EECS 127/227AT

Optimization Models in Engineering

Course Reader

Spring 2024

Acknowledgements

This reader is based on lectures from Spring 2021, Fall 2022, Spring 2023, and Fall 2023 iterations of EECS 127/227A by Prof. Gireeja Ranade. The reader was mostly written by Spring 2023 GSIs Druv Pai, Arwa Alanqary, and Aditya Ramabadran, and reviewed by Prof. Ranade. Fall 2022 tutor Jeffrey Wu collaborated with Druv on a writeup about the Eckart-Young theorem which was folded into the reader. Contributions from Prof. Venkat Anantharam and Chih-Yuan Chiu were also added in Fall 2023.

Contents

Chapter 1

Introduction

Relevant sections of the textbooks:

- [1] Chapter 1.
- [2] Chapter 1.

1.1 What is Optimization?

Try to see what the following "problems" have in common.

- A statistical model, such as a neural network, trains using finite data samples.
- A robot learns a strategy using the environment, so that it does what you want.
- A major gas company decides what mixture of different fuels to process in order to get maximum profit.
- The EECS department decides how to set class sizes in order to maximize the number of credits offered subject to budget constraints.

While it might seem that these four examples are very distinct, they can all be formulated as minimizing an *objective function* over a *feasible set*. Thus, they can all be put into the framework of optimization.

To develop the basics of optimization, including precisely defining an objective function and a feasible set, we use some motivating examples from the third and fourth "problems". (The first and second "problems" will be discussed at the very end of the course.)

Example 1 (Oil and Gas)**.** Say that we are a gas company with 10^5 barrels of crude oil that we *must* refine by an expiration date. There are two refineries: one which processes crude oil into jet fuel, and one which processes crude oil into gasoline. We can sell a barrel of jet fuel to consumers for $\$0.10$, while we can sell a barrel of gasoline fuel for $\$0.20$. So, letting x_1 be a variable denoting the number of barrels of jet fuel produced, and x_2 be a variable denoting the number of barrels of gasoline produced, we aim to solve the problem:

$$\begin{aligned} \max_{x_1,x_2} \quad & \frac{1}{10}x_1 + \frac{1}{5}x_2 \\ \text{s.t.} \quad & x_1 \geq 0 \\ & x_2 \geq 0 \end{aligned} \tag{1.1}$$

$$x_1 + x_2 = 10^5.$$

That is, we aim to choose x_1 and x_2 which maximize the *objective function* $\frac{1}{10}x_1 + \frac{1}{5}x_2$, but with the caveat that they must obey the *constraints* $x_1 \geq 0$, $x_2 \geq 0$, and $x_1 + x_2 = 10^5$. The *feasible set* is the set of all (x_1, x_2) pairs which obey the constraints. As you may have noticed, constraints can be equalities or inequalities in the x_i, which we formalize shortly.

The solution to this problem can be seen to be $(x_1^\star, x_2^\star) = (0, 10^5)$, which corresponds to refining all the crude oil into gasoline. This makes sense – after all, gasoline sells for more! And with all else equal between gasoline and jet fuel, to maximize our profit, we just need to produce gasoline.

To model another constraint, say that we need at least 10^3 gallons of jet fuel and $5 \cdot 10^2$ gallons of gasoline, we can directly incorporate them into the constraint set:

$$\begin{aligned} \max_{x_1, x_2} \quad & \frac{1}{10}x_1 + \frac{1}{5}x_2 \\ \text{s.t.} \quad & x_1 \geq 0 \\ & x_2 \geq 0 \\ & x_1 \geq 10^3 \\ & x_2 \geq 5 \cdot 10^2 \\ & x_1 + x_2 = 10^5. \end{aligned} \tag{1.2}$$

We then notice that $x_1 \geq 0$ is made redundant by the constraint $x_1 \geq 10^3$. That is, no pair (x_1, x_2) which satisfies $x_1 \geq 10^3$ is not going to satisfy $x_1 \geq 0$. Thus, we can eliminate the latter constraint, since it defines the same feasible set. We can do the same thing for the constraints $x_2 \geq 0$ and $x_2 \geq 5 \cdot 10^2$, the latter making the former redundant. Thus, we can simplify the above problem to only include the redundant constraints:

$$\begin{aligned} \max_{x_1, x_2} \quad & \frac{1}{10}x_1 + \frac{1}{5}x_2 \\ \text{s.t.} \quad & x_1 \geq 10^3 \\ & x_2 \geq 5 \cdot 10^2 \\ & x_1 + x_2 = 10^5. \end{aligned} \tag{1.3}$$

Let's say that we want to incorporate one final business need. Before, we were modeling that the oil refinement is free, since we don't have an objective or constraint term which involves this cost. Now, let us say that we can transport a total of $2 \cdot 10^6$ "barrel-miles" – that is, the number of barrels times the number of miles we can transport is no greater than $2 \cdot 10^6$. Let us further say that the jet fuel refinery is 10 miles away from the crude oil storage, and the gasoline refinery is 30 miles away from the crude oil storage. We can incorporate this further constraint into the constraint set directly:

$$\begin{aligned} \max_{x_1, x_2} \quad & \frac{1}{10}x_1 + \frac{1}{5}x_2 \\ \text{s.t.} \quad & x_1 \geq 10^3 \\ & x_2 \geq 5 \cdot 10^2 \\ & 10x_1 + 30x_2 \leq 2 \cdot 10^6 \\ & x_1 + x_2 = 10^5. \end{aligned} \tag{1.4}$$

This is a good first problem; we have a non-trivial objective function, non-trivial inequality and equality constraints, and even got to work with manipulating constraints (so as to remove redundant ones)!

This type of optimization problem is called a *linear program*. We will learn more about how to formulate and solve linear programs later in the course.

A more generic reformulation of the above optimization problem is the following "standard form".

