ACM 122: Mathematical Optimization
Lecture Notes

Lecturer: Venkat Chandrasekaran, Author: Massimiliano de Sa

Contents

1 Lecture 1: Introduction to Optimization

Broadly, we define optimization as the discipline concerned with identifying the "best" element according to some criterion from a collection of possible choices. More concretely, we describe optimization as the following: over all $x \in G$, we wish to minimize $f(x)$ subject to x belonging to some constrain set S.
Here, G is known as the **ground set**, $f : G \to \mathbb{R}$ is known as the **objective function**, and $S \subset G$ is known as the constraint set. S consists of all possible choices of x.
Now, we consider some common classes of optimization problems.

1. **Continuous Optimization:** First, we consider the ground set $G = \mathbb{E}^d$, d dimensional Euclidean space. This class of optimization problems is called the set of continuous optimization problems. The style of analysis used in this area is mainly real analysis.

2. **Integer Programming:** Next, we consider $G = \mathbb{Z}^d$, the set of d dimensional integers. This class of problems is called integer programming. The math used in this are mainly comes from number theory and abstract algebra.

3. **Combinatorial Optimization:** G can also be the set of all graphs on d nodes. This family is called combinatorial optimization. The math used in this field mainly relies on ideas from combinatorics, graph theory, and other areas of discrete math.

4. **Calculus of Variations:** Finally, we consider the set G of all continuous functions from $[0, 1]$ to $[0, 1]$. This ground set is infinite dimensional. The family of problems associated with this ground set is called calculus of variations, and relies on functional analysis.

In this course, we'll mainly focus on the case of continuous optimization, but will see applications to integer programming and combinatorial optimization.
The examples of d dimensional Euclidean space that we'll commonly encounter are the following:

1. $\mathbb{R}^d$, the set of vectors of length d with real entries. The inner product associated with this space is the dot product:

$$\langle x, y \rangle = \sum_{i=1}^{d} x_i y_i \tag{1}$$

2. $\mathbb{R}^{m \times n}$, the set of $m \times n$ matrices with real entries. The associated inner product on this space is the trace inner product:

$$\langle X, Y \rangle = tr(X^* Y) \tag{2}$$

   This space is $m \cdot n$ dimensional.

3. $\mathbb{S}^n$, the set of $n \times n$ symmetric matrices with real entries. The associated inner product is once again the trace inner product:

$$\langle X, Y \rangle = tr(X^*Y) = tr(XY) \tag{3}$$

This space has dimension $n(n+1)/2$, or "n choose 2," in other words. The dimension of this space may be calculated from a symmetry argument. We know there are n elements on the diagonal of $X \in \mathbb{S}^n$ and the same number of elements p above and below the diagonal. Thus:

$$2p + n = n^2 \tag{4}$$

$$p = n(n-1)/2 \tag{5}$$

The total number of "free" elements in $X \in \mathbb{S}^n$ is therefore:

$$p + n = n(n+1)/2 \tag{6}$$

This defines the dimension of the space.

Now, we discuss the general form of a continuous optimization problem. Formally, a continuous optimization problem is described:

$$v = \inf_{x \in \mathbb{E}^d} f(x) \tag{7}$$

$$s.t.\ x \in S \tag{8}$$

Here, $f : \mathbb{E}^d \to \mathbb{R}$ is the objective function. Its domain, $dom(f)$, can be a proper subset of $\mathbb{E}^d$. Note that the constraint set S is *not necessarily* a subset of $dom(f)$ - it's possible that there is no intersection at all between the two! They are formulated entirely separately.
On the second line of this problem, $s.t.\ x \in S$, we specify that the optimization problem is "subject to" the constraint that $x \in S \subset \mathbb{E}^d$. The *s.t.* can also be thought of as meaning "such that."
An element of $S \cap dom(f)$ is known as a **feasible point**, or **feasible solution**, of the optimization problem. This is because it is in the constraint set and the domain of f.
v is called the **optimal value**, and is defined:

$$v = \inf\{f(x) | x \in dom(f) \cap S\} \tag{9}$$

Here, inf denotes infimum, the largest lower bound of a given subset of $\mathbb{R}$. Note that we always take $\mathbb{R}$ to be the codomain of f, as $\mathbb{R}$ has a natural ordering of elements that many other spaces do not.
Note that here, we allow $-\infty$ to be the infimum of a set in the case that the set is not bounded below. Additionally, recall that $\inf S$ is not necessarily contained in S.
For instance, consider the problem:

$$\inf_{x \in \mathbb{R}} \exp(x) \tag{10}$$

The minimum doesn't exist in this case, but the infimum does! The solution to this problem is thus $v = 0$. Remember - for any given subset of $\mathbb{R}$, the infimum will always exist.
A feasible point $\hat{x} \in dom(f) \cap S$ for which $f(\hat{x}) = v$ is called an **optimal solution** of the problem. This is any point $\hat{x}$ which achieves the optimal value. Note that the optimal solution might not always exist, even when the optimal value might! In the case of $\exp x$ above, for instance, the optimal value $v = 0$ exists but an optimal solution does not.
We now provide some more remarks regarding this structure of problem. A problem in which $dom(f) \cap S = \mathbb{E}^d$ is called **unconstrained**, while a problem in which $dom(f) \cap S = \emptyset$ is called **infeasible**.
In the case of infeasibility, we take $v = +\infty$ by convention! To reason about this from an intuitive perspective, we can imagine the solution to all optimization problems "starting at" $+\infty$, and the process of finding an optimal value as "pushing the solution down" from ∞. In the case of infeasibility, we cannot "push the solution down," so it remains at ∞. Mathematically, this corresponds to saying $\inf \emptyset = +\infty$.
If $v = -\infty$, on the other hand, the problem is said to be **unbounded below**, or simply **unbounded**. For example, consider:

$$\inf_{x \in \mathbb{R}} x \tag{11}$$

This problem is unbounded below. Note that by conventions from operations research, the variable x is commonly referred to as the **decision variable**.
We conclude this lecture by asking a few questions. First - what does it mean to solve an optimization problem? In particular, given a candidate solution, how do we *certify* that the solution is optimal? We'd like to find some **certificate** that tells us that the answer is correct.
Secondly, we want to know how to *describe* an optimization problem. We will pick up with these questions in the next lecture.

2 Lecture 2: Certifying Optimality

Last time, we brought up two important questions in continuous optimization: how do we *certify* optimality, and how do we *describe* optimization problems?
In this lecture, we'll begin developing an answer to the first question, on certifying optimality.
Our approach to certifying optimality will be to develop methods to certify lower bounds on the optimal value of a problem. Consider the problem:

$$v = \inf_{x \in \mathbb{E}^d} f(x) \tag{12}$$

$$s.t.\ x \in S \tag{13}$$

For $f : \mathbb{E}^d \to \mathbb{R}$, $dom(f) \subset \mathbb{E}^d, S \subset \mathbb{E}^d$, suppose we're given some γ that is *claimed* to be a lower bound on v. How can we *certify* that γ is indeed a lower

bound on v?
Let's see why the problem of certifying γ is a lower bound is still a hard problem. Suppose we have a proposed lower bound γ and an optimal solution v. For γ to be a lower bound on v, we require that $f(x) \geq \gamma \forall x \in S$! This $\forall x \in S$ is what is challenging - we need to somehow verify that γ is a lower bound for every value of x.
As an exercise, imagine that we have a black box that tells us if γ is a lower bound or not. How could we use this to our advantage? If we can find an $\hat{x}$ that's a lower bound on $f(x)$ and satisfies $f(\hat{x}) \leq f(x)$, then $\hat{x}$ must give a solution to our problem!
In particular, given $\hat{x} \in S$ that is claimed to be an optimal solution, we could try to *certify* that $\hat{x}$ is optimal by setting $\gamma = f(\hat{x})$ and appealing to our lower bound certificate.[1]
As an illustration, consider the problem:

$$v = \inf_{x \in \mathbb{E}^d} ||x - b||^2 \tag{14}$$

$$s.t.\ x \in U \tag{15}$$

Here, $b \in \mathbb{E}^d$ is fixed and $U \subset \mathbb{E}^d$ is a subspace. Geometrically, we know from the orthogonality principle in linear algebra that the error between b and the subspace U is orthogonal to U.
Let's try to derive a lower bound on v, the optimal solution. As an easy lower bound, we know:

$$v \geq 0 \tag{16}$$

This is because the square of a norm is always greater than or equal to zero.
However, this lower bound doesn't use any information about the constraint set or the geometry of the problem! This makes $v \geq 0$ a weak certificate - it's a loose lower bound that doesn't allow us to say much about the solution to the problem. Mathematically, all it tells us is that for all $\gamma \leq 0$, we can certify that γ is a lower bound on v.
Let's derive a stronger certificate. Remember - our goal is to develop a lower bound with no reliance on x - this will enable us to form a bound that holds for all x. Fix any $\mu \in U^\perp$, the orthogonal complement of U in $\mathbb{E}^d$. Then:

$$||x - b||^2 = ||x - b + \mu||^2 + ||x - b||^2 - ||x - b + \mu||^2 \tag{17}$$

$$= ||x - b + \mu||^2 + [||x - b||^2 - (||x - b||^2 + ||\mu||^2 + 2\langle x - b, \mu \rangle)] \tag{18}$$

$$= ||x - b + \mu||^2 - 2\langle x, \mu \rangle + 2\langle b, \mu \rangle - ||\mu||^2 \tag{19}$$

Where we use the "completing the square" identity to get to line 2. So far, we still haven't used any geometric properties of the problem! Let's use these to get rid of the terms involving x.

[1] Notice that this method will only apply when an optimal solution to the problem exists.

We know that $||x - b + \mu||^2 \geq 0$ and $\langle x, \mu \rangle = 0$, since $x \in U$ by the constraint on the problem and $\mu \in U^\perp$ by definition. This forms the inequality:

$$||x - b||^2 \geq 2\langle b, \mu \rangle - ||\mu||^2 \tag{20}$$

We now have a lower bound that doesn't depend on x! It therefore holds for all $x \in U$. This is a stronger lower bound than $v \geq 0$. We can see this by setting $\mu = 0$ as a simple example. In this case, $2\langle b, \mu \rangle - ||\mu||^2 = 0$, which tells us that our new lower bound is *at least* as good as our old lower bound.
If some γ is claimed to be a lower bound on v, one way to certify it is a lower bound is to identify a $\mu \in U^\perp$ such that:

$$2\langle b, \mu \rangle - ||\mu||^2 \geq \gamma \tag{21}$$

Recall that for *every* fixed μ, this expression gives a lower bound on v! This is what allows us to conclude the statement above.
Now, we want to somehow maximize this lower bound on v! This will take us towards a solution. This technique won't give us a solution for every optimization problem, but will in this special case! This is because this family of certificates for this problem is very good.
Let's recap what we've covered so far. So far, we've seen:

1. Different lower bound certification techniques may very in their certification capabilities.

2. To produce a certifiable lower bound, we changed a statement involving a universal quantifier, $\forall x \in U, ||x - b|| \geq \gamma$, to one involving an existential quantifier:

$$\exists \mu \in U^\perp \; s.t. \; 2\langle b, \mu \rangle - ||\mu||^2 \geq \gamma \tag{22}$$

   This has simplified the problem.

If we are given a γ, and find a μ such that $2\langle b, \mu \rangle - ||\mu||^2 \geq \gamma$, then we know that γ is a lower bound! It's important to remember here that the existence of a certificate is not the same as the computation of a certificate. Here, we only care that the certificate "does the job" of verifying a lower bound - we don't care about the details of computing μ.
Consider the following question about the problem above. Can any lower bound on v be certified using $\mu \in U^\perp$? If we can find an $x \in U$ and a $\mu \in U^\perp$ such that $||x - b + \mu||^2 = 0$ and $2\langle x, \mu \rangle = 0$, then our lower bound $2\langle b, \mu \rangle - ||\mu||^2$ will be exact, as in this case:

$$||x - b||^2 = ||x - b + \mu||^2 - 2\langle x, \mu \rangle + 2\langle b, \mu \rangle - ||\mu||^2 \tag{23}$$

$$= 2\langle b, \mu \rangle - ||\mu||^2 \tag{24}$$

This *exact* lower bound will therefore give us a solution to the problem, as it is less than or equal to $f(x) \forall x \in \mathbb{E}^d$ and is an actual value of the objective

function.
Let's find such an x and μ and solve the optimization. We know from linear algebra that $v = ||P_{U^\perp}(b)||^2$, the orthogonal projection of b onto $U^\perp$ *should* be the solution to the problem. Is there a μ for which we can certify this v is a lower bound? Consider $\mu = P_{U^\perp}(b)$.

$$2\langle b, \mu\rangle - ||\mu||^2 = 2\langle b, P_{U^\perp}(b)\rangle - ||P_{U^\perp}(b)||^2 \tag{25}$$

$$= 2||P_{U^\perp}(b)||^2 - ||P_{U^\perp}(b)||^2 = v \tag{26}$$

Thus, such a μ exists! Therefore, v is a lower bound of the optimization problem. Since an x exists such that $v = f(x)$, v must be the solution of the problem.
Going beyond this simple problem, how do we think about certificates for lower bounds more generally? To make progress on this question, we'll reformulate the problem of certifying lower bounds to certifying **infeasibility**.
Specifically, for a problem:

$$v = \inf_{x \in \mathbb{E}^d} f(x) \tag{27}$$

$$s.t.\ x \in S \tag{28}$$

For $f(x) : \mathbb{E}^d \to \mathbb{R}$ with domain $dom(f)$. For $S \subset \mathbb{E}^d$, we have the equivalence:

$$f(x) \geq \gamma\ \forall x \in S \cap dom(f) \Leftrightarrow \{x \in \mathbb{E}^d | f(x) < \gamma, x \in dom(f) \cap S\} = \emptyset \tag{29}$$

In other words, γ is a lower bound over all feasible points if and only if $\{x \in \mathbb{E}^d | f(x) < \gamma, x \in dom(f) \cap S\} = \emptyset$. Now, we must discuss how to actually certify that this set is empty!
As an example, consider the set $\{x \in \mathbb{E}^d | Ax = b\}$, where $A : \mathbb{E}^d \to \mathbb{E}^n$ is a linear map and $b \in \mathbb{E}^n$. To show that the set is empty, we can use Fredholm's alternative, which states the following.

Proposition 1 ***Fredholm's Alternative***
For a linear map $A : \mathbb{E}^d \to \mathbb{E}^n$ and $b \in \mathbb{E}^n$, exactly one of the following is true:

1. *There exists $x \in \mathbb{E}^d$ such that $Ax = b$.*
2. *There exists $\mu \in \mathbb{E}^n$ such that $A^*\mu = 0$ and $\langle b, \mu\rangle = 1$.*

This allows us to conclude emptiness with the following: if $\mu \in \mathbb{E}^n$ such that $A^*\mu = 0$ and $\langle b, \mu\rangle = 1$, then there is no x such that $Ax = b$. By showing such a μ exists, we may conclude emptiness. This defines a certificate.

3 Lecture 3: Convex Sets and Hyperplanes

In this lecture, we'll begin by understanding what's actually going on, geometrically speaking, with Fredholm's alternative. First, we review what Fredholm's

alternative states.
Given a linear map $A : \mathbb{E}^d \to \mathbb{E}^n$ and $b \in \mathbb{E}^n$:

$$\{x \in \mathbb{E}^d | Ax = b\} = \emptyset \tag{30}$$

If and only if $\exists \mu \in \mathbb{E}^n$ such that $A^*\mu = 0$ and $\langle b, \mu \rangle = 1$.
How can we geometrically understand what's happening in this scenario? Fredholm's alternative states that when the set is empty, the vector b doesn't lie in the image of A. Where does μ come into this?
The vector μ provides a geometric certificate of the fact that the image of A is separate from the vector b, since we know μ is orthogonal to the image of A (from $A^*\mu = 0$) and that $\langle \mu, b \rangle = 1$.
Consider the following set, called a hyperplane, which allows us to interpret this geometric property further.

$$\mathcal{H} = \{y \in \mathbb{E}^n | \langle \mu, y \rangle = 1/2\} \tag{31}$$

This set, which has dimension $n - 1$, separates the vector b from the image of A. Since $\langle \mu, y \rangle = 1/2$ and $\langle \mu, b \rangle = 1$, we can visualize $\mathcal{H}$ as a plane with normal vector μ, located halfway between b and the image of A.
What does this hyperplane accomplish in a geometric sense? We know that $image(A)$ lies on one side of the hyperplane, and b lies on the other side of the hyperplane. We can think about this property mathematically as follows:

$$image(A) \subset \{y \in \mathbb{E}^n | \langle \mu, y \rangle < 1/2\} \tag{32}$$

$$\{b\} \subset \{y \in \mathbb{E}^n | \langle \mu, y \rangle > 1/2\} \tag{33}$$

Through these two sets, we observe how the hyperplane $\mathcal{H}$ "splits up" space into two pieces! These pieces don't intersect, except for on the hyperplane itself.
This splitting allows us to gain a full appreciation for what Fredholm is actually saying. Fredholm really says that $\exists \mu$ that is normal to some hyperplane $\mathcal{H}$ that breaks space such that $image(A)$ is in one half and b is in the other.
Let's summarize the steps we took in interpreting Fredholm's alternative. First, to show that $\{x \in \mathbb{E}^d | Ax = b\} = \emptyset$, we showed that:

$$image(A) \cap \{b\} = \emptyset \tag{34}$$

We accomplished this by identifying a hyperplane that separates $image(A)$ and b.
This two-step process suggests a more general strategy for certifying that a given set is empty! Let's work out the steps of this strategy.
Suppose we're given a set $C \subset \mathbb{E}^d$ that we'd like to certify is empty. We may follow these two steps:

1. Identify $C_1, C_2 \subset \mathbb{E}^d$ such that $C_1 \cap C_2 = \emptyset$ implies $C = \emptyset$.
2. Attempt to certify that $C_1 \cap C_2 = \emptyset$ by identifying a hyperplane that separates C_1 and C_2.

