We saw that we can define vector spaces where vectors are actually infinite sequences or even functions! We will not work much with such function spaces in the rest of the class, but you should know that we can define inner products on these vector spaces too.

For example, recall we saw that $C^0[0,1]$, the space of all continuous functions defined over the interval $[0,1]$, is a vector space. A commonly used inner product on this space is

$$\langle f, g \rangle = \int_0^1 f(t)g(t)\,dt \qquad (\text{Int})$$

for any two functions $f, g \in C^0[0,1]$. One can verify that this inner product satisfies Def'n 3.1. We won't do that, but to get

Some intuition, let's consider the sampled function versions $\underline{f}, \underline{g} \in \mathbb{R}^{T+1}$, where remember, we define

$$\underline{f} = \begin{bmatrix} f(0) \\ f(\tau) \\ \vdots \\ f(T\tau) \end{bmatrix} = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_{T+1} \end{bmatrix}, \quad \underline{g} = \begin{bmatrix} g(0) \\ g(\tau) \\ \vdots \\ g(T\tau) \end{bmatrix} = \begin{bmatrix} g_1 \\ g_2 \\ \vdots \\ g_{T+1} \end{bmatrix}$$

for τ a sampling time chosen so that $T\tau = 1$. These two vectors live in $\mathbb{R}^{T+1}$, and let's consider a weighted inner product with all weights $c_i = \tau$. Then the inner product between $\underline{f}$ and $\underline{g}$ is

$$\langle \underline{f}, \underline{g} \rangle = \tau f_0 g_0 + \tau f_1 g_1 + \cdots + \tau f_{T+1} g_{T+1}$$
$$= \sum_{i=0}^{T} \tau f(i\tau) g(i\tau)$$

This should remind you of how the Riemann Integral for the function $h(t) = f(t)g(t)$ is defined. Indeed, if we let $\tau \to 0$, we recover the integral inner product defined above. Since our inner product $\langle \underline{f}, \underline{g} \rangle$ satisfies Defn 3.1 for any τ, it shouldn't be too surprising that the integral inner product (Int) does too.

(Int) then defines the L_2 norm of a function:

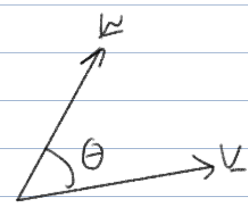
$$\|f\| = \sqrt{\int_0^1 f^2(t)\,dt}$$

which generalizes the notion of length to functions. These ideas might seem very abstract, but they are immensely practical, and lie at the heart of modern applications of Fourier analysis, differential equations, and numerical analysis.

Angles and the Cauchy-Schwarz Inequality

Our starting point in defining the notion of angle in a general inner product space is the familiar formula

$$\underline{v} \cdot \underline{w} = \|\underline{v}\| \cdot \|\underline{w}\| \cos\theta$$



where θ measures the angle between $\underline{v}$ and $\underline{w}$. Since $|\cos\theta| \le 1$, we can bound the magnitude of $\underline{v} \cdot \underline{w}$:

$$|\underline{v} \cdot \underline{w}| \le \|\underline{v}\| \cdot \|\underline{w}\|.$$

This is the simplest form of the general Cauchy - Schwarz inequality, which holds for any inner product. That is, it is always true that

$$|\langle \underline{v}, \underline{w} \rangle| \leq \|\underline{v}\| \cdot \|\underline{w}\| \text{ for all } \underline{v}, \underline{w} \in V \qquad (CS)$$

Here, $\|\underline{v}\| = \sqrt{\langle \underline{v}, \underline{v} \rangle}$ is the norm induced by the inner product, and $|\cdot|$ denotes the absolute value of a real number.

Note that equality holds in (CS) if and only if $\underline{v}$ and $\underline{w}$ are parallel vectors.

This inequality lets us define the following generalized "angle" between any two vectors $\underline{v}$ and $\underline{w}$ in an inner product space:

$$\cos\theta = \frac{\langle \underline{v}, \underline{w} \rangle}{\|\underline{v}\| \cdot \|\underline{w}\|}. \qquad (\text{angle})$$

This definition makes sense because, by (CS), we know that

$$-1 \leq \frac{\langle \underline{v}, \underline{w} \rangle}{\|\underline{v}\| \cdot \|\underline{w}\|} \leq 1$$

and so θ is well defined, and unique if restricted to lie in $[0, \pi]$.