Definition 2 (Standard Form of Optimization Problem)
We say that an optimization problem is written in *standard form* if it is of the form

$$\begin{aligned} \min_{\vec{x}\in\mathbb{R}^n} \quad & f_0(\vec{x}) \\ \text{s.t.} \quad & f_i(\vec{x}) \leq 0, \qquad \forall i \in \{1,\dots,m\} \\ & h_j(\vec{x}) = 0, \qquad \forall j \in \{1,\dots,p\}. \end{aligned} \tag{1.5}$$

Here:

- $\vec{x} \in \mathbb{R}^n$ is the optimization variable.
- $f_1, \dots, f_m$ and $h_1, \dots, h_p$ are functions $\mathbb{R}^n \to \mathbb{R}$.
- f_0 is the objective function.
- f_i are inequality constraint functions; the expression "$f_i(\vec{x}) \leq 0$" is an inequality constraint.
- Similarly, h_j are equality constraint functions, and the expression "$h_j(\vec{x}) = 0$" is an equality constraint.
- The feasible set, i.e., the set of all $\vec{x}$ that satisfy all constraints, is

$$\Omega \doteq \left\{ \vec{x} \in \mathbb{R}^n \;\middle|\; \begin{array}{ll} f_i(\vec{x}) \leq 0, & \forall i \in \{1,\dots,m\} \\ h_j(\vec{x}) = 0, & \forall j \in \{1,\dots,p\} \end{array} \right\}. \tag{1.6}$$

  We can thus also write the problem (1.5) as

$$\min_{\vec{x}\in\Omega} f_0(\vec{x}). \tag{1.7}$$

- A *solution* to this optimization problem is any $\vec{x}^\star \in \Omega$ which attains the minimum value of $f(\vec{x})$ across all $\vec{x} \in \Omega$. Correspondingly, $\vec{x}^\star$ is also called a minimizer of f_0 over Ω.

It's perfectly fine if $m = 0$ (in which case there are no inequality constraints) and/or $p = 0$ (in which case there are no equality constraints). If there are no constraints, then $\Omega = \mathbb{R}^n$ and the problem is called *unconstrained*; otherwise it is called *constrained*.

Let us try another example now, which has vector-valued quantities.

Example 3. Consider the following table of EECS courses:

Class	Size	Credits	Resources per Student
127	x_1	c_1	r_1
126	x_2	c_2	r_2
182	x_3	c_3	r_3
189	x_4	c_4	r_4
162	x_5	c_5	r_5
188	x_6	c_6	r_6
$\vdots$	$\vdots$	$\vdots$	$\vdots$

Suppose there are n classes in total. Let $\vec{x} \doteq \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix}^\top \in \mathbb{R}^n$ be the decision variable, and let $\vec{c} \doteq \begin{bmatrix} c_1 & c_2 & \cdots & c_n \end{bmatrix}^\top \in \mathbb{R}^n$ and $\vec{r} \doteq \begin{bmatrix} r_1 & r_2 & \cdots & r_n \end{bmatrix}^\top \in \mathbb{R}^n$ be constants. Then, in order to maximize the total number of credit hours subject to a total resource budget b, we set up the linear program

$$\begin{aligned} \max_{\vec{x} \in \mathbb{R}^n} \quad & \vec{c}^\top \vec{x} \\ \text{s.t.} \quad & \vec{r}^\top \vec{x} \leq b \\ & x_i \geq 0, \qquad \forall i \in \{1, \ldots, n\}. \end{aligned} \tag{1.8}$$

As notation, instead of the last set of constraints $x_i \geq 0$, we can write the vector constraint $\vec{x} \geq \vec{0}$.

More generally, recall that if we have a vector equality constraint $\vec{h}(\vec{x}) = \vec{0}$, it can be viewed as short-hand for the several scalar equality constraints $h_1(\vec{x}) = 0, \ldots, h_p(\vec{x}) = 0$. Correspondingly, we define the vector inequality constraint $\vec{f}(\vec{x}) \leq \vec{0}$ to be short-hand for the several scalar inequality constraints $f_1(\vec{x}) \leq 0, \ldots, f_m(\vec{x}) \leq 0$.

1.2 Least Squares

We begin with one of the simplest optimization problems, that of least squares. We've probably seen this formulation before. Mathematically, we are given a data matrix $A \in \mathbb{R}^{m \times n}$ and a vector of outcomes $\vec{y} \in \mathbb{R}^m$, and attempt to find a parameter vector $\vec{x} \in \mathbb{R}^n$ which minimizes the residual $\|A\vec{x} - \vec{y}\|_2^2$. Here $\|\cdot\|_2$ is the standard Euclidean norm $\|\vec{z}\|_2 \doteq \sqrt{\vec{z}^\top \vec{z}} = \sqrt{\sum_{i=1}^n z_i^2}$; it is labeled with the 2 for a reason we will see later in the course.

More precisely, we attempt to solve the following optimization problem:

$$\min_{\vec{x} \in \mathbb{R}^n} \|A\vec{x} - \vec{y}\|_2^2 . \tag{1.9}$$

Theorem 4 (Least Squares Solution)
Let $A \in \mathbb{R}^{m \times n}$ have full column rank, and let $\vec{y} \in \mathbb{R}^m$. Then the solution to (1.9), i.e., the solution to

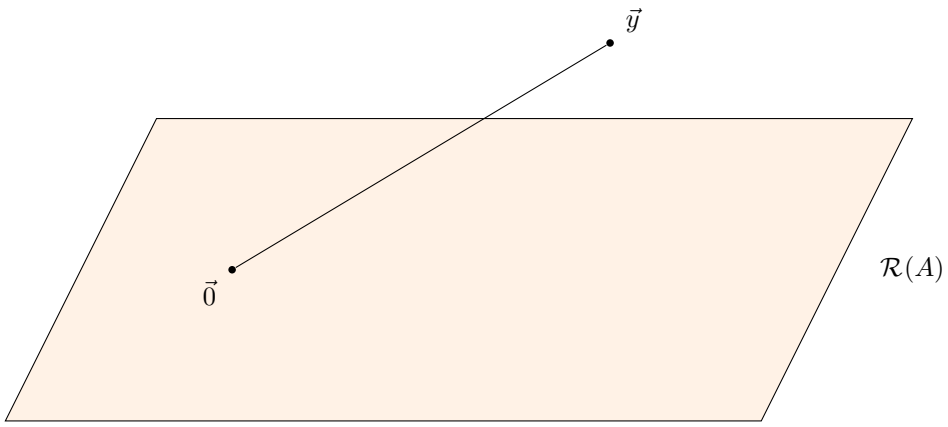
$$\min_{\vec{x} \in \mathbb{R}^n} \|A\vec{x} - \vec{y}\|_2^2 , \tag{1.9}$$