In summary, once we have two disjoint sets such that $C_1 \cap C_2 = \emptyset \Rightarrow C = \emptyset$, we want to find a way to certify that $C_1 \cap C_2$ is indeed empty. A convenient way to do this is by finding a hyperplane that separates the two sets.
As a brief aside, we ask the question - why use a hyperplane? In continuous optimization, hyperplanes are particularly convenient because of their parameterization. Every hyperplane is characterized by a normal vector μ and some "offset." This simple parameterization makes it computationally simpler to search over the space of hyperplanes to find one that separates our sets.
Thus far, we've only introduced hyperplanes in a somewhat informal manner. Now, we formalize our terminiology.

Definition 1 ***Hyperplane***
The set:

$$\mathcal{H} = \{z \in \mathbb{E}^p |\, \langle a, z \rangle = c\} \tag{35}$$

For $a \in \mathbb{E}^p \backslash \{0\}$ and $c \in \mathbb{R}$ is called a hyperplane. The vector a is called the ***normal vector*** *and the scalar c is called the* ***translate****.*

Notice that if $c = 0$, the hyperplane would pass through the origin instead of having some nonzero offset from the origin. Associated with every hyperplane is a set of spaces called the open and closed halfspaces.

Definition 2 ***Open Halfspaces***
Let $\mathcal{H}$ be a hyperplane. The open halfspaces defined by $\mathcal{H}$ are the sets:

$$\{z \in \mathbb{E}^p |\, \langle a, z \rangle < c\} \tag{36}$$
$$\{z \in \mathbb{E}^p |\, \langle a, z \rangle > c\} \tag{37}$$

The word "open" in the name open halfspace refers to the far that the inequalities in the set definition are strict.

Definition 3 ***Closed Halfspaces***
Let $\mathcal{H}$ be a hyperplane. The closed halfspaces defined by $\mathcal{H}$ are the sets:

$$\{z \in \mathbb{E}^p |\, \langle a, z \rangle \leq c\} \tag{38}$$
$$\{z \in \mathbb{E}^p |\, \langle a, z \rangle \geq c\} \tag{39}$$

The closed halfspaces defined by a hyperplane are the exact same as open halfspaces, only having $\leq$ and $\geq$ in the place of the strict inequalities $<$ and $>$. Using the language of halfspaces, we may more precisely state what it means for a hyperplane $\mathcal{H}$ to "separate" two sets.

Definition 4 ***Hyperplane Separation***
A hyperplane $\mathcal{H}$ separates two sets $C_1, C_2 \subset \mathbb{E}^p$ if C_1 and C_2 lie in opposite closed halfspaces defined by $\mathcal{H}$.

Thus, if C_1 is contained in one halfspace defined by $\mathcal{H}$ and C_2 is contained in the other, $\mathcal{H}$ is said to separate C_1 and C_2.
After introducing these concepts formally, a few questions naturally arise. Namely: given any two sets $C_1, C_2 \subset \mathbb{E}^n$ such that $C_1 \cap C_2 = \emptyset$, is there a hyperplane that separates C_1 and C_2?
We can come up with a few simple counterexamples to this question to see that the answer is no! For instance, take C_1 and C_2 to be concentric circles of radii 1 and 2. These sets are disjoint, yet there is no hyperplane that separates them!
For what sets do separating hyperplanes always exist? Let's think about a scenario that might cause a problem for the existence of a hyperplane to answer this question.
Let C_1, C_2 be disjoint sets in $\mathbb{E}^p$, and let $a, b \in C_1$. Now, consider the line segment that connects a and b. If this line segment passes through the set C_2, then no separating hyperplane can exist between the two sets! This is where the problem arises.
Mathematically, we can think about this problem as follows. Suppose we have a hyperplane that separates space into two halfspaces. A line segment connecting two points in one halfspace must remain on that side of the hyperplane! Thus, we can think about this "line segment" condition as giving us what we need to conclude the existence of a separating hyperplane.
Let's describe the line segment condition more formally. What we want is for the sets $C_1, C_2 \subset \mathbb{E}^p$ to be "closed under line segments" between pairs of points. We now define a set that satisfies this "closed under line segments" condition.

Definition 5 ***Convex Set***
A set $C \subset \mathbb{E}^p$ is called convex if $x, y \in C$ implies:

$$\lambda x + (1-\lambda)y \in C \ \forall \lambda \in [0,1] \tag{40}$$

The expression $\lambda x + (1-\lambda)y \in C, \lambda \in [0,1]$ is called a **convex combination** of x and y. Let's think of a couple examples of convex and nonconvex sets. An example of a convex set would be an open ball, as it is closed under line

segments. A nonconvex set would be a hollow ball.
Now, let's think about some operations on convex sets that preserve convexity.

1. Let $C \subset \mathbb{E}^d$ and $D \subset \mathbb{E}^n$ be convex sets. Then, the Cartesian product of C and D, defined:

$$C \times D = \{(c, d) | c \in C, d \in D\} \tag{41}$$

   Is a convex set.

2. Let $\{C^{(i)}\}_{i \in \mathcal{I}}$ be any family of sets where $C^{(i)} \subset \mathbb{E}^d$ is convex. The intersection:

$$\bigcap_{i \in \mathcal{I}} C^{(i)} \tag{42}$$

   Is also convex in $\mathbb{E}^d$. Note that the index set $\mathcal{I}$ can be finite, countable infinite, or uncountably infinite! Note that a consequence of this result is that the empty set $\emptyset$ is convex.

3. Let $A : \mathbb{E}^d \to \mathbb{E}^n$ be an affine map, and $C \subset \mathbb{E}^d$ be convex. Then, the image of C under A, defined:

$$A(C) = \{A(x) | x \in C\} \tag{43}$$

   Is a convex set in $\mathbb{E}^n$.

Now that we've established the definition and basic properties of a convex set, we must think back to the relation between convexity and the existence of a separating hyperplane.
Given two sets $C_1, C_2 \subset \mathbb{E}^p$ that are convex and disjoint, can we put a hyperplane between them? Interestingly, the answer to this question is yes, subject to C_1 and C_2 belonging to opposite closed halfspaces. We formalize this idea in the separation theorem.

Theorem 1 ***Separation Theorem***
Let $C_1, C_2 \subset \mathbb{E}^p$ be convex sets such that $C_1 \cap C_2 = \emptyset$. Then, there exists a hyperplane:

$$\mathcal{H} = \{z \in \mathbb{E}^p | \langle a, z \rangle = c\} \tag{44}$$

With $a \in \mathbb{E}^p \backslash \{0\}$ and $c \in \mathbb{R}$ such that C_1 and C_2 lie in the opposite closed halfspaces defined by $\mathcal{H}$. In particular:

$$\inf_{z^{(1)} \in C_1} \langle a, z^{(1)} \rangle \geq c \geq \sup_{z^{(2)} \in C_2} \langle a, z^{(2)} \rangle \tag{45}$$

Where c is the translate of $\mathcal{H}$.

Before we proceed, we make a few remarks on this theorem. First, note that although $C_1 \cap C_2 = \emptyset$ implies the existence of a separating hyperplane, the reverse direction requires us to be a little bit more careful!
If a separating hyperplane exists, it's not necessarily true that C_1 and C_2 are disjoint. However, this is the case when one of C_1, C_2 belongs to an open halfspace. This subtlety will be discussed in further detail when we apply this theorem to optimization problems.
We now define two more common sets based on the definition of a hyperplane.

Definition 6 ***Polyhedron***
A polyhedron is a finite intersection of closed halfspaces.

Definition 7 ***Cone***
A set $C \subset \mathbb{E}^p$ that is closed under nonnegative scaling is called a cone. Mathematically, if C is a cone:

$$x \in C \Rightarrow ax \in C \ \forall a \geq 0 \tag{46}$$

If C is also convex, C is called a convex cone.

Note that not every cone is convex! For instance, consider a common cone in three dimensions. The set just composed of the surface of the cone is a cone in the mathematical sense, but it not convex. The set composed of the surface and interior of the cone, however, is convex, and thus forms a convex cone.

4 Lecture 4: Convex Functions

In today's lecture, we will develop a method for determining if a function has "convexity structure," based on the notions of convexity we introduced last time. Our main focus in this lecture will therefore be the study of **convex functions**. We'll make the transition from studying convex sets to convex functions through developing a *mapping* between sets and functions! With this mapping, we'll be able to translate any ideas about convex sets into ideas about convex functions.

Definition 8 ***Epigraph of a Function***
The epigraph of a function $f : \mathbb{E}^d \to \mathbb{R}$ with domain $dom(f)$ is denoted $epi(f)$, and is defined as:

$$epi(f) = \{(x,t) \in \mathbb{E}^d \times \mathbb{R} | x \in dom(f),\ f(x) \leq t\} \tag{47}$$

Intuitively, we can reason about the epigraph of a function as follows. It is the set of all pairs of (x, t) values such that t is greater than $f(x)$. This makes the epigraph the set of all values "above" the graph of the function. Note that this concept, by construction, is well-defined for any function.
The epigraph leads us nicely to the definition of a convex function.

Definition 9 ***Convex Function***
A function $f : \mathbb{E}^d \to \mathbb{R}$ with domain $dom(f)$ is called convex if $epi(f)$ is a convex subset of $\mathbb{E}^d \times \mathbb{R}$.

Thus, we can conclude convexity of a *function* by looking at convexity of a set associated with the function. Note that in our definition, we assume that our functions are not able to achieve values of $\pm\infty$.
It's important to note that convexity does *not* imply differentiability! Convex functions, such as $f(x) = |x|, x \in \mathbb{R}$, *can* have corners, and will not necessarily be everywhere-differentiable.

Proposition 2 ***Inequality Characterization of Convexity***
A function $f : \mathbb{E}^d \to \mathbb{R}$ with domain $dom(f)$ is convex if and only if $dom(f) \subset \mathbb{E}^d$ is convex and $\forall x, y \in dom(f)$, we have that:

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y) \; \forall \lambda \in [0, 1] \tag{48}$$

This proposition provides us with an alternate way of characterizing the convexity of functions. It states that if, for all $x, y \in \mathbb{E}^d$, f(line segment connecting x, y) is $\leq$ the line segment in $\mathbb{R}$ connecting $f(x)$ and $f(y)$, then the function is convex. In $\mathbb{R}$, we can visualize this as the line segment between two points on the function always being above the graph of the function in-between those two points.
It's important to note that convexity is a *global* condition on a function, rather than a local one! Additionally, we *must* have a convex domain in order to apply the definition of convexity to a function. Of further interest, note that the convexity of a set implies that the set must be simply connected.
As with convexity of sets, there are a number of operations that preserve the convexity of functions!

1. Let $f_1, f_2 : \mathbb{E}^d \to \mathbb{R}$ be convex functions with domains $dom(f_1) = dom(f_2) = \mathbb{E}^d$. For any scalars $\alpha_1, \alpha_2 \geq 0$, the function:

$$f(x) = \alpha_1 f_1(x) + \alpha_2 f_2(x) \tag{49}$$

   Is a convex function with $dom(f) = \mathbb{E}^d$.

2. Let $g : \mathbb{E}^d \to \mathbb{R}$ with $dom(g) = \mathbb{E}^d$ be a convex function and $A : \mathbb{E}^n \to \mathbb{E}^d$ be an affine map. The function:

$$f(x) = g(A(x)) \tag{50}$$

With domain $dom(f) = \mathbb{E}^n$ is a convex function.

3. Let $\{f^{(i)}\}_{i\in\mathcal{I}}$ be any collection of convex functions with $f^{(i)}\mathbb{E}^d \to \mathbb{R}$ having domain $dom(f^{(i)}$. The pointwise supremum, defined as:

$$f(x) = \sup_{i\in\mathcal{I}} f^{(i)}(x) \tag{51}$$

With domain $dom(f) = \bigcap_{i\in\mathcal{I}} dom(f^{(i)}$ is a convex function. Note that this property can be viewed as a parallel for convex functions of the intersection property of convex sets. There are two methods to approach the proof of this statement - by applying the inequality definition of convexity or by applying the epigraph definition and using the intersection property of convex sets.

Note that these observations also hold for more general convex functions. Using these properties, we can determine the convexity of more complex functions using the convexity of simpler functions.
Let's now turn our attention to differentiability. If f is differentiable, there are a number of additional techniques we can make use of to analyze its convexity. Differentiability generally gives us more refined conditions for determining the convexity of functions.

Proposition 3 ***Convexity for Differentiable Functions***
Let $f : \mathbb{E}^d \to \mathbb{R}$ be a function with domain $dom(f)$ being open and f being differentiable on $dom(f)$. f is convex if and only if:

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle \;\; \forall x, y \in dom(f) \tag{52}$$

We may think about the proposed inequality more intuitively by fixing x and checking that the inequality holds for all $y \in dom(f)$. If we can accomplish this for all $x \in dom(f)$, we can conclude convexity of f.
Let's think about the right hand side of the inequality. For each fixed x, we know that:

$$f(x) + \langle \nabla f(x), y - x \rangle \tag{53}$$

Is an *affine function* of y. In particular, it gives the first order Taylor approximation of f at each point x! We can therefore think of this statement as saying:

if the graph of the function is above the tangent line to the function at each x, then the function is convex.

Proposition 4 ***Hessian of a Convex Function*** *Let $f : \mathbb{E}^d \to \mathbb{R}$ be a twice-differentiable function with open domain $dom(f)$. f is convex if and only if the Hessian $\nabla^2 f$ satisfies:*

$$\nabla^2(f(x)) \succeq 0 \ \forall x \in dom(f) \tag{54}$$

For f is twice differentiable, recall that $\nabla^2 f$ is the symmetric matrix in $\mathbb{R}^{d\times d}$ filled with the partial derivatives of f. Because it is symmetric, talking about its positive definiteness is something that makes sense.
Let's come up with a geometric interpretation of this proposition. We know that the Hessian tells us about the local curvature of the function, and appears in the second order approximation of the function.
This condition tells us that the second order approximation of a convex function must "curve upwards" at every point.
We are now ready to introduce the formulation of a general convex optimization optimization problem. Depending on the source, each of the following types of optimization problems are called convex. First:

$$v = \inf_{x\in\mathbb{E}^d} f(x) \tag{55}$$

$$s.t. \ x \in S \tag{56}$$

For a convex function $f : \mathbb{E}^d \to \mathbb{R}$ and a convex set $S \subset \mathbb{E}^d$. The other common form of a convex optimization is:

$$v = \inf_{x\in\mathbb{E}^d} f_0(x) \tag{57}$$

$$s.t. \ f_i(x) \leq 0, i = 1, ...k \tag{58}$$

$$g_j(x) = 0, j = 1, ..., m \tag{59}$$

For convex functions $f_0, ..., f_k : \mathbb{E}^d \to \mathbb{R}$ and affine functions $g_1, ..., g_m : \mathbb{E}^d \to \mathbb{R}$.
Note that the f_i constraints are known as **inequality constraints**, while the g_j constraints are known as **equality constraints**.
Compared to the first formulation of a convex optimization problem, we have simply elected to describe the constraint set more explicitly. Additionally, the second formulation is more commonly used in the sub-field of continuous optimization known as nonlinear programming. Note, however, that the two formulations are identical.

5 Lecture 5: Fenchel Duality I

In this lecture, we will focus on deriving certificates based on separating hyperplanes. Additionally, we'll come up with a "principled" way of "adding and subtracting" to produce verifiable certificates on our optimization problems, just as we did for the simple projection onto a subspace problem.
First, we'll introduce some new terminology to formalize some concepts we've discussed thus far. First, let's discuss sets of certificates.
A family of certificates for deriving lower bounds on the optimal value of an optimization problem is called a **duality scheme**. The problem of identifying the "best" (largest) lower bound that can be obtained via a duality scheme is called the **dual problem** associated with the given duality scheme. Note that the dual problem is itself an optimization problem.
While this alternate lower bound optimization is called the dual problem, the original optimization problem:

$$v = \inf_{x \in \mathbb{E}^d} f(x) \tag{60}$$

$$s.t.\ x \in S \tag{61}$$

For which we wish to obtain certifiable lower bounds on the optimal value v, is called the **primal problem**.
A duality scheme known as the **Fenchel duality scheme** is central to the study of convex optimization. This approach was popularized by Rockafellar in his text *Convex Analysis*. Let's discuss the Fenchel duality scheme. Consider the following primal optimization problem:

$$v = \inf_{x \in \mathbb{E}^d} f(x) \tag{62}$$

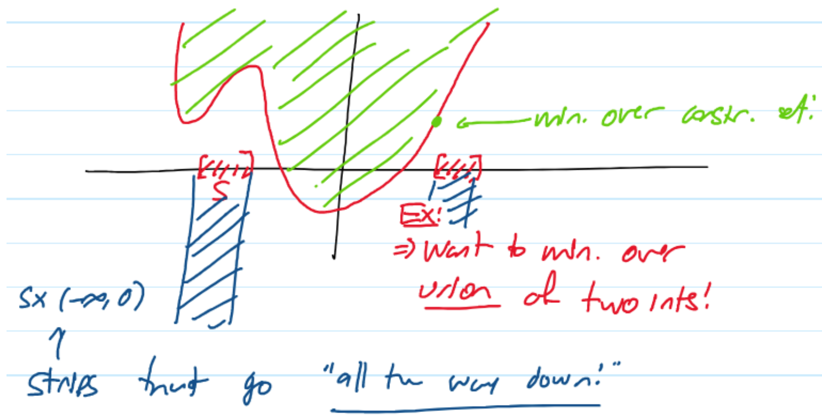
$$s.t.\ x \in S \tag{63}$$

Here, $f : \mathbb{E}^d \to \mathbb{R}$ with domain $dom(f)$ is the objective, and $S \subset \mathbb{E}^d$ is the constraint set. For now, note that this problem is general, and need not be convex.
The Fenchel duality scheme proceeds in two steps:

1. Certify that $\gamma \leq v \in \mathbb{R}$ by certifying that:

$$epi(f - \gamma) \cap S \times (-\infty, 0) = \emptyset \tag{64}$$

   Note that here, we are focusing on two sets in $d + 1$ dimensions! The dimension we focus on is therefore *not* the same as the dimension of the original problem!