For example, the vectors $\underline{v} = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}$ and $\underline{w} = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}$ have dot product $\underline{v} \cdot \underline{w} = 1$, and norms $\|\underline{v}\| = \|\underline{w}\| = \sqrt{2}$, and hence

$$\cos\theta = \frac{1}{\sqrt{2} \cdot \sqrt{2}} = \frac{1}{2} \Rightarrow \theta = \arccos\left(\tfrac{1}{2}\right) = \frac{\pi}{3}$$

which is the usual notion of angle. But, we can also compute the "angle" between $\underline{v}$ and $\underline{w}$ with respect to the weighted inner product $\langle \underline{v}, \underline{w} \rangle = v_1 w_1 + 2 v_2 w_2$. In this inner product, $\langle \underline{v}, \underline{w} \rangle = 3$, $\|\underline{v}\| = 2$, and $\|\underline{w}\| = \sqrt{5}$, and so

$$\cos\theta = \frac{3}{2\sqrt{5}} = .67082\ldots \Rightarrow \theta = \arccos\left(\frac{3}{2\sqrt{5}}\right) = .83548\ldots$$

We can now also define angles between, for example, polynomials. For $p(x) = a_0 + a_1 x + a_2 x^2$, $q(x) = b_0 + b_1 x + b_2 x^2 \in P^{(2)}$, define the inner product $\langle p, q \rangle = a_0 b_0 + a_1 b_1 + a_2 b_2$. Note that this agrees with the standard dot product applied to $\underline{p} = \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix}$, $\underline{q} = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix}$, and hence immediately satisfies def'n 3.1. The angle between $p(x)$ and $q(x)$ is then computed as

$$\cos\theta = \frac{\langle p,q \rangle}{\|p\| \cdot \|q\|} = \frac{\underline{p} \cdot \underline{q}}{\|\underline{p}\| \cdot \|\underline{q}\|}.$$

For example if $p(x) = 1 + x^2$ and $q(x) = x + x^2$, then $\langle p,q \rangle = 1$ and $\|p\| = \|q\| = \sqrt{2}$, and $\cos\theta = \frac{1}{2} \Rightarrow \theta = \frac{\pi}{3}$.

Note that the expression (angle) is called the cosine similarity of two vectors, and measures how "aligned" they are. We will see in this lecture's case study that this plays an important role in modern chatbots like ChatGPT!

Orthogonal Vectors

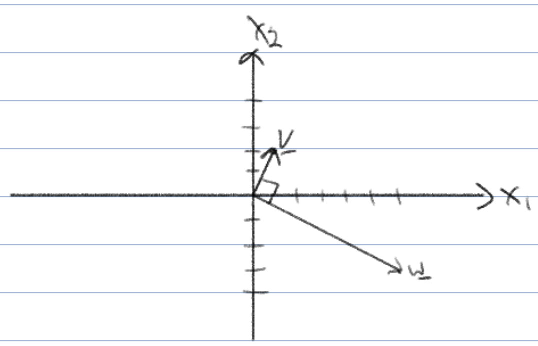
The notion of perpendicular vectors is an important one in Euclidean geometry. These are vectors that meet at a right angle, i.e., $\theta = \frac{\pi}{2}$ or $\theta = -\frac{\pi}{2}$, with $\cos\theta = 0$. This tells us vectors $\underline{v}$ and $\underline{w}$ if and only if their dot product vanishes: $\underline{v} \cdot \underline{w} = 0$ (can you see why via Cauchy-Schwarz?).

We continue with our strategy of extending familiar geometric concepts in Euclidean spaces to general inner product spaces. For historic reasons, we use the term orthogonal instead of perpendicular.

Two elements $\underline{v}, \underline{w} \in V$ of an inner product space are orthogonal (with respect to $\langle \cdot, \cdot \rangle$) if $\langle \underline{v}, \underline{w} \rangle = 0$.

Orthogonality is an incredibly useful and practical idea that appears all over the place in engineering, AI, and economics, which we will explore in detail next lecture.

Example: The vectors $\underline{v} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$ and $\underline{w} = \begin{bmatrix} 6 \\ -3 \end{bmatrix}$ are orthogonal with respect to the dot product: $\underline{v} \cdot \underline{w} = 1 \cdot 6 + 2 \cdot (-3) = 0$. Indeed if we draw them, we see they meet at a right angle:



HOWEVER, $\underline{v}$ and $\underline{w}$ are NOT ORTHOGONAL with respect to the weighted inner product $\langle \underline{v}, \underline{w} \rangle = v_1 w_1 + 2 v_2 w_2$:

$$\langle \underline{v}, \underline{w} \rangle = \left\langle \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \begin{bmatrix} 6 \\ -3 \end{bmatrix} \right\rangle = 1(1 \cdot 6) + 2(2 \cdot -3)$$
$$= 6 - 12 = -6 \neq 0$$

FACT: Orthogonality, like angles in general, depend on the inner product being used!