is given by

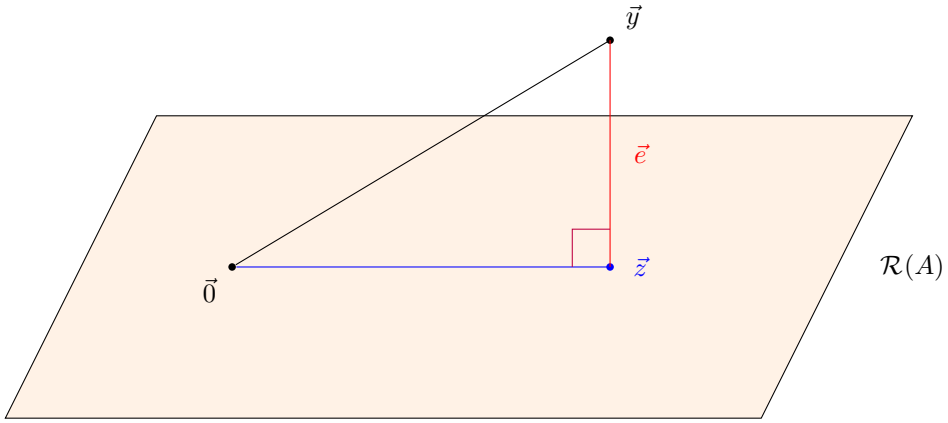
$$\vec{x}^\star = (A^\top A)^{-1} A^\top \vec{y}. \tag{1.10}$$

Proof. The idea is to find $A\vec{x} \in \mathcal{R}(A)$ which is closest to $\vec{y}$. Here $\mathcal{R}(A)$ is the range, or column space, or column span, of A. In general, We have no guarantee that $\vec{y} \in \mathcal{R}(A)$, so there is not necessarily an $\vec{x}$ such that $A\vec{x} = \vec{y}$. Instead, we are finding an approximate solution to the equation $A\vec{x} = \vec{y}$.

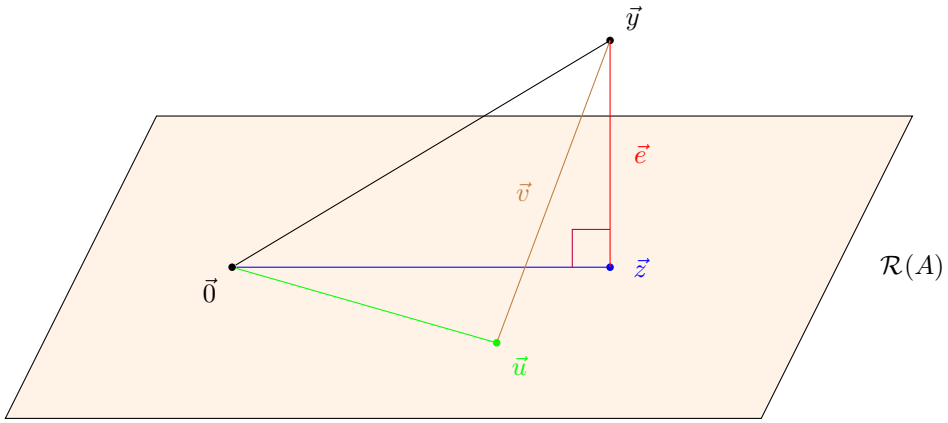
Recall that $\mathcal{R}(A)$ is a subspace, and that $\vec{y}$ itself may not belong to $\mathcal{R}(A)$. Thus we can visualize the geometry of the problem as the following picture:



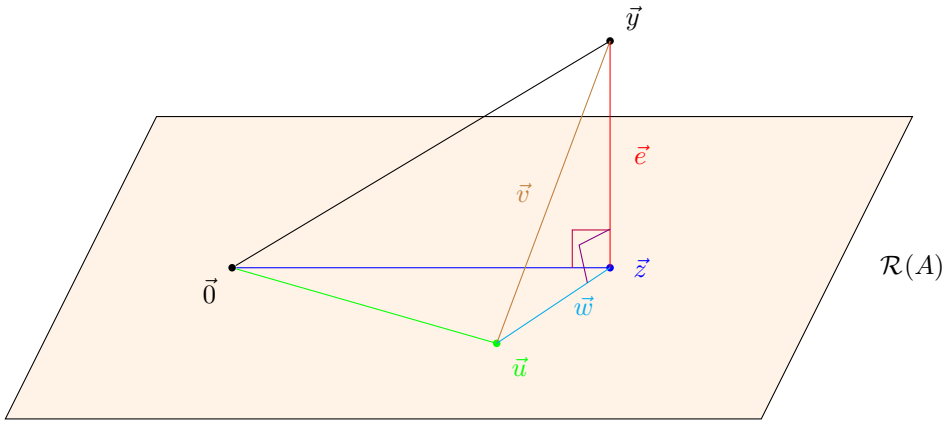
We can now solve this problem using ideas from geometry. We claim that the closest point to $\vec{y}$ contained in $\mathcal{R}(A)$ is the orthogonal projection of $\vec{y}$ onto $\mathcal{R}(A)$; call this point $\vec{z}$. Also, define $\vec{e} \doteq \vec{y} - \vec{z}$. This gives the following diagram.



From this diagram, we see that $\vec{e}$ is orthogonal to any vector in $\mathcal{R}(A)$. But remember that we still have to prove that $\vec{z}$ is the closest point to $\vec{y}$ within $\mathcal{R}(A)$. To see this, consider any another point $\vec{u} \in \mathcal{R}(A)$ and define $\vec{v} \doteq \vec{y} - \vec{u}$. This gives the following diagram:



To complete our proof, we define $\vec{w} \doteq \vec{z} - \vec{u}$, noting that the angle $\vec{u} \to \vec{z} \to \vec{y}$ is a right angle; in other words, $\vec{w}$ and $\vec{e}$ are orthogonal. This gives the following picture.



By the Pythagorean theorem, we see that

$$\begin{aligned}
\|\vec{y} - \vec{u}\|_2^2 &= \|\vec{v}\|_2^2 && (1.11)\\
&= \|\vec{w}\|_2^2 + \|\vec{e}\|_2^2 && (1.12)\\
&= \underbrace{\|\vec{z} - \vec{u}\|_2^2}_{>0} + \|\vec{e}\|_2^2 && (1.13)\\
&> \|\vec{e}\|_2^2 && (1.14)\\
&= \|\vec{y} - \vec{z}\|_2^2 . && (1.15)
\end{aligned}$$

Therefore, $\vec{z}$ is the closest point to $\vec{y}$ within $\mathcal{R}(A)$.