Thinking back to our earlier discussion of empty sets, it seems reasonable that the process of certifying the intersection of two sets is empty can be completed by finding a separating hyperplane. This takes us to step 2.

2. Certify that $epi(f-\gamma) \cap S \times (-\infty, 0) = \emptyset$ by separating the sets using a *non-vertical* hyperplane. In the context of multiple dimensions, a non-vertical hyperplane is a hyperplane that is not parallel to the $d+1$'st axis. Note that this hyperplane will not always exist! We can only apply this procedure for certain types of optimization problems.

Notice that the first step of the Fenchel duality scheme is simply to reformulate the optimization problem at hand, and that the second step - finding a separating hyperplane - is where we *actually* do all of the work. Also of importance is that the Fenchel duality scheme *only* provides us with a method of certifying optimal values, *not* optimal solutions!
Let's reframe the Fenchel duality problem in a more theoretical light. Formally, we wish to identify a tuple $(\lambda, \eta, \delta) \in \mathbb{E}^d \times \mathbb{R} \times \mathbb{R}$ such that:

$$epi(f-\gamma) \subseteq \{(x,t) \in \mathbb{E}^d \times \mathbb{R} | \, \langle(\lambda,\eta),(x,t)\rangle \leq \delta\} \tag{65}$$

$$S \times (-\infty, 0) \subseteq \{(x,t) \in \mathbb{E}^d \times \mathbb{R} | \, \langle(\lambda,\eta),(x,t)\rangle > \delta\} \tag{66}$$

Where $\eta \neq 0$ to enforce that the hyperplane defined by the normal vector (λ, η) is not parallel to the $d+1$'st axis of $\mathbb{E}^{d+1}$. The two constraints above are the *same* as saying that $epi(f-\gamma)$ and $S \times (-\infty, 0)$ lie on opposite sides of a hyperplane! Note that one halfspace is closed and the other is open to avoid intersections between the two sets.
The Fenchel dual optimization problem may now be written as the following:

$$v^* = \sup_{\lambda \in \mathbb{E}^d, \eta, \delta, \gamma \in \mathbb{R}} \gamma \tag{67}$$

$$s.t. \; epi(f-\gamma) \subseteq \{(x,t) \in \mathbb{E}^d \times \mathbb{R} | \, \langle(\lambda,\eta),(x,t)\rangle \leq \delta\} \tag{68}$$

$$S \times (-\infty, 0) \subseteq \{(x,t) \in \mathbb{E}^d \times \mathbb{R} | \, \langle(\lambda,\eta),(x,t)\rangle > \delta\} \tag{69}$$

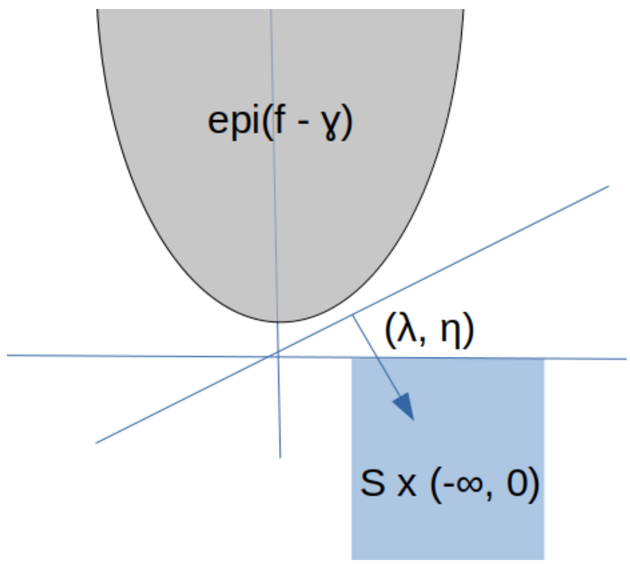
Thus, we want to find the largest amount, γ, that we can "push the function down" such that the epigraph of $f - \gamma$ has no intersection with $S \times (\infty, 0)$. Notice that we do not see x, t as decision variables in this problem, as we want our lower bound on the optimal value of the primal problem to be global, rather than hold for a single value of x.

Let's examine this optimization problem in greater detail. Suppose we have a (λ, η, δ) tuple that satisfies the constraints on the dual problem. Looking at the form of the constraints, we notice that *any* positive scaling of (λ, η, δ) will still satisfy the constraints! What does this tell us about the possible values of η?

If $(\tilde{\lambda}, \tilde{\eta}, \tilde{\delta})$ is feasible for the Fenchel dual problem, then *so is* any positive scaling of the three! Thus, we can set $\eta = \pm 1$ without loss of generality.

Let's first consider the case of $\eta = 1$, and see how this interacts with the constraints. In $S \times (-\infty, 0)$, we know that t can be arbitrarily small. if $\eta = 1$, will we ever get a δ such that $\langle (\lambda, \eta), (x, t) \rangle > \delta$? Since t can be arbitrarily small, we *cannot* get such a δ. For $\eta = 1$, the same problem happens for the other constraint set.

We can also visualize this with the following graphic:



In the above, we see that as we increase γ, such that $epi(f - \gamma)$ and $S \times (-\infty, 0)$ get closer together, the only separating hyperplane that has $epi(f - \gamma)$ in the negative halfspace $(... \leq \delta)$and $S \times (-\infty, 0)$ in the positive halfspace $(... > \delta)$, requires that η points downwards. This tells us that η cannot equal $+1$ and separate the two spaces as we raise γ, as the hyperplane normal vector must point down.

Thus, the two halfspace constraints tell us that $\eta = -1$. This gives us the following simplification to the Fenchel dual problem:

$$v_F^* = \sup_{\lambda \in \mathbb{E}^d, \delta, \gamma \in \mathbb{R}} \gamma \tag{70}$$

$$s.t.\ epi(f-\gamma) \subseteq \{(x,t) \in \mathbb{E}^d \times \mathbb{R} |\ \langle \lambda, x\rangle - t \leq \delta\} \tag{71}$$

$$S \times (-\infty, 0) \subseteq \{(x,t) \in \mathbb{E}^d \times \mathbb{R} |\ \langle \lambda, x\rangle - t > \delta\} \tag{72}$$

Note that we have added a subscript F to the optimal solution to the Fenchel dual. Before we continue, it's important to make a few remarks regarding the dual problem. Note that we will *always* have that:

$$v_F^* \leq v \tag{73}$$

Where v is the solution to the primal problem. This is a fact that holds true for *any* duality scheme, not just Fenchel! Schemes with this inequality are said to exhibit **weak duality**.
For a given primal problem, if the associated Fenchel dual problem satisfies $v_F^* = v$, then the Fenchel dual is said to exhibit **strong duality**. Under what conditions will the Fenchel dual exhibit this type of behavior?

Theorem 2 ***Strong Duality of the Fenchel Duality Scheme***
Consider a primal optimization problem:

$$v = \inf_{x \in \mathbb{E}^d} f(x) \tag{74}$$

$$s.t.\ x \in S \tag{75}$$

With $f : \mathbb{E}^d \to \mathbb{R}$ having domain $dom(f) \subset \mathbb{E}^d$ and constraint set $S \subset \mathbb{E}^d$. Suppose we have that:

1. *f is convex*
2. *S is convex*
3. *$int(dom(f)) \cap S \neq \emptyset$*

The, the Fenchel dual problem is a strong dual, i.e. $v_F^* = v$. Moreover, there exists an optimal solution of the Fenchel dual problem.

This tells us that if the three conditions specified above are satisfied, the sup in the statement of the Fenchel dual problem may be replaced with a max. It's also important to note that the dual may have an optimal solution even if the primal doesn't! A simple example of this is the primal problem $\inf_{x \in \mathbb{E}^d} \exp(x)$, which is a convex problem with no optimal solution.

Proof: To prove this theorem, we must show there exists a hyperplane that separates $epi(f-v)$ and $S \times (-\infty, 0)$. As $epi(f-v)$ and $S \times (-\infty, 0)$ are both convex and have an empty intersection, we have by the separation theorem for convex sets that $\exists(\overline{\lambda}, \overline{\eta}) \in \mathbb{E}^d \times \mathbb{R}/\{0\}$ and $\overline{\delta} \in \mathbb{R}$ such that:

$$epi(f-v) \subseteq \{(x,t) \in \mathbb{E}^d \times \mathbb{R} | \, \langle(\overline{\lambda}, \overline{\eta}), (x,t)\rangle \leq \overline{\delta}\} \tag{76}$$

$$S \times (-\infty, 0) \subseteq \{(x,t) \in \mathbb{E}^d \times \mathbb{R} | \, \langle(\overline{\lambda}, \overline{\eta}), (x,t)\rangle \geq \overline{\delta}\} \tag{77}$$

Remember - the separation theorem just gives us *closed* halfspaces! We must now show one of the inequalities in the halfspace sets above is strict and that the hyperplane $\mathcal{H}$ from above is nonvertical.

First, we'll prove the nonvertical property by showing that $\overline{\eta} \neq 0$. Suppose for contradiction that $\overline{\eta} = 0$. We must then have that $\overline{\lambda} \neq 0$, since it is required that $(\overline{\lambda}, \overline{\eta}) \neq 0$.

If $\overline{\eta} = 0$, then the optimization constraint associated with the bottom set "ignores t" and only deals with S! For the top set, if we "ignore t," we just get $dom(f)$ as the possible values of X. Thus, we would require:

$$dom(f) \subseteq \{x | \, \langle\overline{\lambda}, x\rangle \leq \overline{\delta}\} \tag{78}$$

$$S \subseteq \{x | \, \langle\overline{\lambda}, x\rangle \geq \overline{\delta}\} \tag{79}$$

This tells us that $dom(f)$ and S are in opposite closed halfspaces, which contradicts our assumption:

$$int(dom(f)) \cap S \neq \emptyset \tag{80}$$

Thus, we conclude that $\overline{\eta} \neq 0$. This proves the first part!

Now, we turn our attention to establishing a strict inequality in one of the hyperplanes. If $\overline{\eta} \neq 0$, we may go through the same λ, η, δ "scaling" reasoning as before to establish that $\overline{\eta} = -1$ without loss of generality. For $\overline{\eta} = -1$:

$$epi(f-v) \subseteq \{(x,t) | \, \langle\overline{\lambda}, x\rangle - t \leq \overline{\delta}\} \tag{81}$$

$$S \times (-\infty, 0) \subseteq \{(x,t) | \, \langle\overline{\lambda}, x\rangle - t \geq \overline{\delta}\} \tag{82}$$

Now, as the second constraint holds for $x \in S$ and $t \in (-\infty, 0)$, which is an open interval, we can replace the closed halfspace in the second constraint with an open halfspace:

$$S \times (-\infty, 0) \subseteq \{(x,t) | \, \langle\overline{\lambda}, x\rangle - t > \overline{\delta}\} \tag{83}$$

From the fact that t strictly satisfies $-\infty < t < 0$. This completes the proof! $\square$

6 Lecture 6: Fenchel Duality II

Today, we will derive a simplification of Fenchel dual problems. Our aim here is to *derive* the Fenchel dual more easily.

First, we recall our problem setup. Consider the optimization:

$$v = \inf_{x \in \mathbb{E}^d} f(x) \tag{84}$$

$$s.t.\ x \in S \tag{85}$$

Where $f : \mathbb{E}^d \to \mathbb{R}$ has domain $dom(f) \subset \mathbb{E}^d$ and constraint set $S \subset \mathbb{E}^d$. Again, we will start by assuming this problem is not necessarily convex.
Last time, we derived the Fenchel dual problem as follows:

$$v_F^* = \sup_{\lambda \in \mathbb{E}^d, \delta, \gamma \in \mathbb{R}} \gamma \tag{86}$$

$$s.t.\ epi(f - \gamma) \subset \{(x,t) \in \mathbb{E}^d \times \mathbb{R} | \langle \lambda, x \rangle - t \leq \delta\} \tag{87}$$

$$S \times (-\infty, 0) \subset \{(x,t) \in \mathbb{E}^d \times \mathbb{R} | \langle \lambda, x \rangle - t > \delta\} \tag{88}$$

Remember, there is no variable x in this optimization problem - this is an important general feature of dual problems!
Let's analyze each of the constraints. First, we examine:

$$epi(f - \gamma) \subset \{(x,t) \in \mathbb{E}^d \times \mathbb{R} | \langle \lambda, x \rangle - t \leq \delta\} \tag{89}$$

Any (x,t) pair in the epigraph lies in this closed halfspace. Here, the "worst-case" choice of t is the *smallest* value of t. This "worst case" idea tells us that this constraint is equivalent to:

$$\langle \lambda, x \rangle - (f(x) - \gamma) \leq \delta \ \forall x \in dom(f) \tag{90}$$

Where we substitute in $f(x) - \gamma$ for our "worst-case" t value. This constraint is equivalent to the following:

$$\langle \lambda, x \rangle - f(x) \leq \delta - \gamma \ \forall x \in dom(f) \tag{91}$$

Since this inequality hold for all x, we conclude that the supremum over x of the left hand side of the constraint must be less than or equal to $\delta - \gamma$. For a given λ, this supremum has a particular name.

Definition 10 ***Conjugate Function***
Let $f : \mathbb{E}^d \to \mathbb{R}$ be any function with domain $dom(f) \subset \mathbb{E}^d$. The conjugate of f is denoted f^*, and is defined:

$$f^*(\lambda) = \sup_{x \in dom(f)} \langle \lambda, x \rangle - f(x) \tag{92}$$

Here, $f^* : \mathbb{E}^d \to \mathbb{R}$ and:

$$dom(f^*) = \{\lambda \in \mathbb{E}^d | \sup_{x \in dom(f)} \langle \lambda, x \rangle - f(x) < \infty\} \tag{93}$$

Interestingly, this function is related to the Legendre transformation, which is used in Hamiltonian dynamics.
Before we get into the properties of the conjugate function, let's make note of a couple of things about the definition.
First, we may rewrite f^* in terms of the epigraph of f:

$$\sup_{x \in dom(f)} \langle \lambda, x \rangle - f(x) = \sup_{(x,t) \in epi(f)} \langle (\lambda, -1), (x, t) \rangle \tag{94}$$

Rewriting the conjugate in this manner "lifts it up" a dimension. From this alternate definition, we see that f^* is a linear function in (x, t).
Importantly, we note that f^* is convex function, even if f is not! Looking at the function inside the supremum:

$$\langle \lambda, x \rangle - f(x) \tag{95}$$

We notice that this function is affine in λ. When we take its supremum over x, we're essentially taking its "maximum" over a collection of affine functions. Since the maximum of a set of convex functions is convex, f^* is convex.
Now, we turn our attention back to the constraints on the Fenchel dual problem. In terms of the conjugate function, the first constraint in the Fenchel dual states:

$$f^*(\lambda) \leq \delta - \gamma \tag{96}$$

Now, we rewrite the second constraint, which says:

$$\langle \lambda, x \rangle > \delta + t \; \forall x \in S, t \in (-\infty, 0) \tag{97}$$

This is the same as saying:

$$\langle -\lambda, x \rangle \leq -\delta \; \forall x \in S \tag{98}$$

Now, we define another concept that will assist us in simplifying the Fenchel dual.

Definition 11 ***Support Function***
Let $S \subset \mathbb{E}^d$ be any set. The support function associated to S is denoted h_s, and is defined as:

$$h_s(\mu) = \sup_{x \in S} \langle \mu, x \rangle \tag{99}$$

Here, $h_s : \mathbb{E}^d \to \mathbb{R}$, and:

$$dom(h_s) = \{\mu \in \mathbb{E}^d | \sup_{x \in \mathbb{S}} \langle \mu, x \rangle < \infty\} \tag{100}$$

Looking at this definition, we see that the Fenchel dual may be rewritten using both conjugate and support functions.
Notice that h_s is a convex function, even if S is a nonconvex set! In the definition of h_s, we are taking a supremum over a set of convex functions, and thus get a convex function.
In terms of the support function, the second constraint of the Fenchel dual can now be stated:

$$h_s(-\lambda) \leq -\delta \tag{101}$$

Putting all of these components together, the Fenchel dual problem can be restated as:

$$v_F^* = \sup_{\lambda \in \mathbb{E}^d, \delta, \gamma \in \mathbb{R}} \gamma \tag{102}$$

$$s.t. \ f^*(\lambda) \leq \delta - \gamma \tag{103}$$

$$h_s(-\lambda) \leq -\delta \tag{104}$$

Eliminating γ, we then have that:

$$v_F^* = \sup_{\lambda \in \mathbb{E}^d, \delta \in \mathbb{R}} \delta - f^*(\lambda) \tag{105}$$

$$s.t. \ h_s(-\lambda) \leq \delta \tag{106}$$

We can now also eliminate δ from the problem. This gives us:

Definition 12 ***Fenchel Dual Simplification***
The Fenchel dual of an optimization problem $v = \inf_{x \in S} f(x)$ may be written:

$$v_F^* = \sup_{\lambda \in \mathbb{E}^d} -f^*(\lambda) - h_s(-\lambda) \tag{107}$$

This final simplification has the advantage that we no longer need to rely on hyperplanes to compute the Fenchel dual of a problem! This formulation allows for a significantly simpler computation of the dual, and is convenient for deriving Fenchel dual problems in practice. Notice that no matter *what* the original primal problem is, the Fenchel dual problem is a convex optimization problem.
This reformulation also provides another interpretation of weak duality of the Fenchel dual. Consider any $x \in S \cap dom(f)$. We have that:

$$f(x) = f(x) - \langle \lambda, x \rangle + \langle \lambda, x \rangle \tag{108}$$

The firs two terms are $\geq f^*(\lambda)$, and the third term is $\geq -h_s(-\lambda)$. So, $-f^*(\lambda) - h_s(-\lambda)$ gives a lower bound on v for each fixed λ!