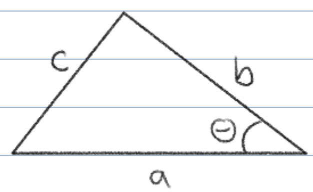
Example: The polynomials $f(x) = x$ and $g(x) = 1 + x^2$ are orthogonal with respect to the inner product on $\mathcal{P}^{(2)}$ defined previously, i.e., $\langle p, q \rangle = a_0 b_0 + a_1 b_1 + a_2 b_2$. Here $a_0 = 0$, $a_1 = 1$, $a_2 = 0$ and $b_0 = 1$, $b_1 = 0$, $b_2 = 1$, so $\langle f, g \rangle = 0 \cdot 1 + 1 \cdot 0 + 0 \cdot 1 = 0$.

However, f and g are not orthogonal with respect to the inner product $\langle p, q \rangle = \int_0^1 p(x) q(x)\, dx$ defined on $C^0[0,1]$:

$$\langle f, g \rangle = \int_0^1 x(1 + x^2)\, dx = \int_0^1 x + x^3\, dx = \frac{x^2}{2} + \frac{x^4}{4} \Big|_0^1$$
$$= \frac{1}{2} + \frac{1}{4} = \frac{3}{4} \neq 0.$$

The Triangle Inequality

We know, e.g., from the law of cosines, that the length of one side of a triangle is at most the sum of the lengths of the other two sides:



$$c^2 = a^2 + b^2 - 2ab\cos\theta$$
$$\leq a^2 + b^2 + 2ab \qquad (\text{since } |\cos\theta| \leq 1)$$
$$= (a+b)^2$$

So that $c \leq a + b$.

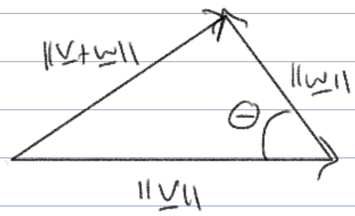
This idea extends directly to the setting where we want to relate the length $\|\underline{v} + \underline{w}\|$ of the sum of vectors $\underline{v}, \underline{w}$ to the lengths $\|\underline{v}\|$ and $\|\underline{w}\|$.

Theorem: The norm associated with an inner product satisfies the triangle inequality

$$\|\underline{v} + \underline{w}\| \leq \|\underline{v}\| + \|\underline{w}\| \quad \text{for all } \underline{v}, \underline{w} \in V.$$

Equality holds if and only if $\underline{v} = c\underline{w}$ for some positive constant $c > 0$.

Proof: This is almost exactly the same as the law of cosines! Set up a triangle as follows



and now use that $\|\underline{v}+\underline{w}\|^2 = \langle \underline{v}+\underline{w}, \underline{v}+\underline{w}\rangle$

$$= \|\underline{v}\|^2 + 2\langle \underline{v},\underline{w}\rangle + \|\underline{w}\|^2$$

Cauchy-Schwarz $(|\cos\theta| \leq 1)$

$$\left\{ \begin{aligned} &= \|\underline{v}\|^2 + 2\|\underline{v}\|\cdot\|\underline{w}\|\cos\theta + \|\underline{w}\|^2 \\ &\leq \|\underline{v}\|^2 + 2\|\underline{v}\|\cdot\|\underline{w}\| + \|\underline{w}\|^2 \end{aligned} \right.$$

$$= (\|\underline{v}\| + \|\underline{w}\|)^2.$$

Example: $\underline{v} = \begin{bmatrix} 1 \\ 2 \\ -1 \end{bmatrix}$, $\underline{w} = \begin{bmatrix} 2 \\ 0 \\ 3 \end{bmatrix}$, $\underline{v}+\underline{w} = \begin{bmatrix} 3 \\ 2 \\ 2 \end{bmatrix}$

$\|\underline{v}\| = \sqrt{6}$, $\|\underline{w}\| = \sqrt{13}$, $\|\underline{v}+\underline{w}\| = \sqrt{17}$

And triangle ineq. tells us that

$$4.123 \approx \sqrt{17} = \|\underline{v}+\underline{w}\| \leq \|\underline{v}\| + \|\underline{w}\| = \sqrt{6} + \sqrt{13} \approx 6.055$$

which is true.