Now, we want to find $\vec{z} \in \mathcal{R}(A)$, i.e., the orthogonal projection of $\vec{y}$ onto $\mathcal{R}(A)$, such that $\vec{e} = \vec{y} - \vec{z}$ is orthogonal to all vectors in $\mathcal{R}(A)$. By the definition of $\mathcal{R}(A)$, it's equivalent to find $\vec{x}^\star \in \mathbb{R}^n$ such that $\vec{y} - A\vec{x}^\star$ is orthogonal to all vectors in $\mathcal{R}(A)$. Since the columns of A form a spanning set for $\mathcal{R}(A)$, it's equivalent to find $\vec{x}^\star \in \mathbb{R}^n$ such that $\vec{y} - A\vec{x}^\star$ is orthogonal to all columns of A. This implies

$$\begin{aligned}
\vec{0} &= A^\top(\vec{y} - A\vec{x}^\star) && (1.16)\\
&= A^\top\vec{y} - A^\top A\vec{x}^\star && (1.17)\\
\implies A^\top A\vec{x}^\star &= A^\top\vec{y} && (1.18)\\
\implies \vec{x}^\star &= (A^\top A)^{-1}A^\top\vec{y}. && (1.19)
\end{aligned}$$

Here $A^\top A$ is invertible because A has full column rank. □

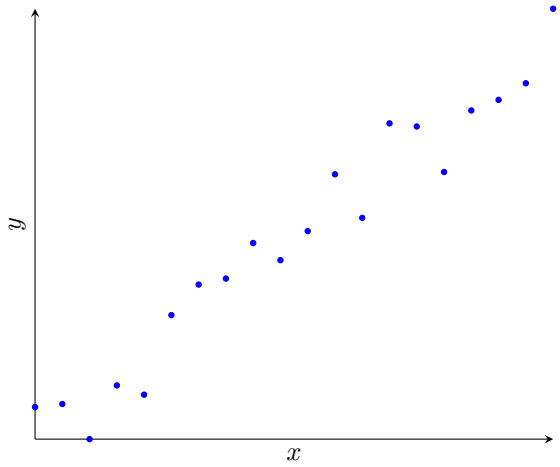
We'll conclude with a statistical application of least squares to linear regression. Suppose we are given data $(x_1, y_1), \ldots, (x_n, y_n)$, and want to fit an *affine* model $y = mx + b$ through these data points. This corresponds to approximately solving the system

$$\begin{aligned}
mx_1 + b &= y_1\\
mx_2 + b &= y_2\\
&\vdots\\
mx_n + b &= y_n.
\end{aligned} \qquad (1.20)$$

Formulating it in terms of vectors and matrices, we have

$$\begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{bmatrix} \begin{bmatrix} m \\ b \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}. \tag{1.21}$$

In the case where the data is noisy or inconsistent with the model, as in the below figure, the linear system will be overdetermined and have no solutions. Then, we find an approximate solution – a line of best fit – via least squares on the above system.



As a last note, solving least squares (and similar problems) is easy because it is a so-called *convex* problem. Convex problems are easy to solve because any local optimum is a global optimum, which allows us to use a variety of simple techniques to find global optima. It is generally much more difficult to solve non-convex problems, though we solve a few during this course.

We discuss much more about convexity and convex problems later in the course.

1.3 Solution Concepts and Notation

Sometimes we assign values to our optimization problems. For example in the framework of (1.5) we may write

$$\begin{aligned} p^\star = \min_{\vec{x} \in \mathbb{R}^n} \quad & f_0(\vec{x}) \\ \text{s.t.} \quad & f_i(\vec{x}) \leq 0 \quad \forall i \in \{1, \dots, m\} \\ & h_j(\vec{x}) = 0 \quad \forall j \in \{1, \dots, p\}. \end{aligned} \tag{1.22}$$

On the other hand, in the framework of (1.7) and using the definition of Ω in (1.6), we may write[1].

$$p^\star = \min_{\vec{x} \in \Omega} f_0(\vec{x}). \tag{1.23}$$

This means that $p^\star \in \mathbb{R}$ is the minimum value of f_0 over all $\vec{x} \in \Omega$; formally,

$$p^\star = \min_{\vec{x} \in \Omega} f_0(\vec{x}) \doteq \min\{f_0(\vec{x}) \mid \vec{x} \in \Omega\}. \tag{1.24}$$

[1]For the case where the minimum does not exist, but the infimum is finite, please see Section 1.4

As an example, consider the two-element set $\Omega = \{0, 1\}$ and $f_0(x) = 3x^2 + 2$. Then $p^\star = \min\{f(0), f(1)\} = \min\{2, 5\} = 2$. We emphasize that $p^\star$ is a *real number*, not a vector.

To extract the minimizers, i.e., the points $\vec{x} \in \Omega$ which minimize $f_0(\vec{x})$, we use the argmin notation, which gives us the set of arguments which minimize our objective function. Formally, we define:

$$\operatorname*{argmin}_{\vec{x}\in\Omega} f_0(\vec{x}) \doteq \left\{ \vec{x} \in \Omega \,\middle|\, f_0(\vec{x}) = \min_{\vec{u}\in\Omega} f_0(\vec{u}) \right\} \tag{1.25}$$

We can thus write the set of solutions to (1.5) as

$$\begin{aligned} \operatorname*{argmin}_{\vec{x}\in\mathbb{R}^n} \quad & f_0(\vec{x}) \\ \text{s.t.} \quad & f_i(\vec{x}) \leq 0 \quad \forall i \in \{1, \ldots, m\} \\ & h_j(\vec{x}) = 0 \quad \forall j \in \{1, \ldots, p\}. \end{aligned} \tag{1.26}$$

And, as just discussed, we can write the set of solutions to (1.7) as

$$\operatorname*{argmin}_{\vec{x}\in\Omega} f_0(\vec{x}). \tag{1.27}$$

We emphasize that the argmin is a *set of vectors*, any of which are an optimal solution, i.e., a minimizer, of the optimization problem at hand. It is possible for the argmin to contain 0 vectors (in which case the minimum value is not realized and the problem has no global optima), any positive number of vectors, or an infinite number of vectors.

Let us consider the same example as before. In particular, consider the two-element set $\Omega = \{0, 1\}$ and $f_0(x) = 3x^2 + 2$. Then $\operatorname{argmin}_{x\in\Omega} f_0(x) = \{0\}$. But, in different scenarios, the argmin can have zero elements; for example, if $f_0(x) = 3x$, then $\operatorname{argmin}_{x\in\mathbb{R}} f_0(x) = \emptyset$. And it can have multiple elements; for example, if $f_0(x) = 3x^2(x-1)^2$, then $\operatorname{argmin}_{x\in\mathbb{R}} f_0(x) = \{0, 1\}$. It can even have infinitely many elements; for example, if $f_0(x) = 0$, then $\operatorname{argmin}_{x\in\mathbb{R}} f_0(x) = \mathbb{R}$.