Now, let's turn our attention back to the convex case. Consider a convex optimization problem:

$$v = \inf_{x \in \mathbb{E}^d} f(x) \tag{109}$$

$$s.t.\ x \in S \tag{110}$$

Here, $f : \mathbb{E}^d \to \mathbb{R}$ with domain $dom(f)$ is a convex function and S is a convex set. Further, suppose $int(dom(f)) \cap S \neq \emptyset$. We know that under these conditions, the Fenchel dual exhibits strong duality and has an optimal solution for $\hat{\lambda}$.
Now, suppose further that the primal problem has an optimal solution $\hat{x}$. We have that:

$$f(\hat{x}) = f(\hat{x}) - \langle \hat{\lambda}, \hat{x} \rangle + \langle \hat{\lambda}, \hat{x} \rangle \tag{111}$$

$$\geq -f^*(\hat{\lambda}) - h_s(-\hat{\lambda}) \tag{112}$$

$$= v_F^* \tag{113}$$

Where in the last step, we use the optimality of $\hat{\lambda}$. We know the under these constraints $v = v_F^*$, and that the inequality above is an equality.
As $v_F^* = f(\hat{x})$, we conclude the following two facts:

$$f^*(\hat{\lambda}) = \langle \hat{\lambda}, \hat{x} \rangle - f(\hat{x}) \tag{114}$$

$$h_s(-\hat{\lambda}) = \langle -\hat{\lambda}, \hat{x} \rangle \tag{115}$$

Let's analyze each of these conditions. First:

$$\langle \hat{\lambda}, \hat{x} \rangle - f(\hat{x}) \geq \langle \lambda, x \rangle - f(x)\ \forall x \in dom(f) \tag{116}$$

Stated differently, we have that:

$$f(x) \geq f(\hat{x}) + \langle \hat{\lambda}, x - \hat{x} \rangle \ \ x \in dom(f) \tag{117}$$

This statement looks similar to the first order optimality conditions for f, where instead of a λ, we had a gradient. The condition above allows us to use a similar first order optimality condition to this *but* without the requirement of differentiability!

Definition 13 ***Subdifferential***
Let $f : \mathbb{E}^d \to \mathbb{R}$ be convex with $int(dom(f)) \neq \emptyset$. For all $x \in int(dom(f))$, the subdifferential of f at x is denoted $\partial f(x)$ and is defined as:

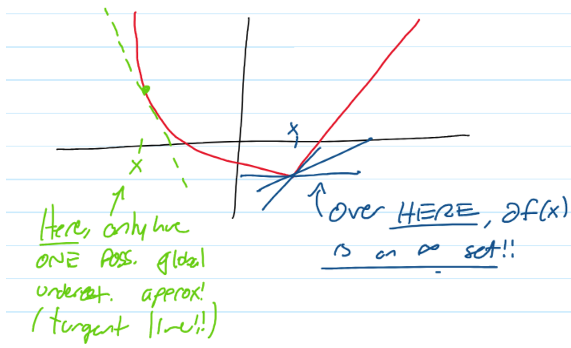
$$\partial f(x) = \{\mu \in \mathbb{E}^d | f(y) \geq f(x) + \langle \mu, y - x \rangle \ \ \forall y \in dom(f)\} \tag{118}$$

The subdifferential is a generalization of the notion of "gradient" to arbitrary *convex* functions. Note that convexity is required to apply this definition.

It's important to note that for each fixed x, $\partial f(x)$ gives a *collection* of μ satisfying the property above, not necessarily just a single μ!
At any differentiable point, however, this collection will only contain one element, corresponding to the gradient line to the function at that point. Thus, if f is differentiable at x:

$$\partial f(x) = \nabla f(x) \tag{119}$$

At any non-differentiable point, we will not get this uniqueness.



Because of its close ties to the gradient, the subdifferential obeys many of the standard rules for gradients. For instance:

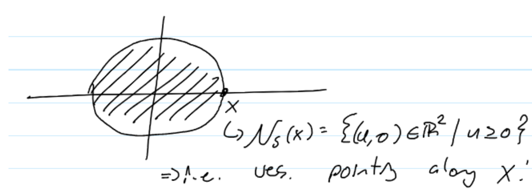
$$\partial(f_1 + f_2)(x) = \partial f_1(x) + \partial f_2(x) \tag{120}$$

Where the + sign above denotes the Minkowski sum of sets.
Now, we state another definition, which together with the subdifferential will help us describe convex optimization problems.

Definition 14 ***Normal Cone***
Let $S \subset \mathbb{E}^d$ be a convex set. For any $x \in S$, the normal cone is denoted $\mathcal{N}_s(x)$ and is defined as:

$$\mathcal{N}_s(x) = \{\mu \in \mathbb{E}^d | \langle \mu, y - x \rangle \leq 0 \; \forall y \in S\} \tag{121}$$

From this definition, we observe that $\mathcal{N}_s(x)$ is the set of linear functionals that attain their maximum over S at x.
Consider the following simple example:



From its definition, we notice that $\mathcal{N}_s(x)$ is a convex cone! Fixing x, y, we see that $\mathcal{N}_s(x)$ provides a halfspace condition. For all y, we then get an intersection of halfspaces:

$$\langle \mu, y - x \rangle \leq 0 \; \forall y \in S \tag{122}$$

Now, we put this definition together to reach a conclusion above convex optimization problems. In a convex optimization problem in which $int(dom(f)) \cap S \neq \emptyset$ for which $\hat{x}$ is optimal, there exists a $\hat{\lambda}$ such that $\hat{\lambda} \in \partial f(\hat{x})$ and $-\hat{\lambda} \in \mathcal{N}_s(\hat{x})$. Alternatively, we may express this as:

$$0 \in \partial f(\hat{x}) + \mathcal{N}_s(\hat{x}) \tag{123}$$

7 Lecture 7: Describing Convex Sets

Previously, we came up with an efficient description of the Fenchel dual, which allowed us to compute the dual problem with relative ease. However, a key issue that remains at this point in our development of optimization is that of computation.
Just because convex problems have "nice" mathematical properties *doesn't* mean that we can tractably find an optimal solution. Furthermore, we might not have a certificate of optimality that we can use to verify optimal solutions.
Let's think about the challenging nature of convex optimization problems further by thinking about the different representations of convex sets.
First, consider an arbitrary (not necessarily convex) optimization problem:

$$v = \inf_{x \in \mathbb{E}^d} f(x) \tag{124}$$

$$s.t. \; x \in S \tag{125}$$

We will show that this problem, though arbitrary in nature, can be *reformulated* as a convex optimization problem!
First, we'll rewrite the problem with a linear objective function. We may write the problem equivalently as:

$$v = \inf_{x \in \mathbb{E}^d, t \in \mathbb{R}} t \tag{126}$$

$$s.t. \; (x, t) \in epi(f) \tag{127}$$

$$x \in S \tag{128}$$

Here, we introduce an additional variable, t, to reformulate the optimization problem with a linear objective in $d + 1$ dimensions. This trick of changing to a linear optimization over t using the epigraph is aptly called the "epigraph trick." Notice that we may rewrite the set of two constraints in this formulation as the single optimization constraint:

$$(x, t) \in epi(f) \cap S \times \mathbb{R} \tag{129}$$

Next, we will rewrite this modified problem as an equivalent problem with a convex constraint. Consider the following lemma.

Lemma 1 ***Recasting Problems with a Convex Constraint***
Consider the following optimization problem with a linear objective function:

$$v = \inf_{z \in \mathbb{E}^n} \langle c, z \rangle \tag{130}$$

$$s.t.\ z \in \mathcal{T} \tag{131}$$

Here, $\mathcal{T} \subset \mathbb{E}^n$ *is a constraint set and* $c \in \mathbb{E}^n$ *defines a linear objective function. We have that:*

$$v = \inf_{z \in \mathbb{E}^n} \langle c, z \rangle \tag{132}$$

$$s.t.\ z \in conv(\mathcal{T}) \tag{133}$$

This lemma states that we may equivalently rewrite an optimization problem with a constraint set $\mathcal{T}$ as an optimization problem with a constraint set of the convex hull of $\mathcal{T}$.

Proof: As $\mathcal{T} \subset conv(\mathcal{T})$, we have that:

$$\inf_{z \in \mathbb{E}^n} \langle c, z \rangle \ \ s.t.\ z \in \mathcal{T} \tag{134}$$

Has a larger optimal value than the problem:

$$\inf_{z \in \mathbb{E}^n} \langle c, z \rangle \ \ s.t.\ z \in conv(\mathcal{T}) \tag{135}$$

This gives us one direction of an inequality. We now show the other direction of inequality holds to prove equality between the optimal values of the two optimization problems.
In the other direction, consider any $\overline{z} \in conv(\mathcal{T})$. We have by the definition of the convex hull that there exist $\overline{z}^{(1)}, ..., \overline{z}^{(k)} \in \mathcal{T}$ and $\lambda_1, ..., \lambda_k \geq 0$ such that $\sum_{i=1}^{k} \lambda_i = 1$ and:

$$\overline{z} = \sum_{i=1}^{k} \lambda_i \overline{z}^{(i)} \tag{136}$$

We can then conclude that:

$$\langle c, \overline{z} \rangle = \sum_{i=1}^{k} \lambda_i \langle c, \overline{z}^{(i)} \rangle \geq \min_i \langle c, \overline{z}^{(i)} \rangle \tag{137}$$

So, this point in $\mathcal{T}$ has a *smaller* objective, even though $\mathcal{T} \subset conv(\mathcal{T})$. Therefore:

$$\inf_{z \in \mathbb{E}^d} \langle c, z \rangle \;\; s.t. z \in conv(\mathcal{T}) \tag{138}$$

Has a larger optimal value than the other. $\square$

The key challenge with this convex reformulation in practice is that it is difficult to compute convex hulls of arbitrary sets.
This challenge leads us to the following question: how can we derive "good" ways to describe and represent convex sets? To describe any set, we some way to identify whether a given point is an element of the set or not.
Our task is therefore the following: can we obtain or derive certificates of set membership in the convex case? We've already seen that hyperplanes can be used to certify that a point does not belong to a convex set - let's see if we can expand on this idea.
The **external description** of a convex set is used to certify that points are *not* in a convex set. To begin with, consider a polyhedron:

$$S = \{x \in \mathbb{E}^d | \, \langle a^{(i)}, x \rangle \leq c_i, i = 1, ..., k\} \tag{139}$$

This set is the intersection of finitely many halfspaces. Here, $a^{(1)}, ..., a^{(k)} \in \mathbb{E}^d$ and $c_1, ..., c_k \in \mathbb{R}$.
So, with this description, certifying a point $\overline{x} \notin S$ amounts to providing an $i \in \{1, ..., k\}$ such that:

$$\langle a^{(i)}, \overline{x} \rangle > c_i \tag{140}$$

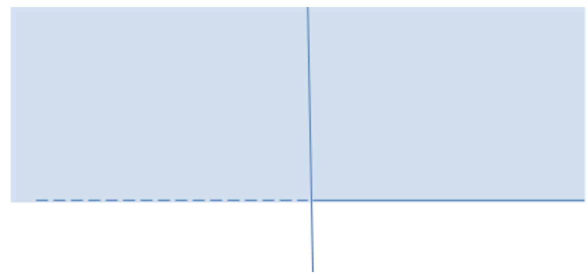
This inequality *explicitly* gives us a certificate that $\overline{x}$ is not in the set.
Let's extend this procedure to arbitrary convex sets.

Proposition 5 ***Halfspace Representation of Convex Sets***
Let $S \subset \mathbb{E}^d$ be a closed convex set. We have that:

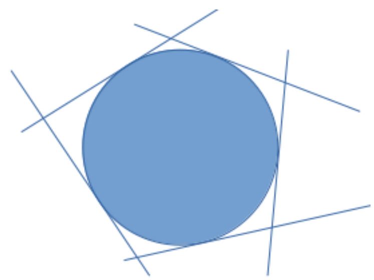
$$S = \bigcap_{i \in \mathcal{I}} \{x \in \mathbb{E}^d | \, \langle a^{(i)}, x \rangle \leq c_i\} \tag{141}$$

With $a^{(i)} \in \mathbb{E}^d$ and $c_i \in \mathbb{R}$ for all $i \in \mathcal{I}$, where $\mathcal{I}$ is a (potentially infinite) index set.

Note that in this proposition, we require that the set is closed to ensure a set of halfspaces will exist. For instance, consider the following example of a convex but not closed set:



This set cannot be expressed as an intersection of halfspaces, as in the left half of the plane, the axis is not included, while on the right half of the plane, the axis is included. Requiring closure eliminates possibilities of this form.
Now, consider the following example, where we need an infinite number of halfspaces to represent the closed ball:



Despite needing an infinite number of halfspaces, we can still apply the same idea of external representation! All we need to do to show $\overline{x} \notin S$ is provide an $i \in \mathcal{I}$ such that:

$$\langle a^{(i)}, \overline{x} \rangle > c_i \tag{142}$$

Let's think more explicitly about how to represent this closed ball with halfspaces. The closed unit ball in $\mathbb{E}^d$ is the set:

$$S = \{x \in \mathbb{E}^d | \ ||x|| \leq 1\} \tag{143}$$

Here, we have that:

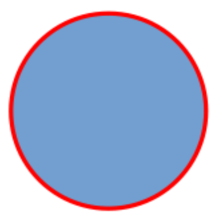
$$S = \bigcap_{a \in \mathbb{E}^d, ||a||=1} \{x \in \mathbb{E}^d | \ \langle a, x \rangle \leq 1\} \tag{144}$$

This is a "halfspace way" of describing the Euclidean ball. To show that some $\overline{x} \notin S$, we could use the halfspace corresponding to $a = \overline{x}/||\overline{x}||$ for verification! This provides the same form of verification as checking that the length of x is > 1, simply using halfspaces instead of explicitly checking the norm.
Now, we consider an **internal description** of convex sets. The internal description of a convex set is concerned with methods for verifying that a point *is* a member of a convex set.

Consider the following basic idea. Let S be a convex set. To certify that some $\overline{x} \in S$, we could describe $\overline{x}$ as a convex combination of points that are *known* to be in S. This suggests the idea of describng a convex set as a convex hull of *some collection* of points we know to be in the convex set.
Consider the closed Euclidean ball, for example. Here, we could look at the boundary of the ball, as the convex hull of the boundary will contain the interior of the ball.
This inspires the following question: for a convex set S, what is a "minimal" collection of points in S whose convex hull equals S?

Definition 15 ***Extreme Point of a Convex Set***
Let $S \subset \mathbb{E}^d$ be a convex set. A point $x \in S$ is called an extreme point of S if x cannot be expressed as a non-trivial convex combination of other points in S.

It's important to note that the definition of an extreme point is one that *only* applies to convex sets! Let's think of a few simple examples to get a feel for what the extreme points of convex sets might be.
First, consider the disk in the plane.



The extreme points of this set, as discussed earlier, are the set of boundary points of the disk, highlighted above in red. Every point in the disk is a convex combination of the points on the boundary, and no point on the boundary is a convex combination of any other point.
Now, consider the box in the plane.

Here, the extreme points are the corners of the box.
At this point, it's important to ask the following question: is any convex set equal to the convex hull of its extreme points? In general, the answer to this

question is no! Consider the following counterexample, which represents the first quadrant of the plane:

Although this set is convex, its only extreme point is the origin! Thus, we cannot generate every point in this convex set with convex combinations of its extreme points.
So, the application of the idea of extreme points *seems* to require that the set in question be closed and *bounded.*

Theorem 3 ***Krein-Milman Theorem***
Let $S \subset \mathbb{E}^d$ be a compact (closed and bounded) convex set. We have that S is equal to the convex hull of its extreme points.

This theorem confirms our "closed and bounded" idea from above! Next, consider the following theorem, which is often simply referred to as Caratheodory's theorem.

Theorem 4 ***Minkowski-Caratheodory Theorem***
Let $S \subset \mathbb{E}^d$ be a compact convex set. Any $x \in S$ can be expressed as a convex combination of at most $d+1$ extreme points of S.

It's important to note that in the theorem above, the collection of $d+1$ points is not necessarily the same for all $x \in S$! A different collection of $d+1$ points could be used to represent each $x \in S$.
This theorem provides us with a certificate of membership for compact convex sets! To certify $x \in S$, we could explicitly pick these $d+1$ points.
Next, we'd like to work with specific families of sets that we know *everything* about, both internally and externally!
Once we know how to work with simple "primitives" that we have characterized internally and externally, we may reach conclusions about more complex sets by considering the intersections of these primitive sets. For instance, if we have good descriptions of sets S_1 and S_2, then we can verify membership for $S_1 \cap S_2$ by checking that the extreme point constraints are satisfied for both S_1 and S_2.

8 Lecture 8: Cone Programming

In this lecture, we'll begin developing the structure of a common set of optimization problems called **cone programs**. These types of optimization problems are natural generalizations of another type of problem called a **linear program**. First, we introduce the linear program. A linear program is an optimization problem in which a linear function is optimized over a constraint set that is a polyhedron.

$$v = \inf_{x \in \mathbb{E}^d} \langle c, x \rangle \tag{145}$$

$$s.t.\ \langle a^{(i)}, x \rangle + b_i \geq 0,\ i = 1, ..., k \tag{146}$$

This problem is a convex optimization problem, as it has a linear objective function and a convex constraint set.
The main constraint of a linear program relative to the general convex optimization problem is that the constraint set is a polyhedron - a set composed of a finite number of intersections of halfspaces.
For the general convex optimization problem, constraint sets may be written an an infinite number of intersections of halfspaces - this is a much more challenging condition to work with than the finite case.
Cone programs generalize linear programs by considering more general types of constraints. Specifically, let $K \subset \mathbb{E}^n$ be a convex cone. A cone program with respect to K is an optimization problem of the form:

$$v = \inf_{x \in \mathbb{E}^d} \langle c, x \rangle \tag{147}$$

$$s.t. Ax + b \in K \tag{148}$$

Here, $A : \mathbb{E}^d \to \mathbb{E}^n$ is a linear map, $b \in \mathbb{E}^n$, and $c \in \mathbb{E}^d$. Notice that to obtain the previous linear program as a special case of a general cone program, we let:

$$A = \begin{bmatrix} - & a^{(1)*} & - \\ & \vdots & \\ - & a^{(n)*} & - \end{bmatrix},\ b = \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix} \tag{149}$$

$$K = \mathbb{R}^n_+ = \{x \in \mathbb{R}^n | x_i \geq 0,\ i = 1, ..., n\} \tag{150}$$

This cone, $K = \mathbb{R}^n_+$, is called the **non-negative orthant in** $\mathbb{R}^n$. Hence, a linear program is a cone program with respect to the non-negative orthant.
Now that we've set up the general cone program framework, we ask - what is the key source of complexity in cone programs? Does our cone have an efficient internal or external description? For general convex cones, we need a way to efficiently certify is a given point is in the cone or not. This ability will enable us to more easily analyze cone programs.
For instance, let's think about how we could certify inclusion of a point into the non-negative orthant. To certify that some $x \notin K$, we can use a hyperplane to separate x from K.