Norms

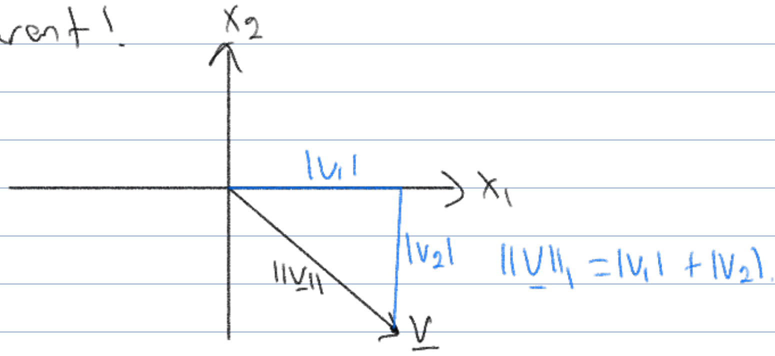
We have seen that inner products allow us to define a natural notion of length, which we called a norm. However, there are other sensible ways of measuring the size of a vector that do not arise from an inner product. For example suppose we choose to measure the size of a vector by its "taxi cab distance" where we pretend we are a cab driver in Manhattan, and we can only drive east/west and then north south. We then end up with a different measure of length that makes lots of sense!

Example Consider the vector $\underline{v} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$. Its Euclidean norm is

$\|\underline{v}\| = \sqrt{1^2 + (-1)^2} = \sqrt{2}$. Its taxi cab distance, which we will label $\|\underline{v}\|_1$ (for reasons that become clear soon), is

$$\|\underline{v}\|_1 = \underbrace{|1|}_{\substack{\text{drive 1 unit} \\ \text{east}}} + \underbrace{|-1|}_{\substack{\text{drive} \\ \text{1 unit south}}} = 2$$

These are different!



To define a general norm on a vector space, we will extract properties that "make sense" as a measure of distance but that do not directly rely on inner product structure (like angles).

Definition 3.12. A *norm* on a vector space V assigns a non-negative real number $\|\mathbf{v}\|$ to each vector $\mathbf{v} \in V$, subject to the following axioms, valid for every $\mathbf{v}, \mathbf{w} \in V$ and $c \in \mathbb{R}$:

(*i*) *Positivity*: $\|\mathbf{v}\| \geq 0$, with $\|\mathbf{v}\| = 0$ if and only if $\mathbf{v} = \mathbf{0}$.

(*ii*) *Homogeneity*: $\|c\,\mathbf{v}\| = |c|\,\|\mathbf{v}\|$.

(*iii*) *Triangle inequality*: $\|\mathbf{v} + \mathbf{w}\| \leq \|\mathbf{v}\| + \|\mathbf{w}\|$.

Axiom (i) says "length" should always be non-negative, and only the zero vector has zero length (seems reasonable!).

Axiom (ii) says if I stretch/shrink a vector $\underline{v}$ by a factor $c \in \mathbb{R}$ then the length should scale accordingly (this is why we call $c \in \mathbb{R}$ a scalar!). Note that $c < 0$ means we stretch/shrink and flip $\underline{v}$, but flipping shouldn't affect length, so $\|c\underline{v}\| = \|-c\underline{v}\| = |c|\|\underline{v}\|$.

Axiom (iii) tells us that lengths of sums of vectors should "behave as if there is a cosine rule" even if there is no notion of angle. This is a less intuitive property, but has been identified as a key property to make norms useful to work with.

We will introduce two other commonly used norms in practice, but you should know that there are many many more.

Example: The 1-norm of a vector $\underline{v} = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix} \in \mathbb{R}^n$ is the sum of the absolute values of its entries:

$$\|\underline{v}\|_1 = |v_1| + |v_2| + \cdots + |v_n|$$

which we recognize as our taxi cab distance!

The ∞-norm or max-norm is given by the maximal entry in absolute value:

$$\|\underline{v}\|_\infty = \max\{|v_1|, |v_2|, \ldots, |v_n|\}$$

Checking the axioms of Defn 3.12 is a good exercise for you. The basic inequality $|a+b| \leq |a| + |b|$ for $a, b \in \mathbb{R}$ is all you need.

The 1-norm, ∞-norm and Euclidean norm (also called the 2-norm) are examples of the general p-norm:

$$\|\underline{v}\|_p = \sqrt[p]{\sum_{i=1}^{n} |v_i|^p}, \qquad (\text{p-norm})$$

which can be shown to be a valid norm for all $1 \leq p < \infty$ (the ∞-norm is a limiting case of (p-norm) as $p \to \infty$).

The hard part in showing (p-norm) is a norm is verifying the triangle inequality, which is also known as Minkowski's inequality.