Though we must remember to keep in mind that technically argmin is a set, in the problems we study, it usually contains exactly one element. Thus, instead of writing, for example, $\vec{x}^\star \in \operatorname{argmin}_{\vec{x}\in\Omega} f_0(\vec{x})$, we may also write $\vec{x}^\star = \operatorname{argmin}_{\vec{x}\in\Omega} f_0(\vec{x})$. The former expression is technically more correct, but both usages are fine, if — and only if — the argmin in question contains exactly one element.

1.4 (OPTIONAL) Infimum Versus Minimum

There is one remaining issue with our formulation, which we can conceptually consider as a "corner case". What happens if the minimum does not exist? This may seem like a very esoteric case, yet one can construct a straightforward example, such as the following. We know that the minimum of any set of numbers must be contained in the set. But what happens if we try to find the minimum of the open interval $(0, 1)$? For any $x \in (0, 1)$ which we claim to be our minimum, we see that $\frac{x}{2}$ is also contained in $(0, 1)$ and is smaller than x, which is a contradiction to our claim. Thus the set $(0, 1)$ has no minimum.

It seems like 0 is a useful notion of "minimum" for this set — that is, it's the largest number which is $\leq$ all numbers in the set, i.e., its "greatest lower bound" — but it isn't contained in the set and thus cannot be the minimum. Fortunately, this notion of greatest lower bound of a set is formalized in real analysis as the concept of an "infimum", denoted $\inf$. For our purposes, we can think of the infimum as a generalization of the minimum which takes care of these corner cases and always exists. When the minimum exists, it is always equal to the infimum.

Based on this discussion, we can write our optimization problems as

$$p^\star = \inf_{\vec{x}\in\mathbb{R}^n} \quad f_0(\vec{x}) \tag{1.28}$$

$$\begin{aligned} \text{s.t.} \quad & f_i(\vec{x}) \leq 0 \quad \forall i \in \{1, \ldots, m\} \\ & h_j(\vec{x}) = 0 \quad \forall j \in \{1, \ldots, p\}. \end{aligned}$$

and

$$p^\star = \inf_{\vec{x} \in \Omega} f_0(\vec{x}). \tag{1.29}$$

However, the argmin retains the same definition. In fact, one can prove that if we replaced the $\min$ in the argmin definition (1.25) with $\inf$, that this "new" argmin would be exactly equivalent in every case to the "old" argmin, which we use henceforth. The analogous quantity to infimum for maximization — that is, the appropriate generalization of $\max$ — is the supremum, denoted $\sup$.

Interested readers are encouraged to consult a real analysis textbook such as [3] for a more comprehensive coverage.

Though we have gone over the technical details here, for the rest of the course we will omit them for simplicity, and stick to using $\min$ and $\max$ (meaning $\inf$ and $\sup$ when the minimum and maximum do not exist).

Chapter 2

Linear Algebra Review

Relevant sections of the textbooks:

- [1] Appendix A.
- [2] Chapters 2, 3, 4, 5.

2.1 Norms

2.1.1 Definitions

Definition 5 (Norm)
Let $\mathcal{V}$ be a vector space over $\mathbb{R}$. A function $f\colon \mathcal{V} \to \mathbb{R}$ is a norm if:

- Positive definiteness: $f(\vec{x}) \geq 0$ for all $\vec{x} \in \mathcal{V}$, and $f(\vec{x}) = 0$ if and only if $\vec{x} = \vec{0}$.
- Positive homogeneity: $f(\alpha\vec{x}) = |\alpha|\, f(\vec{x})$ for all $\alpha \in \mathbb{R}$ and $\vec{x} \in \mathcal{V}$.
- Triangle inequality: $f(\vec{x} + \vec{y}) \leq f(\vec{x}) + f(\vec{y})$ for all $\vec{x}, \vec{y} \in \mathcal{V}$.

We can check that the familiar Euclidean norm $\|\cdot\|_2 : \vec{x} \mapsto \sqrt{\sum_{i=1}^n x_i^2}$ satisfies these properties. A generalization of the Euclidean norm is the following very useful class of norms.

Definition 6 (ℓ^p Norms)
Let $1 \leq p < \infty$. The ℓ^p-norm on $\mathbb{R}^n$ is given by

$$\|\vec{x}\|_p \doteq \left(\sum_{i=1}^n |x_i|^p\right)^{1/p}. \tag{2.1}$$

The ℓ^∞-norm on $\mathbb{R}^n$ is given by

$$\|\vec{x}\|_\infty \doteq \max_{i\in\{1,\ldots,n\}} |x_i|\,. \tag{2.2}$$

Example 7 (Examples of ℓ^p Norms)**.**

(a) The Euclidean norm, given by $\|\vec{x}\|_2 = \sqrt{\sum_{i=1}^n x_i^2}$, is an ℓ^p-norm for $p = 2$. (This is why we gave the subscript 2 to the Euclidean norm previously).

(b) The ℓ^1-norm is given by $\|\vec{x}\|_1 = \sum_{i=1}^n |x_i|$.

(c) The ℓ^∞-norm, given by $\|\vec{x}\|_\infty = \max_{i\in\{1,\dots,n\}} |x_i|$, is the limit of the ℓ^p norms as $p \to \infty$:

$$\|\vec{x}\|_\infty = \lim_{p\to\infty} \|\vec{x}\|_p . \tag{2.3}$$

We do not prove this here; it is left as an exercise.

2.1.2 Inequalities

There are a variety of useful *inequalities* which are associated with the ℓ^p norms. Before we provide them, we will take a second to discuss the importance of inequalities for optimization.

A priori, it may not be clear why we need to care about inequalities; why does it matter whether one arrangement of variables is always greater or less than another arrangement? It turns out that such inequalities are very helpful for characterizing the minimum and maximum of a given set of things; we can obtain upper bounds and lower bounds for things using these inequalities. This is definitely very helpful for optimization.

With that out of the way, let us get to the first major inequality.

Theorem 8 (Cauchy-Schwarz Inequality)
For any $\vec{x}, \vec{y} \in \mathbb{R}^n$, we have

$$\left|\vec{x}^\top \vec{y}\right| \le \|\vec{x}\|_2 \|\vec{y}\|_2 . \tag{2.4}$$

Proof. Let θ be the angle between $\vec{x}$ and $\vec{y}$. We write

$$\begin{aligned} \left|\vec{x}^\top \vec{y}\right| &= \left|\|\vec{x}\|_2 \|\vec{y}\|_2 \cos\theta\right| && (2.5)\\ &= \|\vec{x}\|_2 \|\vec{y}\|_2 |\cos\theta| && (2.6)\\ &\le \|\vec{x}\|_2 \|\vec{y}\|_2 . && (2.7) \end{aligned}$$

□

We can get this result for ℓ^2 norms. A natural next question is whether we can generalize it to ℓ^p norms for $p \neq 2$. It turns out that we can, as we demonstrate shortly.