On the other hand, to certify that some $x \in K$, we need an analogue of the idea of an extreme point for cones! Unfortunately, the basic idea of an extreme point is *not* one that translates well to cones, as typically, cones will only have a single extreme point.
Consider the following definition, which generalizes the concept of an extreme point to convex cones.

Definition 16 ***Extreme Ray***
Let $K \subset \mathbb{E}^n$ be a convex cone. An extreme ray of K is a set:

$$\{\alpha u | \alpha \geq 0, u \in K/\{0\}\} \tag{151}$$

Such that no point in this set can be described in a convex combination of points in K that lie outside of this set.

Let's generate a few examples of extreme rays of convex cones. First, consider the non-negative orthant $K = \mathbb{R}^n_+$. Each set:

$$\{\alpha e^{(i)} | \alpha \geq 0\} \tag{152}$$

Where $e^{(i)}$ is the i'th standard basis vector for $\mathbb{R}^n$, is an extreme ray of $\mathbb{R}^n_+$.
It's important to note that this *entire set* is a ray. Some texts will refer to this set as "the ray generated by $e^{(i)}$."
As a second example, consider the Euclidean space $\mathbb{S}^m$ of $m \times m$ real symmetric matrices equipped with the trace inner product. Let's consider the set of positive semidefinite matrices in $\mathbb{S}^m$:

$$\mathbb{S}^m_+ = \{x \in \mathbb{S}^m | X \succeq 0\} \tag{153}$$

This set forms a convex cone. Imagine we wanted to certify that some $X \notin \mathbb{S}^m_+$. If $X \notin \mathbb{S}^m_+$, then there must exist $u \in \mathbb{R}^m$ such that:

$$u^* X u < 0 \tag{154}$$

Let's rephrase this condition in terms of the inner product. If $X \notin \mathbb{S}^m_+$, then there exists $u \in \mathbb{R}^m$ such that:

$$\langle uu^*, X \rangle < 0 \tag{155}$$

Here, we may think of uu^* as a rank one matrix that *defines the normal of a hyperplane!* Remember - since $\mathbb{S}^d$ is a Euclidean space, separating hyperplanes on $\mathbb{S}^m$ follow the same definitions as those on the more general space $\mathbb{E}^d$!
To efficiently find such a u satisfying the condition above, we would need to compute an eigendecomposition of X. Fortunately, this is a computationally efficient process that may be done in polynomial time.

This gives us a method of certifying that $X \notin \mathbb{S}^m+$ - how could we instead certify that $X \in \mathbb{S}^m_+$? To do this, we first need to characterize the extreme rays of $\mathbb{S}^m_+$.
The extreme rays of $\mathbb{S}^m_+$ are the sets:

$$\{\alpha uu^* | \alpha \geq 0, u \in \mathbb{R}^n\{0\}\} \tag{156}$$

Now, we state a theorem regarding membership in convex cones.

Theorem 5 ***Krein-Milman/Minkowski-Caratheodory Theorem for Convex Cones***
Let $K \subset \mathbb{E}^n$ be a closed convex cone that is pointed ($K \cap -K = \{0\}$). Any element of K can be expressed as a convex combination of at most n elements that lie in extreme rays of K.

Let's consider a few examples of this theorem. First, for the non-negative orthant in $\mathbb{R}^n$, we can express each point in $\mathbb{R}^n_+$ using at most n extreme rays.
Secondly, for the PSD cone in $\mathbb{S}^m_+$, we can express each point in $\mathbb{S}^m_+$ using at most m extreme rays - this may be shown using an eigendecomposition. Notice that this is in spite of the dimension of the PSD cone being larger than m - the theorem above only gives us an *upper bound* on the number of elements we need - here, we can be more efficient!
The PSD cone is associated with a special type of optimization problem. Cone programs with respect to the PSD cone are called **semidefinite programs (SDPs)**. They have the form:

$$v = \inf_{x \in \mathbb{E}^d} \langle c, x \rangle \tag{157}$$

$$s.t.\ Ax + b \in \mathbb{S}^m_+ \tag{158}$$

Here, $A : \mathbb{E}^d \to \mathbb{S}^m$ is a linear map, $b \in \mathbb{S}^m_+$, and $c \in \mathbb{E}^d$.
The constraint set for an SDP has a special name. Sets of the form:

$$\{x \in \mathbb{E}^d | Ax + b \in \mathbb{S}^m_+\} \tag{159}$$

Are called **spectrahedra**. Note that all polyhedra are spectrahedra, but spectrahedra are not necessarily polyhedra.

9 Lecture 9: Conic Duality

Let's recap what we covered last time. Previously, we introduced cone programs, which minimize a linear function over a constraint set that is the intersection of a convex cone and an affine space. Cone programs generalize linear programs, and contain semidefinite programs as a special case. Further, cone programs are

tractable if the cones in the constraint set have an efficient internal and external description.
In this lecture, we'll study duality and optimality conditions for cone programming. Consider the standard cone program:

$$v = \inf_{x \in \mathbb{E}^d} \langle c, x \rangle \tag{160}$$

$$s.t.\ Ax + b \in \mathcal{K} \tag{161}$$

Here, $A : \mathbb{E}^d \to \mathbb{E}^n$ is a linear map, $\mathcal{K} \subset \mathbb{E}^n$ is a convex cone, $b \in \mathbb{E}^n$, and $c \in \mathbb{E}^d$.
Now, we ask the following question: How can we write the support function for the constraint set $Ax + b \in \mathcal{K}$? To derive the dual of this problem, we'll instead consider the constraint set to be an intersection fo simpler constraint sets. Consider the following reformulation:

$$v = \inf_{(x,y) \in \mathbb{E}^d \times \mathbb{E}^n} \langle c, x \rangle \tag{162}$$

$$s.t.\ Ax + b - y = 0,\ y \in \mathcal{K} \tag{163}$$

Here, our objective remains linear in x. The two constraints, $Ax + b - y = 0$ and $y \in \mathcal{K}$, form an affine space in $\mathbb{E}^d \times \mathbb{E}^n$. This second constraint $y \in \mathcal{K}$ may be thought of as a conic condition on both x and y if we represent it as $(x, y) \in \mathbb{E}^d \times \mathcal{K}$. This is a conic condition on both x and y.
We now have all of the pieces in place to write down the Fenchel dual of this problem. First, we must take the conjugate of a linear function - we can do this without too much trouble. Secondly, we must write the support function of an affine space. Thirdly, we must write the support function of an convex cone.
There are two main paths that we can take to deriving the Fenchel dual. Once we have the support functions of each of these sets individually, we can find the support function of their intersection with the following result:

Proposition 6 ***Support of an Intersection of Sets***
Let $S, T \subset \mathbb{E}^d$ be two convex sets satisfying $int(S) \cap T \neq \emptyset$. Then:

$$h_{S \cap T}(x) = \inf_{y,z \in \mathbb{E}^d} h_S(y) + h_T(z)\ s.t.\ y + z = x \tag{164}$$

This result, when combined with the above, allows us to derive the Fenchel dual of the cone program. This is a perfectly valid approach to the problem, and is one way to arrive at the dual!
In this lecture, we'll follow a *slightly* different approach that involves somewhat less notational overhead. This approach will involve a trick that is commonly used with linear objective functions.
This path towards deriving the Fenchel dual will rely on the following definition:

Definition 17 ***Polar of a Convex Cone***
Let $\mathcal{K} \subset \mathbb{E}^n$ be a convex cone. The polar of $\mathcal{K}$, denoted $\mathcal{K}^\circ$, is defined as:

$$\mathcal{K}^\circ = \{y \in \mathbb{E}^n | \langle y, x \rangle \leq 0 \ \forall x \in \mathcal{K}\} \tag{165}$$

Now that we have this definition, let's get started on deriving the dual! Consider the convex program with objective and constraint set defined as follows:

$$f(x, y) = \langle c, x \rangle + \langle 0, y \rangle, \ dom(f) = \mathbb{E}^d \times \mathcal{K} \tag{166}$$
$$S = \{(x, y) | Ax - y + b = 0\} \tag{167}$$

Notice that we have incorporated the constraint $y \in \mathcal{K}$ into the domain of the objective function! To obtain the dual of the optimization with this objective and constraint, we need to derive the conjugate of f and the support of S. First, we derive the conjugate of f.

$$f^*(\lambda, \mu) = \sup_{(x,y) \in \mathbb{E}^d \times \mathcal{K}} \langle (\lambda, \mu), (x, y) \rangle - \langle (c, 0), (x, y) \rangle \tag{168}$$
$$= \sup_{(x,y) \in \mathbb{E}^d \times \mathcal{K}} \langle (\lambda - c, \mu), (x, y) \rangle \tag{169}$$

Now, we notice that $\mathbb{E}^d \times \mathcal{K}$ is a convex cone, and that we are maximizing a linear function over a convex cone. Thus, we get the following definition of f^*:

$$f^*(\lambda, \mu) = \begin{cases} 0, (\lambda - c, \mu) \in (\mathbb{E}^d \times \mathcal{K})^\circ \\ \infty, \ o.w. \end{cases} \tag{170}$$

Note that here, $\circ$ refers to the *polar* of a cone! This is derived using the fact that the inersection between a cone and its polar is simply the set $\{0\}$. Now, we examine this function further. We may break up the polar of the Cartseian product into the polars of each individual set.

$$f^*(\lambda, \mu) = \begin{cases} 0, (\lambda - c) \in (\mathbb{E}^d)^\circ, \ \mu \in (\mathcal{K})^\circ \\ \infty, \ o.w. \end{cases} \tag{171}$$

But, the polar of $\mathbb{E}^d$ is simply the set $\{0\}$. Thus, we have that:

$$f^*(\lambda, \mu) = 0, \ dom(f^*) = \{(\lambda, \mu) | \lambda = c, \mu \in \mathcal{K}^\circ\} \tag{172}$$

This completes the derivation of the conjugate function. Now, we need to deal with the support function. The support function is defined:

$$h_s(\lambda, \mu) = \sup_{(x,y) \in \mathbb{E}^d \times \mathbb{E}^n} \langle (\lambda, \mu), (x, y) \rangle \ s.t. \ Ax - y + b = 0 \tag{173}$$

$Ax - y + b = 0$ forms an affine space in terms of x and y. How can we find the support function of an affine space? Let's think about the solution space of $Ax - y + b = 0$. First, we rewrite this in terms of the (x, y) vector as:

$$[A, -I](x, y) + b = 0 \tag{174}$$

We know that the solution space of this linear system of equations may be written in the form:

$$N([A, -I]) + (x_0, y_0) \tag{175}$$

Where $N([A, -I])$ is the null space of $[A, -I]$ and (x_0, y_0) is a particular solution to the equation. Note that $+$ here denotes that we are shifting the entire null space by the vector (x_0, y_0). Using this representation of the constraint, the support of the space above may be equivalently written:

$$= \sup_{(x,y) \in \mathbb{E}^d \times \mathbb{E}^n} \langle (\lambda, \mu), (x, y) \rangle \ \ s.t. \ (x, y) \in N([A, -I]) + (x_0, y_0) \tag{176}$$

$$= \langle (\lambda, \mu), (x_0, y_0) \rangle + \sup_{(x,y) \in \mathbb{E}^d \times \mathbb{E}^n} \langle (\lambda, \mu), (x, y) \rangle \ \ s.t. \ (x, y) \in N([A, -I]) \tag{177}$$

Let's think about the different possibilities for this function. If (λ, μ) is orthogonal to the null space of $[A, -I]$, then the remaining supremum will be zero. Otherwise, this supremum term will go to ∞, as we may scale up a component of (λ, μ) in the null space until this term goes to infinity.
Under this logic, the support function of the set is defined:

$$h_s(\lambda, \mu) = \langle (\lambda, \mu), (x_0, y_0) \rangle + \begin{cases} 0, \ (\lambda, \mu) \in Row([A, -I]) \\ \infty, \ o.w. \end{cases} \tag{178}$$

$$\tag{179}$$

Where $Row([A, -I])$ denotes the row space of $[A, -I]$. We may rewrite this using the column space of $[A, -I]^*$ and define the domain of the resulting function as follows:

$$h_s(\lambda, \mu) = \langle (\lambda, \mu), (x_0, y_0) \rangle \,, \ dom(h_s) = \left\{ \begin{bmatrix} A^* \omega \\ -\omega \end{bmatrix} \,\middle|\, \omega \in \mathbb{E}^n \right\} \tag{180}$$

Let's put this support function together with the conjugate to obtain the dual. The Fenchel dual of our cone program can be expressed as follows:

$$v_F^* = \sup_{\lambda \in \mathbb{E}^d, \mu, \omega \in \mathbb{E}^n} -0 - [\langle -(\lambda, \mu), (x_0, y_0) \rangle] \tag{181}$$

$$s.t. \ \lambda = c, \mu \in \mathcal{K}^\circ, \lambda = A^* \omega, \mu = -\omega \tag{182}$$

Notice that, as with all duals, there are no appearances of x or y as decision variables. Let's simplify this problem by substituting some of the constraints into the objective function. This yields:

$$v_F^* = \sup_{\omega \in \mathbb{E}^n} \langle (A^*\omega, -\omega), (x_0, y_0) \rangle \tag{183}$$

$$s.t.\ A^*\omega = c, \mu \in \mathcal{K}^\circ \tag{184}$$

$$v_F^* = \sup_{\omega \in \mathbb{E}^n} \langle \omega, Ax_0 - y_0 \rangle \tag{185}$$

$$s.t.\ A^*\omega = c, \omega \in -\mathcal{K}^\circ \tag{186}$$

$$v_F^* = \sup_{\omega \in \mathbb{E}^n} - \langle b, \omega \rangle \tag{187}$$

$$s.t.\ A^*\omega = c, \omega \in -\mathcal{K}^\circ \tag{188}$$

In this final simplification, the decision variable is of dimension n instead of d. We can check that this dual problem is a strong dual, $(v_F^* = v)$, provided that there exists $\tilde{x}$ such that $A\tilde{x} + b \in int(\mathcal{K})$.
Let's perform a quick check of *weak* duality. Recall that in all cases, we should have $v_F^* \leq v$. Starting with the primal objective:

$$\langle c, x \rangle = \langle c, x \rangle + \langle b, \omega \rangle - \langle b, \omega \rangle \tag{189}$$

We'd like the last term, $- \langle b, \omega \rangle$, to be left over, as this is the Fenchel dual objective. To show weak duality holds, we'd like to bound the sum of the first two terms below by zero over all x feasible for the primal and ω feasible for the dual. Substituting the constraints for this feasibility, we see:

$$\langle c, x \rangle = \langle A^*\omega, x \rangle + \langle b, \omega \rangle - \langle b, \omega \rangle \tag{190}$$

$$= \langle \omega, Ax + b \rangle - \langle b, \omega \rangle \tag{191}$$

We know that for feasible ω and x, $\omega \in -\mathcal{K}^\circ$ and $Ax + b \in \mathcal{K}$. By the definition of the polar, this inner product must be greater than or equal to zero. Therefore:

$$\langle c, x \rangle \geq - \langle b, \omega \rangle \tag{192}$$

For all primal feasible x and dual feasible ω.
Let's recap the process that we took in this check of weak duality. We observe that we can use Fenchel duality to systematically check what to "add and subtract" to derive a lower bound on the primal objective. In this case, we knew what to add and subtract to get the Fenchel dual objective to appear on the right hand side.
At optimality, there must be points where the dual and primal objective are equal. Recall that for primal optimal $\hat{x}$ and dual optimal $\hat{\omega}$, the inequality above holds with equality in the case that strong duality holds. Analyzing this case for specific cone programs helps us find the optimality conditions associated with that problem.

10 Lecture 10: Linear & Semidefinite Programs

Last lecture, we derived the dual form of a cone program. Recall that the standard primal form of a cone program is:

$$v = \inf_{x \in \mathbb{E}^d} \langle c, x \rangle \tag{193}$$

$$s.t.\ Ax + b \in \mathcal{K} \tag{194}$$

Where $A : \mathbb{E}^d \to \mathbb{E}^n$ is a linear map, $b \in \mathbb{E}^n$, $c \in \mathbb{E}^d$, and $\mathcal{K} \subset \mathbb{E}^n$ is a convex cone.
The dual problem that we derived via Fenchel duality was:

$$v_F^* = \sup_{\omega \in \mathbb{E}^n} - \langle b, \omega \rangle \tag{195}$$

$$s.t.\ A^* \omega = c,\ \omega \in -\mathcal{K}^\circ \tag{196}$$

In this lecture, we'll discuss linear programs, a special case of the cone program in which the cone $\mathcal{K}$ is the nonnegative orthant, $\mathbb{R}^n_+$. Following this we'll examine semidefinite programming, in which $\mathcal{K}$ is the cone of PSD matrices.
We'll begin with linear programming. This type of problem, which aims to optimize a linear objective over a polyhedron, may be written in the standard cone program form as an optimization over $Ax + b \in \mathcal{K}$, where $\mathcal{K}$ is the nonnegative orthant.
Let's think of some examples of common constraint sets for linear programs. First, we consider the **simplex** in $\mathbb{R}^n$, defined:

$$\Delta^n = \{x \in \mathbb{R}^n |\, \langle 1, x \rangle = 1, x \geq 0\} \tag{197}$$

This set is a polyhedron. This may be seen by reformulating the definition of the set as follows

$$\Delta^n = \{x \in \mathbb{R}^n |\, \langle 1, x \rangle \geq 1, \langle 1, x \rangle \leq 1, x \geq 0 \tag{198}$$

This forms a system of linear inequalities that may be represented in the standard polyhedron form $Ax \leq b$. Looking at this expanded representation of the simplex, we see that we need $n + 2$ inequality constraints to represent the set (2 from the $\langle 1, x \rangle$ conditions and n from $x_i \geq 0\ \forall i$). This means that we need a nonnegative orthant of size $n + 2$ to describe the simplex in the standard cone programming form.
Next, we consider the l_∞ ball in $\mathbb{R}^n$. This set is defined:

$$B_{l_\infty} = \{x \in \mathbb{R}^n |\, \max_i |x_i| \leq 1\} \tag{199}$$

$$= \{x \in \mathbb{R}^n | -1 \leq x_i \leq 1, i = 1, ..., n\} \tag{200}$$

From this second representation of B_{l_∞}, we see that we may write this set as a polyhedron with $2n$ linear inequality constraints.

Next, we examine the l_1 ball in $\mathbb{R}^n$. This set is described:

$$B_{l_1} = \{x \in \mathbb{R}^n | \sum_i |x_i| \leq 1\} \tag{201}$$

$$= \{x \in \mathbb{R}^n | \max_{u \in \{-1,+1\}^n} \langle u, x \rangle \leq 1\} \tag{202}$$

Here, our alternate representation of the set takes the maximum over all signed vectors in $\mathbb{R}^n$, the vectors where each element is either $+1$ or -1. Using a counting argument, we can show that this involves 2^n inequalities for the linear program constraint, which is challenging to deal with in high dimensions!
To reduce the amount of inequalities needed to describe this constraint set, we may write the l_1 ball as a projection of a higher dimensional object onto a lower dimensional space. Consider the following representation of the ball:

$$B_{l_1} = \{x \in \mathbb{R}^n | \; \exists y \in \mathbb{R}^n \; s.t. \; -y_i \leq x_i \leq y_i, i = 1, ..., n, \; \langle 1, y \rangle \leq 1\} \tag{203}$$

This is a description of the l_1 ball with n extra variables but only $2n+1$ linear inequalities!
Let's show that this representation of the l_1 ball is equivalent to our standard representation. First, suppose $x \in B_{l_1}$, the standard representation of the l_1 ball. If y is chosen such that $y_i = |x_i|$, then the constraint on the alternate representation is satisfied! Thus, we conclude that our standard representation is a subset of the alternate representation.
Now, we go the other way. Let x be in the alternate representation of the l_1 ball. If $\sum_i y_i \leq 1$, then we know $\sum_i |x_i| \leq 1$. This tells us that the alternate representation is a subset of the standard representation. Thus, the two are equivalent!
These examples conjure the following question regarding the constraint sets of linear programs: given a particular constraint set for a linear program, what is the "minimal representation" of the constraint set? The **linear program rank** of a polyhedron is defined to be the smallest size description of a polyhedron. This provides a numerical answer to this question.
Note that although the norm balls discussed above have polyhedron representations, this is *not* the case for all norm balls! For instance, the Euclidean ball has no representation as a polyhedron.
Let's now turn our attention back to the case of a specific linear program for the case of the l_1 ball. Concretely, suppose that we wish to solve:

$$\inf_{x \in \mathbb{R}^n} \langle c, x \rangle \;\; s.t. \; x \in B_{l_1} \tag{204}$$

$$= \inf_{(x,y) \in \mathbb{R}^{2n}} \langle (c, 0), (x, y) \rangle \tag{205}$$

$$s.t. \; (x, y) \in \{(x, y) \in \mathbb{R}^{2n} | \; -y_i \leq x_i \leq y_i, \; \langle 1, y \rangle \leq 1\} \tag{206}$$

Here, to accommodate the extra y, we simply add a vector of zeros to the end of c. Such descriptions in which we add extra variables and obtain a "simplex" set in a higher dimensional case are called **extended formulations**.

Optimality Conditions for a Linear Program

Recall that the problem of interest is:

$$\inf_{x\in\mathbb{R}^d} \langle c, x\rangle \tag{207}$$

$$s.t.\ Ax + b \in \mathbb{R}^n_+ \tag{208}$$

Where the dual of this problem is given by:

$$\sup_{\omega\in\mathbb{R}^n} -\langle b, \omega\rangle \tag{209}$$

$$s.t.\ A^*\omega = c,\ \omega \in -(\mathbb{R}^n_+)^\circ \tag{210}$$

Let's perform a quick check that weak duality holds for this dual formulation. Weak duality follows via the sequence of inequalities:

$$\langle c, x\rangle = \langle c, x\rangle + \langle b, \omega\rangle - \langle b, \omega\rangle \tag{211}$$

$$= \langle A^*\omega, x\rangle + \langle b, \omega\rangle - \langle b, \omega\rangle \tag{212}$$

$$= \langle \omega, Ax + b\rangle - \langle b, \omega\rangle \tag{213}$$

Here, we know that any feasible ω satisfies $\omega \in -(\mathbb{R}^n_+)^\circ$ and any feasible x satisfies $Ax + b \in \mathbb{R}^n_+$. Without too much trouble, we can actually check that $-(\mathbb{R}^n_+)^\circ = \mathbb{R}^n_+$! This tells us that the first inner product must be non-negative, by definition of the non-negative orthant. Thus, we have that:

$$\langle c, x\rangle \geq -\langle b, \omega\rangle \tag{214}$$

For primal feasible x and dual feasible ω. Thus, weak duality holds in this problem.

Let's examine duality further for this problem. Suppose strong duality holds, with $\hat{x}$ primal optimal and $\hat{\omega}$ dual optimal. In this case, we should have by strong duality that:

$$\langle \hat{\omega}, A\hat{x} + b\rangle = 0 \tag{215}$$

Let's rewrite this inner product condition element-wise. We know that both $\hat{\omega}$ and $A\hat{x} + b$ must be entry-wise non-negative, since they are in the non-negative orthant. Thus, for the condition above to hold, we require that for all $i \in 1, ..., n$:

$$\hat{\omega}_i \cdot (A\hat{x} + b)_i = 0 \tag{216}$$

This condition is known as **complementarity**. Note that this is one particular instance of a complementarity condition for linear programs - conditions of this form also exist for more general problems.

Let's summarize the set of optimality conditions for a linear program. We have:

$$\text{Primal Feasibility: } A\hat{x} + b \in \mathbb{R}^n_+ \tag{217}$$

$$\text{Dual Feasibility: } A^*\omega = c,\ \omega \in \mathbb{R}^n_+ \tag{218}$$

$$\text{Complementarity: } \hat{\omega}_i \cdot (A\hat{x} + b)_i = 0,\ i = 1, ..., n \tag{219}$$

So, any $(\hat{\omega}, \hat{x})$ pair satisfying these three will mean that $\hat{x}$ is optimal. These conditions therefore form a certificate of optimality of $\hat{x}$.

Semidefinite Programming

Now, we shift our attention to semidefinite programming (SDP). Here, the constraint sets are of the form:

$$S = \{x \in \mathbb{E}^d | Ax + b \in \mathbb{S}^m_+\} \tag{220}$$

Where $A : \mathbb{E}^d \to \mathbb{S}^m_+$ is a linear map and $b \in \mathbb{S}^m$. We recall that sets of the form S are called spectrahedra. Semidefinite programs are the class of problems in which a linear function is optimized over a spectrahedron. Let's see if we can use the same trick as before to rewrite this constraint set S as the projection of a higher dimensional spectrahedron.
Before we do this, we recall the following key fact about polyhedrons: a projection of a polyhedron is always a polyhedron, as linear transformations preserve polyhedron structure. However, this is *not* true in general for spectrahedra!
Now, we consider some examples of spectrahedra. First, consider the set:

$$S = \{x \in \mathbb{E}^d | Ax + b \in \mathbb{R}^n_+\} \tag{221}$$

Here, $A : \mathbb{E}^n \to \mathbb{R}^n$ is linear and $b \in \mathbb{R}^n$. This set is a polyhedron! Can we verify that this is a spectrahedron? To show this, we must find some map $A : \mathbb{E}^d \to \mathbb{S}^m_+$ and a $b \in \mathbb{S}^m$ that completely characterize the set.
For convenience, we consider the case where $A : \mathbb{R}^d \to \mathbb{R}^n$. Then, we decompose A using its columns as $Ax := \sum_{i=1}^d A^{(i)} x_i$. We then define $\overline{A} : \mathbb{R}^d \to \mathbb{S}^n_+$ as:

$$\overline{A}x = \sum_{i=1}^{d} \overline{A}^{(i)} x_i \tag{222}$$

Where $\overline{A}^{(i)} = diag(A^{(i)})$ and $B = diag(b)$, the diagonal matrices with $A^{(i)}$ and b as their diagonal entries. Thus, we see that we may write this polyhedron as a spectrahedron! For the more general case, we may show that all polyhedra are spectrahedra. However, not all spectrahedra are polyhedra.
Let's now consider the following example:

$$B_{l_2} = \{x \in \mathbb{R}^d | x^T x \leq 1\} \tag{223}$$

We know that this set is *not* a polyhedron, since it requires an infinite number of halfspace constraints to represent. How else can we represent this set? We may alternatively write this set as:

$$B_{l_2} = \left\{ x \in \mathbb{R}^d \middle| \begin{bmatrix} 1 & x^T \\ x & I \end{bmatrix} \succeq 0 \right\} \tag{224}$$

Thus, this set is a spectrahedron, but not a polyhedron! This rewriting is accomplished using **Schur complements**. The Schur complement rule states that for a symmetric matrix:

$$\begin{bmatrix} A & B^T \\ B & C \end{bmatrix} \tag{225}$$

Where $C \succ 0$, we have that:

$$A - B^T C^{-1} B \succeq 0 \Leftrightarrow \begin{bmatrix} A & B^T \\ B & C \end{bmatrix} \succeq 0 \tag{226}$$

This is a useful tool that enables us to rewrite many sets as spectrahedra.

11 Lecture 11: Lagrange Duality I

Over the last few lectures, we've looked at optimization problems of the form:

$$v = \inf_{x \in \mathbb{E}^d} f(x) \tag{227}$$

$$s.t. \; x \in S \tag{228}$$

Now, we'll turn our attention to optimization problems of the form:

$$v = \inf_{x \in \mathbb{E}^d} f_0(x) \tag{229}$$

$$s.t. \; f_i(x) \leq 0, \; i = 1, ..., k, \; g_i(x) = 0, \; j = 1, ..., m \tag{230}$$

Here, $f_0, ..., f_k, g_1, ..., g_m : \mathbb{E}^d \to \mathbb{R}$. Optimization problems of this form are called **nonlinear programs (NLPs)**.
In this lecture, we'll examine the **Lagrange duality scheme**, the most commonly used duality scheme in nonlinear programming. We'll follow the same two steps in coming up with this duality scheme as we did in Fenchel duality.
Recall: to certify that *some* γ is a lower bound on the optimal value of the primal problem, v (and hence establishes a duality scheme), we will follow a two step process.

1. First, we reformulate the problem of proving a lower bound as a separation of two special sets. For the Lagrange duality scheme, we define two sets, S and T. S is defined:

$$S = \{(t, u) \in \mathbb{R}^{k+1} \times \mathbb{R}^m | \; \exists x \in \mathbb{E}^d \; s.t.(x, t_0) \in epi(f_0 - \gamma), \tag{231}$$

$$(x, t_i) \in epi(f_i), \; i = 1, ..., j \tag{232}$$

$$g_j(x) = u_j, \; j = 1, ..., m\} \tag{233}$$

   While T is defined:

$$T = \{(t, u) \in \mathbb{R}^{k+1} \times \mathbb{R} | \; t_0 < 0, \tag{234}$$

$$t_i \leq 0, \; i = 1, ..., k, \tag{235}$$

$$u_j = 0, \; j = 1, ..., m\} \tag{236}$$

   Observe that $\gamma \leq v$ is equivalent to the sets S and T being disjoint.

2. Secondly, we certify that $S \cap T = \emptyset$ using a separating hyperplane with normal vector $(\lambda, \mu) \in \mathbb{R}^{k+1} \times \mathbb{R}^m$ such that $\lambda_0 \neq 0$.

This is the *same* two step process that we took when developing Fenchel duality: reformulate $\gamma \leq v$ as showing two sets are disjoint, and find a hyperplane to certify that the sets are disjoint.
Let's think about why these S and T give us a valid duality scheme. When $S \cap T = \emptyset$, there is no x in $\mathbb{E}^d$ such that $(f_0 - \gamma)(x) \leq t_0 < 0$, $f_i(x) \leq t_i \leq 0$, and $g_j(x) = u_j = 0$. This implies that γ gives a lower bound on v.
When can we expect the procedure of hyperplane separation, used in step 2 of the duality scheme, to work? We know that S and T must be convex to apply the separating hyperplane theorem. Is this the case? Looking at the sets above, we see that T is *always* convex, and that the convexity of S depends on the convexity of the optimization problem.
Now that we've discussed these two points, let's examine step 2 of the procedure: separating S and T with a hyperplane. We aim to identify a $(\lambda, \mu) \in \mathbb{R}^{k+1} \times \mathbb{R}^m$ with $\lambda_0 \neq 0$ such that:

$$S \subset \{(t,u) | \langle \lambda, t\rangle + \langle \mu, u\rangle \geq \delta\} \tag{237}$$

$$T \subset \{(t,u) | \langle \lambda, t\rangle + \langle \mu, u\rangle < \delta\} \tag{238}$$

We may now frame this as an optimization problem. The dual problem corresponding to the Lagrange duality scheme is:

$$v_L^* = \sup_{(\lambda,\mu)\in\mathbb{R}^{k+1}\times\mathbb{R}^m,\ \delta,\gamma\in\mathbb{R}} \gamma \tag{239}$$

$$s.t.\ S \subset \{(t,u) | \langle \lambda, t\rangle + \langle \mu, u\rangle \geq \delta\} \tag{240}$$

$$T \subset \{(t,u) | \langle \lambda, t\rangle + \langle \mu, u\rangle < \delta\},\ \lambda_0 \neq 0 \tag{241}$$

How can we simplify this problem? If $\lambda_0 \neq 0$, and λ, μ, δ are feasible, then we may multiply (λ, μ, δ) by any positive number and retain feasibility. Thus, without loss of generality, we may assume λ_0 is equal to 1 or -1.
Let's examine the case where $\lambda = -1$. Based on the T constraint, we require:

$$\langle \lambda, t\rangle + \langle \mu, u\rangle < \delta \tag{242}$$

Here, t can be made arbitrarily small, since $t_0 < 0$. This means that if $\lambda_0 = -1$, then the term $\langle \lambda, t\rangle$ can be made to grow arbitrarily large! Thus, if $\lambda_0 = -1$, it would be impossible to find an upper bound δ on this expression. We therefore conclude that λ_0 must be equal to 1.
Now, we have a reduction of the complexity of the problem by one variable. Now, the constraint in the Lagrange dual is simplified as:

$$S \subset \{(t,u) \in \mathbb{R}^{k+1} \times \mathbb{R}^m | t_0 + \langle \lambda_{1:k}, t_{1:k}\rangle + \langle \mu, u\rangle \geq \delta\} \tag{243}$$

Since $(x, t_0) \in epi(f_0 - \gamma)$, we conclude that the smallest value of t_0 is $f_0(x) - \gamma)$. thus, this constraint implies:

$$(f_0(x) - \gamma) + \sum_{i=1}^{k} \lambda_i t_i + \sum_{j=1}^{m} \mu_j g_j(x) \geq \delta \tag{244}$$

$$\forall x \in \mathbb{E}^d\ s.t.\ (x, t_i) \in epi(f_i), i = 1, ..., k \tag{245}$$

Where we have simply expanded the inner products to get the two sum terms. Next, to simplify this further, we'd like to replace the instances of t_i with $f_i(x)$. However, we can only do this if $\lambda_i \geq 0$, as this will preserve the direction of the inequalities. Let's see if we can show $\lambda_i \geq 0$.
To do this, we examine the constraint on T:

$$T \subset \{(t,u) | t_0 + \langle \lambda_{1:k}, t_{1:k} \rangle + \langle \mu, u \rangle < \delta\} \tag{246}$$

This constraint implies that $\lambda_i \geq 0$, since t_i can be made arbitrarily small. This implies that if $\lambda_i < 0$, we could get an arbitrarily large quantity that could not be bounded above by a δ. This gives the desired constraint on λ_i.
Notice that $\mu \in \mathbb{R}^m$ is unrestricted, as we take its inner product with u, which is constrained to be a vector of zeros.
Thus, we may replace the instances of t_i with f_i in the expression above to get:

$$f_0(x) + \sum_{i=1}^{k} \lambda_i f_i(x) + \sum_{j=1}^{m} \mu_j g_j(x) \geq \delta + \gamma \tag{247}$$

$$\forall x \in dom(f_0) \cap ... \cap dom(g_m) \tag{248}$$

Notice that here, we have also moved γ to the other side of the inequality. This simplification leads us to the following important function:

Definition 18 ***Lagrangian of an NLP***
The Lagrangian associated with a nonlinear program is the function:

$$L(x, \lambda, \mu) = f_0(x) + \sum_{i=1}^{k} \lambda_i f_i(x) + \sum_{j=1}^{m} \mu_j g_j(x) \tag{249}$$

Where f_0 is the objective of the NLP, f_i are from its inequality constraints, and g_j are from its equality constraints.