Theorem 9 (Hölder's Inequality)
Let $1 \le p, q \le \infty$ such that $\frac{1}{p} + \frac{1}{q} = 1$.[a] Then for any $\vec{x}, \vec{y} \in \mathbb{R}^n$, we have

$$\left|\vec{x}^\top \vec{y}\right| \le \sum_{i=1}^n |x_i y_i| \le \|\vec{x}\|_p \|\vec{y}\|_q . \tag{2.8}$$

[a] Such pairs (p, q) are called *Hölder conjugates.*

This inequality collapses to Cauchy-Schwarz Inequality when $p = q = 2$. The proof is out of scope for now since it uses convexity.

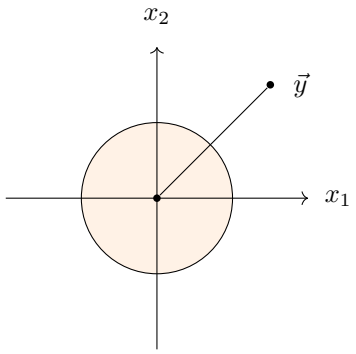
Example 10 (Dual Norms)**.** Fix $\vec{y} \in \mathbb{R}^n$. Let us solve the problem:

$$\max_{\substack{\vec{x}\in\mathbb{R}^n \\ \|\vec{x}\|_p \leq 1}} \vec{x}^\top \vec{y}. \tag{2.9}$$

It is initially difficult to see how to proceed, so let us simplify the problem to get back onto familiar territory. We start with $p = 2$, so that the problem becomes:

$$\max_{\substack{\vec{x}\in\mathbb{R}^n \\ \|\vec{x}\|_2 \leq 1}} \vec{x}^\top \vec{y}. \tag{2.10}$$

For $n = 2$, the feasible set and $\vec{y}$ together look like the following:



For any $\vec{x} \in \mathbb{R}^n$, and θ the angle between $\vec{x}$ and $\vec{y}$, we have

$$\vec{x}^\top \vec{y} = \|\vec{x}\|_2 \|\vec{y}\|_2 \cos\theta. \tag{2.11}$$

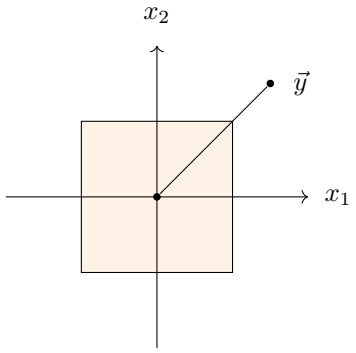
This term is maximized when $\cos\theta = 1$, or equivalently $\theta = 0$. Thus $\vec{x}$ and $\vec{y}$ must point in the same direction, i.e., $\vec{x}$ is a scalar multiple of $\vec{y}$. And since we want to maximize this dot product, we must choose $\vec{x}$ to maximize $\|\vec{x}\|_2$ subject to the constraint $\|\vec{x}\|_2 \leq 1$. Thus, we choose an $\vec{x}$ which has $\|\vec{x}\|_2 = 1$ and points in the same direction as $\vec{y}$. This gives $\vec{x}^\star = \vec{y}/\|\vec{y}\|_2$. Thus,

$$\max_{\substack{\vec{x}\in\mathbb{R}^n \\ \|\vec{x}\|_2 \leq 1}} \vec{x}^\top \vec{y} = (\vec{x}^\star)^\top \vec{y} = \left(\frac{\vec{y}}{\|\vec{y}\|_2}\right)^\top \vec{y} = \frac{\vec{y}^\top \vec{y}}{\|\vec{y}\|_2} = \frac{\|\vec{y}\|_2^2}{\|\vec{y}\|_2} = \|\vec{y}\|_2 . \tag{2.12}$$

Now let us try $p = \infty$. The problem becomes

$$\max_{\substack{\vec{x}\in\mathbb{R}^n \\ \|\vec{x}\|_\infty \leq 1}} \vec{x}^\top \vec{y}. \tag{2.13}$$

The feasible set and $\vec{y}$ are given by the following diagram.



Motivated by this diagram, we see that the constraint $\|\vec{x}\|_\infty \leq 1$ is equivalent to the $2n$ constraints $-1 \leq x_i$ and $x_i \leq 1$. Also, writing out the objective function

$$\vec{x}^\top \vec{y} = \sum_{i=1}^{n} x_i y_i = x_1 y_1 + x_2 y_2 + \cdots + x_n y_n, \tag{2.14}$$

we see that the problem is

$$\begin{aligned} \max_{\vec{x} \in \mathbb{R}^n} \quad & (x_1 y_1 + x_2 y_2 + \cdots + x_n y_n) \\ \text{s.t.} \quad & -1 \leq x_i \leq 1, \qquad \forall i \in \{1, \ldots, n\}. \end{aligned} \tag{2.15}$$

This problem has an interesting structure that will be repeated several times in the problems we discuss in this class. Namely, the objective function is the sum of several terms, each of which involves only one x_i. And the constraints are able to be partitioned into some groups, where the constraints in each group constrain only one x_i. Thus, this problem is *separable* into n different scalar problems, such that the optimal solutions for each scalar problem form an optimal solution for the vector problem. Namely, the problems are

$$\max_{\substack{x_i \in \mathbb{R} \\ -1 \leq x_i \leq 1}} \quad x_i y_i \tag{2.16}$$

We solve this much simpler problem by hand. If $y_i > 0$ then $x_i^\star = 1$; on the other hand, if $y_i \leq 0$ then $x_i^\star = -1$. To summarize, $x_i^\star = \operatorname{sgn}(y_i)$, so that $x_i^\star y_i = |y_i|$.