The Lagrangian involves both (λ, μ) and the primal variable x. Before we proceed, let's redefine λ to be a vector in $\mathbb{R}^k$ instead of $\mathbb{R}^{k+1}$ - we can do this since we know that its first element is 1 without loss of generality.
Now, we define another important function:

Definition 19 ***Dual Function of an NLP***
Let $L(x, \lambda, \mu)$ be the Lagrangian associated with a nonlinear program. The dual function $\nu(\lambda, \mu)$ is defined:

$$\nu(\lambda, \mu) = \inf_{x \in \mathbb{E}^d} L(x, \lambda, \mu) \tag{250}$$

$$s.t.\ x \in dom(f_0) \cap ... \cap dom(g_m) \tag{251}$$

Notice that the dual function *eliminates* x from the Lagrangian by minimizing over it. With these two functions, we may now reformulate the problem associated with the Lagrange duality scheme as follows:

$$v_L^* = \sup_{\lambda \in \mathbb{R}^k, \mu \in \mathbb{R}^m, \delta, \gamma \in \mathbb{R}} \gamma \tag{252}$$

$$s.t. \nu(\lambda, \mu) \geq \delta + \gamma, \ \lambda \in \mathbb{R}_+^k, \ \delta \geq 0 \tag{253}$$

Here, the constraint $\nu(\lambda, \mu) \geq \delta + \gamma$ comes from the S halfspace constraint, while $\lambda \in \mathbb{R}_+^k, \ \delta \geq 0$ come from the T halfspace constraint.
Let's see if we can eliminate δ and γ from this expression. We know from the constraint that the largest value of γ we can pick is $\nu(\lambda, \mu) - \delta$. Thus, eliminating γ, we get:

$$v_L^* = \sup_{\lambda \in \mathbb{R}^k, \mu \in \mathbb{R}^m, \delta \in \mathbb{R}} \nu(\lambda, \mu) - \delta \tag{254}$$

$$s.t. \ \lambda \in \mathbb{R}_+^k, \ \delta \geq 0 \tag{255}$$

Now, we may eliminate δ using $\delta = 0$. This gives the optimization:

$$v_L^* = \sup_{\lambda \in \mathbb{R}^k, \mu \in \mathbb{R}^m} \nu(\lambda, \mu) \tag{256}$$

$$s.t. \ \lambda \in \mathbb{R}_+^k \tag{257}$$

This is the Lagrange dual problem associated with the primal problem.

Definition 20 ***Lagrange Dual***
The Lagrange dual problem associated with a given primal problem is:

$$v_L^* = \sup_{\lambda \in \mathbb{R}^k, \mu \in \mathbb{R}^m} \nu(\lambda, \mu) \tag{258}$$

$$s.t. \ \lambda \in \mathbb{R}_+^k \tag{259}$$

Where $\nu(\lambda, \mu)$ is the dual function of the nonlinear program.

To summarize, the Lagrange dual of an NLP is derived as follows:

1. Formulate the Lagrangian associated with the NLP:

$$L(x, \lambda, \mu) = f_0(x) + \sum_{i=1}^{k} \lambda_i f_i(x) + \sum_{j=1}^{m} \mu_j g_j(x) \tag{260}$$

2. Derive the dual function:

$$\nu(\lambda, \mu) = \inf_{x \in \mathbb{E}^d} L(x, \lambda, \mu) \tag{261}$$

$$s.t. \ x \in dom(f_0) \cap ... \cap dom(g_m) \tag{262}$$

3. Write down the Lagrange dual problem:

$$v_L^* = \sup_{\lambda \in \mathbb{R}^k, \mu \in \mathbb{R}^m} \nu(\lambda, \mu) \tag{263}$$

$$s.t.\ \lambda \in \mathbb{R}_+^k \tag{264}$$

Let's check that the Lagrange dual problem satisfies weak duality with respect to the primal. We know that weak duality is satisfied by construction, since the Lagrange dual was constructed by forming lower bounds on the primal optimal value using the Lagrange duality scheme. Thus, we know that weak duality, $v_L^* \leq v$, should hold.
We can also verify that weak duality holds with our usual "add and subtract method." We know:

$$f(x) \geq f_0(x) + \sum_{i=1}^{k} \lambda_i f_i(x) + \sum_{j=1}^{m} \mu_j g_j(x) \tag{265}$$

This is because $\lambda_i \geq 0$, $f_i \leq 0$, which makes $\sum \lambda_i f_i \leq 0$, and because $\sum \mu_j g_j = 0$. Since the quantity on the right hand side is the Lagrangian, we have that:

$$f(x) \geq \inf_{x \in dom(f_1) \cap \ldots \cap dom(g_m)} L(x, \lambda, \mu) = \nu(\lambda, \mu) \tag{266}$$

This holds for all primal feasible x and dual feasible (λ, μ). Thus, weak duality must hold.

12 Lecture 12: Lagrange Duality II

First, we recall from the previous lecture that a nonlinear program is an optimization of the form:

$$v = \inf_{x \in \mathbb{E}^d} f_0(x) \tag{267}$$

$$s.t.\ f_i(x) \leq 0,\ i = 1, ..., k,\ g_j(x) = 0,\ j = 1, ..., m \tag{268}$$

Here, $f_0, ..., f_k, g_1, ..., g_m : \mathbb{E}^d \to \mathbb{R}$. The Lagrangian associated with this nonlinear program is:

$$L(x, \lambda, \mu) = f_0(x) + \sum_{i=1}^{k} \lambda_i f_i(x) + \sum_{j=1}^{m} \mu_j g_j(x) \tag{269}$$

And the dual function is:

$$\nu(\lambda, \mu) = \inf_{x \in dom(f_0) \cap \ldots \cap dom(g_m)} L(x, \lambda, \mu) \tag{270}$$

Putting these all together, the Lagrange dual problem is:

$$v_L^* = \sup_{\lambda \in \mathbb{R}^k, \mu \in \mathbb{R}^m} \nu(\lambda, \mu) \tag{271}$$

$$s.t. \ \lambda \in \mathbb{R}_+^m \tag{272}$$

Last lecture, we showed that weak duality holds for the Lagrange dual problem, which implies:

$$v_L^* \leq v \tag{273}$$

In this lecture, we'll consider the conditions under which *strong duality*, $v_L^* = v$, holds.

Theorem 6 ***Strong Duality of the Lagrange Duality Scheme***
Consider a nonlinear program:

$$v = \inf_{x \in \mathbb{E}^d} f_0(x) \tag{274}$$

$$s.t. \ f_i(x) \leq 0, \ i = 1, ..., k, \ \langle a^{(j)}, x \rangle + b_j = 0, \ j = 1, ..., m \tag{275}$$

Here, $f_0, f_1, ..., f_k : \mathbb{E}^d \to \mathbb{R}$ with domains $dom(f_i) \subset \mathbb{E}^d$ and $a^{(j)} \in \mathbb{E}^d, b_j \in \mathbb{R}$ for $j = 1, ..., m$. Suppose the following conditions hold:

1. *$f_0, ..., f_k$ are convex functions.*
2. *$a^{(1)}, ..., a^{(m)}$ are linearly independent.*
3. *There exists an $\overline{x} \in int(dom(f_0) \cap ... \cap dom(f_k))$ such that $f_i(\overline{x}) < 0, i = 1, ..., k$, and $\langle a^{(j)}, \overline{x} \rangle + b_j = 0, j = 1, ..., m$. These two conditions are collectively called **Slater's conditions**.*

Then, we have that the Lagrange duality scheme exhibits strong duality, i.e. $v_L^* = v$. Moreover, provided v is finite, there exists an optimal solution for the Lagrange dual problem.

Before we prove this theorem, let's make a couple of notes about what it states. First, notice that the third condition for strong duality, regarding existence of an $\overline{x}$, essentially states that the constraint set is feasible in a "non-degenerate" way - there are feasible points in its interior.
Secondly, note that under strong duality, there exists an optimal solution for the Lagrange dual problem *even if* there is no optimal solution to the primal.
Thirdly, note that the theorem above simply provides sufficient conditions for strong duality - these are *not* necessary.
As a quick side note, let's think about why the equality constraints in the problem formulation have an explicit form. We know that we can write down any

equality constraint using inequalities as $-f(x) \leq 0$ and $f(x) \leq 0$. However, for strong duality, we require both $-f$ and f to be convex functions! If we consider the set of functions where $-f$ and f are both convex, we find that we are restricted to the set of affine functions. This is what provides the explicit affine form of the inequality constraints in the problem above. When a nonlinear program is written in this form, it is called a **convex nonlinear program**.
Let's now work on developing a proof for this theorem.

Proof: Recall that the Lagrange duality scheme entails separating the two sets:

$$S = \{(t,u) \in \mathbb{R}^{k+1} \times \mathbb{R}^m | \; \exists x \in \mathbb{E}^d \; s.t. \; (x,t_0) \in epi(f-\gamma), \tag{276}$$

$$(x_i, t_i) \in epi(f_i) \; i = 1, ..., k, \tag{277}$$

$$\langle a^{(i)}, x\rangle + b_j = u_j \; j = 1, ..., m\} \tag{278}$$

$$T = \{(t,u) \in \mathbb{R}^{k+1} \times \mathbb{R}^m | t_0 \leq 0, t_{1:k} \leq 0, u = 0\} \tag{279}$$

Where $\gamma \leq v$. To separate these two sets, we need to show that there exists a hyperplane, given by a normal $(\lambda, \mu) \in \mathbb{R}^{k+1} \times \mathbb{R}^m$ with $\lambda_0 \neq 0$ and translate $s \in \mathbb{R}$, such that:

$$S \subset \{(t,u) | \; \langle \lambda, t\rangle + \langle \mu, \lambda\rangle \geq \delta\} \tag{280}$$

$$T \subset \{(t,u) | \; \langle \lambda, t\rangle + \langle \mu, \lambda\rangle < \delta\} \tag{281}$$

If we show this for the case where $\gamma = v$, where strong duality holds, then the proof of this theorem is complete! This is because this will show that the conditions proposed in the theorem allow for separation of S and T in the case of strong duality.
First, we'll show that S and T are separated by a hyperplane. Since $f_0, ..., f_k$ are convex, we know that S is convex. This is because the set is convex with respect to (x,t,u), so we can "project out" x to get that S is convex in t and u. By definition of T, we also know that T is convex.
We also know that $S \cap T = \emptyset$. This comes from construction of the Lagrange duality scheme - that for any $\gamma \leq v$, S and T are disjoint.
Therefore, by the separation theorem, we know there exists a hyperplane given by normal $(\lambda, \mu) \in \mathbb{R}^{k+1} \times \mathbb{R}^m \backslash \{0\}$ and a $\delta \in \mathbb{R}$ such that:

$$S \subset \{(t,u) | \; \langle \lambda, t\rangle + \langle \mu, \lambda\rangle \geq \delta\} \tag{282}$$

$$T \subset \{(t,u) | \; \langle \lambda, t\rangle + \langle \mu, \lambda\rangle \leq \delta\} \tag{283}$$

Now, we will show that the inequality on the T halfspace is strict. Suppose that $\lambda_0 \neq 0$. Then, we have, by our earlier discussion of Lagrange duality, that $\lambda_0 = 1$ without loss of generality. Then, by the definition of T, since $t_0 < 0$, we can conclude that:

$$T \subset \{(t,u) | t_0 + \langle \lambda_{1:k}, t_{1:k}\rangle < \delta\} \tag{284}$$

Since we know that t_0 is strictly less than 0. So, if $\lambda_0 \neq 0$, we have our desired condition!
Let's show that $\lambda_0 \neq 0$ in all cases. Suppose for contradiction that $\lambda_0 = 0$. Then, we must have that:

$$(\lambda_{1:k}, \mu) \in \mathbb{R}^k \times \mathbb{R}^m \backslash \{0\} \tag{285}$$

This is because the entire λ cannot be the zero vector. Next, we have from $T \subset \{(t, u) | \langle \lambda_{1:k}, t_{1:k} \rangle + \langle \mu, u \rangle \leq \delta\}$ that $\lambda_{1:k} \geq 0$. This is because $t_{1:k}$ can be arbitrarily negative, and we must bound the left hand side above by some δ. Also, we know that $\delta \geq 0$, since if $\delta < 0$, we would not be able to satisfy the inequality.
Applying these two conditions to the halfspace condition involving S, we have that:

$$\sum_{i=1}^{k} \lambda_i f_i(x) + \sum_{j=1}^{m} \mu_j (\langle a^{(j)}, x \rangle + b_j) \geq \delta \; \forall x \in dom(f_0) \cap ... \cap dom(f_k) \tag{286}$$

Then, there exists an $\overline{x} \in int(dom(f_0) \cap ... \cap dom(f_k))$ such that $f_i(\overline{x}) < 0$ and $\langle a^{(j)}, \overline{x} \rangle + b_j = 0$ for all i and j.
But, since $\delta \geq 0$ and the sum of the f_i's and inner products must be ≤ 0, we have that $\lambda_{1:k} = 0$ if $\lambda_0 = 0$.
Now, we will look at μ. We must have that $\mu \neq 0$, since $\lambda = 0$. Then, we are left with:

$$\langle \sum_{j=1}^{m} \mu_j a^{(j)}, x \rangle + \langle \mu, b \rangle \geq \delta \; \forall x \in dom(f_0) \cap ... \cap dom(f_k) \tag{287}$$

If $\mu \neq 0$ and each $a^{(j)}$ are linearly independent, then $\sum \mu_j a^{(j)} \neq 0$. So, this sum can define a valid normal vector for a hyperplane! This tells us that the conditions above define a valid hyperplane such that $dom(f_0) \cap ... \cap dom(f_k)$ lies in one halfspace.
However, since $\overline{x}$ satisfies $\langle a^{(j), \overline{x}} \rangle + b_j = 0 \; \forall j$ from the equality constraints on the optimization, we know that $\overline{x}$ *must* be on the hyperplane. But, by our earlier development, it must also be in the interior of $dom(f_0) \cap ... \cap dom(f_k)$. Since it cannot be in the interior of this set *and* on the hyperplane, we have reached a contradiction.
This provides the final piece of the strong duality theorem. □

13 Lecture 13: Optimality Conditions for NLPs

Consider a nonlinear program:

$$v = \inf_{x \in \mathbb{E}^d} f(x) \tag{288}$$

$$s.t. \; f_i(x) \leq 0, \; i = 1, ..., k \tag{289}$$

$$g_j(x) = 0, \; j = 1, ..., m \tag{290}$$

Recall that the Lagrangian associated with this nonlinear program is:

$$L(x, \lambda, \mu) = f_0(x) + \sum_{i=1}^{k} \lambda_i f_i(x) + \sum_{j=1}^{m} \mu_j g_j(x) \tag{291}$$

While the dual function is:

$$\nu(\lambda, \mu) = \inf_{x \in \mathbb{E}^d} L(x, \lambda, \mu) \tag{292}$$

Using these two functions, we define the Lagrange dual of the primal problem as:

$$v_L^* = \sup_{\lambda \in \mathbb{R}^k, \mu \in \mathbb{R}^m} \nu(\lambda, \mu) \ s.t. \ \lambda \in \mathbb{R}_+^n \tag{293}$$

Last time, we proved that under the following conditions, the Lagrange dual is a *strong dual*:

1. **Convex NLP:** $f_0, ..., f_k$ are convex and $g_j(x) = \langle a^{(j)}, x \rangle + b_j, j = 1, ..., m$.
2. **Slater's Conditions:** There exists $\overline{x} \in int(dom(f_0) \cap ... \cap dom(f_k)$ such that $f_i(\overline{x}) < 0, \ i = 1, ..., k,$ and $\langle a^{(j)}, \overline{x} \rangle + b_j = 0, \ j = 1, ..., m$, where the set of $a^{(j)}$ is linearly independent.

Now that we've established these conditions, we can derive some optimality conditions for convex nonlinear programs. Consider an arbitrary convex nonlinear program satisfying Slater's conditions, and suppose that:

$$\hat{x} \in int(dom(f_0) \cap ... \cap dom(f_k)) \tag{294}$$

Is an optimal solution to the primal problem. By weak duality, we know that for an optimal solution $(\hat{\lambda}, \hat{\mu})$ of the dual problem, we have that:

$$f_0(\hat{x}) \geq f_0(\hat{x}) + \sum_{i=1}^{k} \hat{\lambda}_i f_i(\hat{x}) + \sum_{j=1}^{m} \hat{\mu}_j g_j(\hat{x}) \tag{295}$$

$$= L(\hat{x}, \hat{\lambda}, \hat{\mu}) \tag{296}$$

$$\geq \inf_{x \in dom(f_0) \cap ... \cap dom(f_k)} L(x, \hat{\lambda}, \hat{\mu}) \tag{297}$$

$$= \nu(\hat{\lambda}, \hat{\mu}) \tag{298}$$

Note that we know an optimal $(\hat{\lambda}, \hat{\mu})$ exists because Slater's conditions hold. Since the problem is convex and Slater's conditions hold, we know that strong duality holds! Thus, we have that:

$$f_0(\hat{x}) = \nu(\hat{\lambda}, \hat{\mu}) \tag{299}$$

Therefore, the two inequalities in the preceding reasoning hold with equality! This allows us to conclude from the chain of equalities that:

$$\sum_{i=1}^{k} \hat{\lambda}_i f_i(\hat{x}) = 0 \tag{300}$$

Under what conditions does this hold? Let's think about the different implications of this constraint, bearing in mind that $\lambda_i \in \mathbb{R}^n_+$. Since $\lambda_i \in \mathbb{R}^n_+$, and $f_i(\hat{x}) \leq 0$ by definition of the inequality constraint, we know that:

$$\hat{\lambda}_i f_i(\hat{x}) = 0, i = 1, ..., k \tag{301}$$

Since if this were not the case, the sum of $\hat{\lambda}_i f_i(\hat{x})$ would not equal zero. What implications does the sign of the inequality constraint have on this result?
First, consider the case where $f_i(\hat{x}) < 0$. In this case, we will have that $\hat{\lambda}_i = 0$. On the other hand, if $\hat{\lambda}_i = 0$, then $f_i(\hat{x}) = 0$, so the constraint will be tight.
Now, we consider a second consequence of the chain of inequalities from strong duality. Because the inequalities are satisfied with equality, we require that $\hat{x}$ minimizes $L(x, \hat{\lambda}, \hat{\mu})$ over $dom(f_0) \cap ... \cap (f_k)$. For a fixed $\hat{\lambda}, \hat{\mu}$, we know that $L(x, \hat{\lambda}, \hat{\mu})$ is a convex function for a convex NLP. Thus, because $\hat{x}$ is a minimizer, 0 must be in its subdifferential.
Putting all of these conditions together, we get the following optimality conditions for primal optimal $\hat{x}$ and dual optimal $(\hat{\lambda}, \hat{\mu})$:

Definition 21 ***Karush-Kuhn-Tucker (KKT) Conditions***
Any $\hat{x}, (\hat{\lambda}, \hat{\mu})$ satisfying the following conditions will be optimal for a convex nonlinear program:

1. *Primal Feasibility:* $f_i(\hat{x}) \leq 0, i = 1, ..., k,\ \langle a^{(j)}, x \rangle + b_j = 0, j = 1, ..., m.$
2. *Dual Feasibility:* $\hat{\lambda}_i \in \mathbb{R}^k_+$.
3. *Complementary Slackness:* $\hat{\lambda}_i \cdot f_i(\hat{x}) = 0,\ i = 1, ..., k.$
4. *Subdifferential:* $0 \in \partial_x L(x, \hat{\lambda}, \hat{\mu})$ *at* $\hat{x}$.