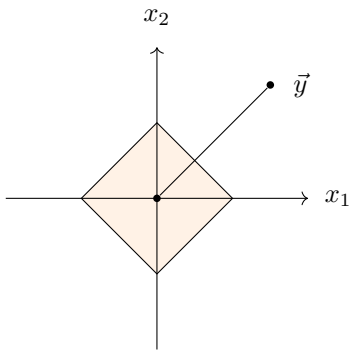
Putting all the scalar problems together, we see that $\vec{x}^\star = \operatorname{sgn}(\vec{y})$, and the vector problem's optimal value is given by

$$\max_{\substack{\vec{x} \in \mathbb{R}^n \\ \|\vec{x}\|_\infty \leq 1}} \vec{x}^\top \vec{y} = (\vec{x}^\star)^\top \vec{y} = \sum_{i=1}^{n} x_i^\star y_i = \sum_{i=1}^{n} \operatorname{sgn}(y_i) y_i = \sum_{i=1}^{n} |y_i| = \|\vec{y}\|_1 . \tag{2.17}$$

As a final exercise, we consider $p = 1$, so that the problem becomes

$$\max_{\substack{\vec{x} \in \mathbb{R}^n \\ \|\vec{x}\|_1 \leq 1}} \vec{x}^\top \vec{y}. \tag{2.18}$$

For $n = 2$, the feasible set and $\vec{y}$ together look like the following:



We now bound the objective as

$$\vec{x}^\top \vec{y} \leq \left| \vec{x}^\top \vec{y} \right| \tag{2.19}$$

$$= \left| \sum_{i=1}^{n} x_i y_i \right| \tag{2.20}$$

$$\leq \sum_{i=1}^{n} |x_i y_i| \qquad \text{by triangle inequality} \tag{2.21}$$

$$= \sum_{i=1}^{n} |x_i| \, |y_i| \tag{2.22}$$

$$\leq \sum_{i=1}^{n} |x_i| \left(\max_{i \in \{1,\ldots,n\}} |y_i| \right) \tag{2.23}$$

$$= \left(\max_{i \in \{1,\ldots,n\}} |y_i| \right) \sum_{i=1}^{n} |x_i| \tag{2.24}$$

$$= \|\vec{y}\|_\infty \, \|\vec{x}\|_1 \tag{2.25}$$

$$\leq \|\vec{y}\|_\infty . \tag{2.26}$$

Thus we have

$$\max_{\substack{\vec{x} \in \mathbb{R}^n \\ \|\vec{x}\|_1 \leq 1}} \vec{x}^\top \vec{y} \leq \|\vec{y}\|_\infty . \tag{2.27}$$

This inequality is actually an equality. To show this, we need to show the reverse inequality

$$\max_{\substack{\vec{x} \in \mathbb{R}^n \\ \|\vec{x}\|_1 \leq 1}} \vec{x}^\top \vec{y} \underbrace{\geq}_{\text{need to show}} \|\vec{y}\|_\infty . \tag{2.28}$$

And showing *this* inequality amounts to choosing, for our fixed $\vec{y}$, a $\vec{x}$ such that $\|\vec{x}\|_1 \leq 1$ and $\vec{x}^\top \vec{y} \geq \|\vec{y}\|_\infty$. This is also called "showing the maximum is attained". To do this, we can find an $\vec{x}$ such that $\|\vec{x}\|_p \leq 1$ and all the inequalities in the chain are met with equality.

- First, the inequality in (2.21) is a triangle inequality with the absolute value, i.e., $|\sum_{i=1}^n x_i y_i| \leq \sum_{i=1}^n |x_i y_i|$. To make sure this is an equality, it's enough to make sure that all terms $x_i y_i$ are the same sign or 0.
- Next, the inequality in (2.23) says that $\sum_{i=1}^n |x_i| \, |y_i| \leq \sum_{i=1}^n |x_i| \left(\max_{i \in \{1,\ldots,n\}} |y_i| \right)$. The most obvious instance in which this inequality is met with equality is when $|y_i| = \max_{j \in \{1,\ldots,n\}} |y_j|$ for all i. But we can't choose $\vec{y}$, as it's fixed, so we can't be assured that this holds. An alternate way in which this holds is that $|x_i| = 0$ for all i for which $|y_i| \neq \max_{j \in \{1,\ldots,n\}} |y_j|$, i.e., $i \notin \operatorname{argmax}_{j \in \{1,\ldots,n\}} |y_j|$.
- Finally, the inequality in (2.26) says that $\|\vec{x}\|_1 \, \|\vec{y}\|_\infty \leq \|\vec{y}\|_\infty$; to meet this inequality with equality, it is sufficient to have $\|\vec{x}\|_1 = 1$.

To meet all three of these constraints, we can construct $\vec{x}^\star$ via the following process:

- For each $i \notin \operatorname{argmax}_{j \in \{1,\ldots,n\}} |y_j|$, set $\widetilde{x}_i = 0$, as per the second bullet point above.
- For each $i \in \operatorname{argmax}_{j \in \{1,\ldots,n\}} |y_j|$, set $\widetilde{x}_i = \operatorname{sgn}(y_i)$, as per the first bullet point above.
- To get the true solution vector $\vec{x}^\star$, divide $\vec{\widetilde{x}}$ by $\left\| \vec{\widetilde{x}} \right\|_1$; that is, $\vec{x}^\star = \vec{\widetilde{x}} / \left\| \vec{\widetilde{x}} \right\|_1$. This ensures that $\|\vec{x}^\star\|_1 = 1$, as per the third bullet point above.

This $\vec{x}^\star$ "achieves the maximum", showing that

$$\max_{\substack{\vec{x} \in \mathbb{R}^n \\ \|\vec{x}\|_1 \leq 1}} \vec{x}^\top \vec{y} = \|\vec{y}\|_\infty . \tag{2.29}$$

This notion where the ℓ^2-norm constraint leads to the ℓ^2-norm objective, the ℓ^∞-norm constraint leads to the ℓ^1-norm objective, and the ℓ^1-norm constraint leads to the ℓ^∞-norm objective, hints at a greater pattern. Indeed, one can show that for $1 \leq p, q \leq \infty$ such that $\frac{1}{p} + \frac{1}{q} = 1$, an ℓ^p-norm constraint leads to an ℓ^q-norm objective:

$$\max_{\substack{\vec{x} \in \mathbb{R}^n \\ \|\vec{x}\|_p \leq 1}} \vec{x}^\top \vec{y} = \|\vec{y}\|_q . \tag{2.30}$$

As before, we can prove this equality by proving the two constituent inequalities:

$$\max_{\substack{\vec{x} \in \mathbb{R}^n \\ \|\vec{x}\|_p \leq 1}} \vec{x}^\top \vec{y} \leq \|\vec{y}\|_q \qquad \text{and} \qquad \max_{\substack{\vec{x} \in \mathbb{R}^n \\ \|\vec{x}\|_p \leq 1}} \vec{x}^\top \vec{y} \geq \|\vec{y}\|_q . \tag{2.31}$$

The proof of the first inequality ($\leq$) follows from applying Hölder's inequality to the objective function:

$$\max_{\substack{\vec{x} \in \mathbb{R}^n \\ \|\vec{x}\|_p \leq 1}} \vec{x}^\top \vec{y} \leq \max_{\substack{\vec{x} \in \mathbb{R}^n \\ \|\vec{x}\|_p \leq 1}} \|\vec{x}\|_p \|\vec{y}\|_q = \|\vec{y}\|_q \cdot \max_{\substack{\vec{x} \in \mathbb{R}^n \\ \|\vec{x}\|_p \leq 1}} \|\vec{x}\|_p = \|\vec{y}\|_q . \tag{2.32}$$

The second inequality ($\geq$) can follow if, for our fixed choice of $\vec{y}$, we produce some $\vec{x}$ such that $\|\vec{x}\|_p \leq 1$ and $\vec{x}^\top \vec{y} \geq \|\vec{y}\|_q$, i.e., "the maximum is attained". This is more complicated to do, and we won't do it here.

The above equality (2.30) means that the norms $\|\cdot\|_p$ and $\|\cdot\|_q$ are so-called *dual* norms. We will explore aspects of duality later in the course, though frankly we are just scratching the surface.

These problems, which are short and easy to state, contain a couple of core ideas within their solutions, which are broadly generalizable to a lot of optimization problems. For your convenience, we discuss these explicitly below.

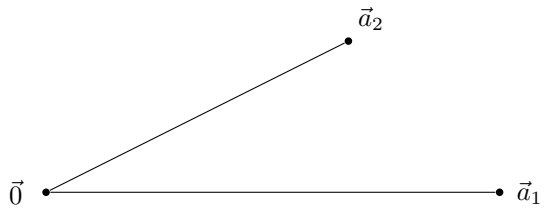
Problem Solving Strategy 11 (Separating Vector Problems into Scalar Problems)**.** *When trying to simplify an optimization problem, try to see if you can simplify it into several independent scalar problems. Then solve each scalar problem — this is usually much easier than solving the whole vector problem at once. The optimal solutions to each scalar problem will then form the optimal solution to the whole vector problem.*

Problem Solving Strategy 12 (Proving Optimality in an Optimization Problem)**.** *To solve an optimization problem, you can use inequalities to bound the objective function, and then try to show that this bound is tight by finding a feasible choice of optimization variable which makes all the inequalities into equalities.*

2.2 Gram-Schmidt and QR Decomposition

The Gram-Schmidt algorithm is a way to turn a linearly independent set $\{\vec{a}_1, \ldots, \vec{a}_k\}$ of vectors into an *orthonormal* set $\{\vec{q}_1, \ldots, \vec{q}_k\}$ which spans the same space. To reiterate, an orthonormal set is a set of vectors in which each vector has norm 1 and is orthogonal to all others in the basis.

Suppose for simplicity that $n = k = 2$, and that we have the following vectors.

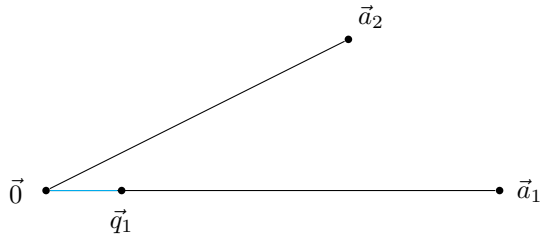


We begin with $\vec{a}_1$. We want to construct a vector $\vec{q}_1$ such that

- it's orthogonal to all the $\vec{q}_i$ which came before it — which is none of them, so we don't have to worry; and
- it has unit norm, so $\|\vec{q}_1\|_2 = 1$.

To achieve this, the simplest choice is

$$\vec{q}_1 \doteq \frac{\vec{a}_1}{\|\vec{a}_1\|_2}. \tag{2.33}$$



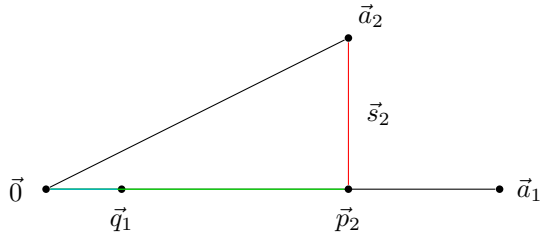
Then we go to $\vec{a}_2$. To find $\vec{q}_2$ which is orthogonal to all the $\vec{q}_i$ before it — that is, $\vec{q}_1$ — we subtract off the orthogonal projection of $\vec{a}_2$ onto $\vec{q}_1$ from $\vec{a}_2$. The orthogonal projection of $\vec{a}_2$ onto $\vec{q}_1$ is given by

$$\vec{p}_2 \doteq \vec{q}_1(\vec{q}_1^{\top}\vec{a}_2) \tag{2.34}$$

and so the projection residual is given by

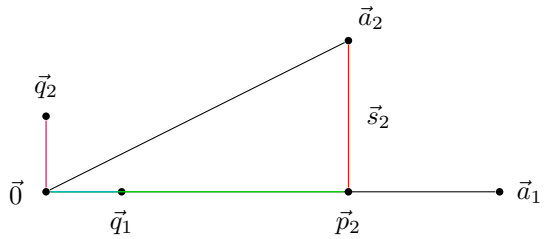
$$\vec{s}_2 \doteq \vec{a}_2 - \vec{p}_2 = \vec{a}_2 - \vec{q}_1(\vec{q}_1^{\top}\vec{a}_2). \tag{2.35}$$

Note that these formulas *only* hold because $\vec{q}_1$ is normalized, i.e., has norm 1.



While $\vec{s}_2$ is orthogonal to $\vec{q}_1$, because we want a $\vec{q}_2$ that is normalized, we normalize $\vec{s}_2$ to get $\vec{q}_2$:

$$\vec{q}_2 \doteq \frac{\vec{s}_2}{\|\vec{s}_2\|_2}. \tag{2.36}$$



If we had a vector $\vec{q}_3$ (and weren't limited by drawing in 2D space), we would ensure that $\vec{q}_3$ were orthogonal to $\vec{q}_1$ and $\vec{q}_2$, as well as normalized, in a similar way as before. First we would compute the projection

$$\vec{p}_3 \doteq \vec{q}_1(\vec{q}_1^{\top}\vec{a}_3) + \vec{q}_2(\vec{q}_2^{\top}\vec{a}_3). \tag{2.37}$$