So, suppose we have a convex NLP satisfying Slater's conditions. If $\tilde{x} \in int(dom(f_0) \cap ... \cap dom(f_k))$ and $(\tilde{\lambda}, \tilde{\mu}) \in \mathbb{R}^k \times \mathbb{R}^m$ satisfy the KKT conditions, then $\tilde{x}$ is primal optimal and $(\tilde{\lambda}, \tilde{\mu})$ are dual optimal. Notice that "*int*(...)" is only used here since the subdifferential of a function, as required in the KKT conditions, is only defined on the interior of a function's domain.

Comparing & Contrasting Fenchel and Lagrange Duality

Consider the following optimization problem:

$$v = \inf_{x \in \mathbb{R}^d} \langle x, Mx \rangle \ \ s.t. x \in S \tag{302}$$

Here, let $S := \{x \in \mathbb{R}^d | \ ||x||_{l_2} \leq 1\}$ and $M \in \mathbb{S}^d$. We know that the primal solution, v, is equal to $v = \min\{0, \lambda_{min}(M)\}$. What are v_F^* and v_L^*, the optimal values of the Fenchel and Lagrange dual problems?
First, let's consider v_F^*. We know that if $M \succeq 0$, then the objective function of the primal is convex, which implies strong duality. This provides $v_F^* = v$. On the other hand, if M is *not* PSD, we can show that the Fenchel dual gives $v_F^* = -\infty$.
Now, consider the Lagrange dual. For v_L^*, we once again have that $v_L^* = v$ when $M \succeq 0$, as the objective will be convex in this case. If M is not PSD, then we don't have a convex problem. Let's derive the Lagrangian and dual function for this case!

$$L(x, \lambda) = \langle x, Mx \rangle + \lambda(||x||_{l_2} - 1) \tag{303}$$

$$\nu(\lambda) = \inf_{x \in \mathbb{R}^d} \langle x, Mx \rangle + \lambda(||x||_{l_2} - 1) = -\infty \tag{304}$$

We get $-\infty$ for the dual function since we can make $\langle x, Mx \rangle$ go to $-\infty$ quadratically fast in x, while $||x||_{l_2}$ only grows linearly in x. So, just like the Fenchel dual, the Lagrange dual problem also has $v_L^* = \infty$ for any $\lambda \geq 0$.
Let's adjust our statement of the problem slightly. Consider the equivalent primal problem:

$$v = \inf_{x \in \mathbb{R}^d} \langle x, Mx \rangle \tag{305}$$

$$s.t. \ ||x||_{l_2}^2 \leq 1 \tag{306}$$

Notice that this is the *same* exact problem as the primal, just with a square in the constraint. With this reformulated constraint, we have that:

$$L(x, \lambda) = \langle x, Mx \rangle + \lambda(\langle x, x \rangle - 1) \tag{307}$$

$$\nu(\lambda) = \inf_{x \in \mathbb{R}^d} \langle x, (M + \lambda I)x \rangle - \lambda \tag{308}$$

$$= \begin{cases} -\lambda, \ M + \lambda I \succeq 0 \\ -\infty, \ o.w. \end{cases} \tag{309}$$

Now that we have a finite dual function for a range of λ, we may formulate the Lagrange dual problem:

$$v_L^* = \sup_{\lambda \in \mathbb{R}} -\lambda \tag{310}$$

$$s.t. \ M + \lambda I \succeq 0, \ \lambda \geq 0 \tag{311}$$

$$= \min\{0, \lambda_{min}(M)\} \tag{312}$$

Here, 0 is the optimal value when M is PSD, and $\lambda_{min}(M)$ is the optimal value otherwise. This is because it is the smallest λ such that $M + \lambda I \succeq 0$. So, we see that for this reformulated constraint, we get the *same* result as the primal problem! Interestingly, this tells us that when we write the constraint differently for the Lagrangian, we get a different optimal solution.
Let's summarize some differences between the Fenchel and Lagrange duality schemes using this example. The Fenchel duality scheme requires less information regarding the description of the constraint set, as it simply relies on a support function. The Lagrange duality scheme may require more information about the constraint set or a different problem description *but* this extra information can be used to get a better lower bound.

14 Lecture 14: Lagrangian Relaxation

In this lecture, we'll discuss a Lagrangian perspective on conic duality. Recall that a cone program is an optimization of the form:

$$v = \inf_{x \in \mathbb{E}^d} \langle c, x \rangle \tag{313}$$

$$s.t.\ Ax + b \in \mathcal{K} \tag{314}$$

Where $A : \mathbb{E}^d \to \mathbb{E}^n$ is a linear map, $b \in \mathbb{E}^n$, and $K \subset \mathbb{E}^n$ is a convex cone.
To derive a dual of this problem using the Lagrange duality scheme, we'll introduce an additional variable, just as we did before with the Fenchel duality scheme. Consider the reformulation:

$$v = \inf_{x \in \mathbb{E}^d, y \in \mathbb{E}^n} \langle (c, 0), (x, y) \rangle \tag{315}$$

$$s.t.\ Ax + b - y = 0 \tag{316}$$

$$(x, y) \in \mathbb{E}^d \times \mathcal{K} \tag{317}$$

This problem can be reformulated as:

$$v = \inf_{x \in \mathbb{E}^d, y \in \mathbb{E}^n} f_0(x, y) \tag{318}$$

$$s.t. Ax + b - y = 0 \tag{319}$$

Here, $f_0(x, y) = \langle c, x \rangle$ with $dom(f_0) = \mathbb{E}^d \times \mathcal{K}$ - we simply incorporate the second constraint into the domain of the objective to achieve this reformulation.
Now, we'll compute the Lagrangian and dual function:

$$L(x, y, \mu) = \langle (c, 0), (x, y) \rangle + \langle \mu, Ax + b - y \rangle \tag{320}$$

$$\nu(\mu) = \inf_{(x,y) \in \mathbb{E}^d \times \mathcal{K}} L(x, y, \mu) \tag{321}$$

When determining the dual function, it'll be instructive to *only* eliminate the y variable (the extra variable we added to the reformulation) first. This will eliminate the extra variable we added in and bring the problem back to its original set of variables. We define:

$$\overline{L}(x,\mu) = \inf_{y\in\mathcal{K}} L(x,y,\mu) \tag{322}$$

$$= \inf_{y\in\mathcal{K}} \langle (c,0),(x,y)\rangle + \langle \mu, Ax+b-y\rangle \tag{323}$$

$$= \left[\inf_{y\in\mathcal{K}} \langle \mu,-y\rangle\right] + \langle c,x\rangle + \langle \mu, Ax+b\rangle \tag{324}$$

$$= \langle c,x\rangle + \langle \mu, Ax+b\rangle + \begin{cases} 0, \ \mu\in\mathcal{K}^\circ \\ -\infty, \ o.w. \end{cases} \tag{325}$$

This observation suggests a natural Lagrangian for a cone program! $\overline{L}$ suggests that μ should be in the polar cone of $\mathcal{K}$. So, we may now get rid of our extra variable y and write down the Lagrangian:

$$L(x,\mu) = \langle c,x\rangle + \langle \mu, Ax+b\rangle\,, \ \mu\in\mathcal{K}^\circ \tag{326}$$

When $\mu\in\mathcal{K}^\circ$ is satisfied, then we get a Lagrangian *just* in terms of μ, x. With this Lagrangian, the dual function becomes:

$$\nu(\mu) = \inf_{x\in\mathbb{E}^d} \langle c,x\rangle + \langle \mu, Ax+b\rangle \tag{327}$$

$$\nu(\mu) = \begin{cases} \langle b,\mu\rangle\,, \ A^*\mu + c = 0 \\ -\infty, \ o.w. \end{cases} \tag{328}$$

The Lagrange dual problem is then:

$$v_L^* = \sup_{\mu\in\mathbb{E}^n} \langle b,\mu\rangle \tag{329}$$

$$s.t.\ A^*\mu + c = 0, \mu\in\mathcal{K}^\circ \tag{330}$$

To see how flexible this approach is, consider the following problem:

$$v = \inf_{x\in\mathbb{E}^d} \langle c,x\rangle \tag{331}$$

$$s.t.\ A^{(1)}x + b^{(1)} \in \mathcal{K}^{(1)} \tag{332}$$

$$\vdots \tag{333}$$

$$A^{(m)}x + b^{(m)} \in \mathcal{K}^{(m)} \tag{334}$$

We can write the following Lagrangian:

$$L(x,\mu^{(1)},...,\mu^{(m)}) = \langle c,x\rangle + \sum_{j=1}^{m} \langle \mu_j, A^{(j)}x + b^{(j)}\rangle\,, \ \mu^{(j)}\in\mathcal{K}_j^\circ \tag{335}$$

We define the dual function accordingly as:

$$\nu(\mu^{(1)},...,\mu^{(m)}) = \begin{cases} \sum_{j=1}^m \langle \mu^{(j)}, b^{(j)}\rangle\,, \ c + \sum_{j=1}^m A^{(j)*}\mu^{(j)} = 0 \\ -\infty, \ o.w. \end{cases} \tag{336}$$

The Lagrange dual problem is then:

$$v_L^* = \sup_{\mu^{(1)},...,\mu^{(m)}} \sum_{j=1}^{m} \langle \mu^{(j)}, b^{(j)} \rangle \tag{337}$$

$$s.t\ \mu^{(j)} \in \mathcal{K}_j^{\circ},\ j = 1, ..., m \tag{338}$$

$$c + \sum_{j=1}^{m} A^{(j)*} \mu^{(j)} = 0 \tag{339}$$

This gives a simple resolution of the dual in the case where there are multiple conic constraints.

Lagrangian Relaxation for Non-Convex NLPs

We now consider a non-convex nonlinear program of the form:

$$v = \inf_{x \in \mathbb{E}^d} f_0(x) \tag{340}$$

$$s.t.\ f_i(x) \leq 0,\ i = 1, ..., k \tag{341}$$

$$g_j(x) = 0,\ j = 1, ..., m \tag{342}$$

The Lagrangian and dual function are defined according to the standard definitions as:

$$L(x, \lambda, \mu) = f(x) + \sum_{j=1}^{k} \lambda_j f_j(x) + \sum_{j=1}^{m} \mu_j g_j(x) \tag{343}$$

$$\nu(\lambda, \mu) = \inf_{x \in dom(f_0) \cap ... \cap dom(g_m)} L(x, \lambda, \mu) \tag{344}$$

So, for each fixed x, we observe that $L(x, \lambda, \mu)$ is affine in λ, μ. Since the pointwise minimum of a set of affine functions is *concave* (i.e. the negative of the function is convex), we conclude that $\nu(\lambda, \mu)$ is concave. This implies that the Lagrange dual problem, which is the maximization of a concave problem, is a convex optimization problem.
Lagrangian relaxation builds on this observation by formulating lower bounds on the optimal value of an NLP (that may or may not be convex) via a convex optimization problem using Lagrange duality.
Consider the following example:

$$v = \inf_{x \in \mathbb{R}} -x^2 \tag{345}$$

$$s.t.\ -1 \leq x \leq 1 \tag{346}$$

This is *not* a convex NLP. We compute the optimal value of the problem as $v = -1$.
The Lagrangian for this problem is:

$$L(x, \lambda_1, \lambda_2) = -x^2 + \lambda_1(-x - 1) + \lambda_2(x - 1),\ \lambda_1, \lambda_2 \geq 0 \tag{347}$$

Optimizing for the dual function, we get:

$$\nu(\lambda_1, \lambda_2) = -\infty \tag{348}$$

Since x can be made arbitrarily small in the dual function optimization. This implies that $v_L^* = -\infty$ - not a useful lower bound!
To make the lower bound tighter, we add the *redundant* constraint $1 - x^2 \geq 0$ to the NLP. Our problem is now:

$$v = \inf_{x \in \mathbb{R}} -x^2 \tag{349}$$

$$s.t. \; -1 \leq x \leq 1, \; 1 - x^2 \geq 0 \tag{350}$$

Notice how this new constraint doesn't impact the original constraint set - it is already implied by the original two constraints. The Lagrangian of this reformulated problem is:

$$L(x, \lambda_1, \lambda_2, \lambda_3) = -x^2 + \lambda_1(-x-1) + \lambda_2(x-1) + \lambda_3(x^2 - 1) \tag{351}$$

Where $\lambda_1, \lambda_2, \lambda_3 \geq 0$. The dual function is then computed:

$$\nu(\lambda_1, \lambda_2, \lambda_3) = \inf_{x \in \mathbb{R}} L(x, \lambda_1, \lambda_2, \lambda_3) \tag{352}$$

$$= \inf_{x \in \mathbb{R}} -x^2 + \lambda_1(-x-1) + \lambda_2(x-1) + \lambda_3(x^2-1) \tag{353}$$

By noting that $\nu(0, 0, 1) = -1$, we can check that under this dual function, $v_L^* = -1$. Thus, we see that this dual is actually a strong dual! This problem encapsulates the idea of Lagrangian relaxation - by adding redundant constraints, we can get a tighter lower bound on the optimal value of our primal problem.

15 Lecture 15: Mixed Integer Programs I

In this lecture, we'll introduce the concept of integer and mixed integer linear programming. A **(mixed) integer linear program (MILP/ILP)** is an optimization problem with a linear objective and constraint, and integer and real variables. The typical form of such a problem is:

$$v = \inf_{x \in \mathbb{Z}^d, y \in \mathbb{R}^k} \langle c, (x, y) \rangle \tag{354}$$

$$s.t. \; A(x, y) \leq b \tag{355}$$

Here, $c \in \mathbb{R}^{d+k}$, $A : \mathbb{R}^{d+k} \to \mathbb{R}^n$ is a linear map and $b \in \mathbb{R}^n$. Note that here, we'll *only* consider problems with linear objective and constraint. More general problems with integers (even those with quadratic objectives), are typically undecidable.
Should we expect to be able to solve mixed integer linear programs? Let's think of a typical example: the Knapsack problem. This is a mixed integer linear program that optimizes over binary variables. As the Knapsack problem can be

proven to be NP-Hard, we should *not* expect to be able to solve integer linear programs in polynomial time.

Despite this setback, we can still reformulate integer linear programs as convex optimization problems! Recall that we may equivalently formulate linear programming problems as optimizations over the convex hull of their constraint sets. With this in mind, we may reformulate integer linear programs as:

$$v = \inf_{z \in \mathbb{R}^{d+k}} \langle c, z \rangle \tag{356}$$

$$s.t.\ z \in conv\{(x, y) \in \mathbb{R}^d \times \mathbb{R}^k \mid A(x, y) \leq b, x \in \mathbb{Z}^d\} \tag{357}$$

Despite this convex reformulation, due to the undecidable nature of these problems, we should *not* expect to be able to compute the convex hull efficiently.

Instead, we'll aim to identify subclasses of integer linear programs that have a constraint set whose convex hull can be described efficiently. More generally, we can use a method called **branch and bound** to solve integer linear programs that don't have an efficient convex description.

Let's determine which integer linear programs have an efficient convex description.

Theorem 7 ***Fundamental Theorem of Integer Programming (Meyer)*** *Consider a (mixed) integer linear set:*

$$S = \{(x, y) \in \mathbb{R}^d \times \mathbb{R}^k \,|A(x, y) \leq b,\ x \in \mathbb{Z}^d\} \tag{358}$$

If $A \in \mathbb{Z}^{n \times (d+k)}$ *and* $b \in \mathbb{Z}^n$*, then* $conv(S)$ *is a polyhedron, and may therefore be described as a finite number of halfspaces.*

Let's make a few remarks about this theorem. First, notice that a finite number of halfspaces *could* imply that the number of halfspaces required is exponential in the dimension of the problem! This theorem says nothing about the NP-hard nature of a mixed integer program.

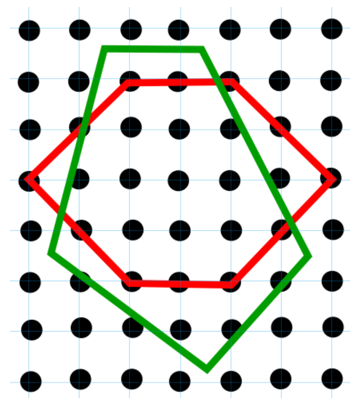
The conclusion of this theorem *does not* hold if A and b do not have integer entries. This result states that integer linear programs with integral problem parameters are *decidable*, although they are still NP-hard in general.

Now, we ask the following important question: when is the convex hull of an integer linear set with integral problem parameters a polyhedron that may be efficiently described? We'll answer this question for the case of "pure" integer linear sets, rather than the more general mixed integer case.

Consider an integer set:

$$S = \{x \in \mathbb{R}^d \mid Ax \leq b,\ x \in \mathbb{Z}^d\} \tag{359}$$

How can we reason about the existence of an efficient description of the convex hull of S? Consider the following visualization, where we plot a polyhedral set on top of a grid of integer vectors.



First, consider the polygon sketched in red. Let's think about the integer vectors contained in the polygon, i.e. the grid points that are contained in the polygon. If we take the convex hull of these points, it's clear that we get back the original hexagon shape! However, this is *not* the case for the green polygon, as the convex hull of the green polygon is not generated by its integer points.
This sketch leads us to the following definition:

Definition 22 ***Integral Set***
A convex set $G \subset \mathbb{R}^d$ is called integral if:

$$G = conv(G \cap \mathbb{Z}^d) \tag{360}$$

In words, a convex set is said to be integral if it is the convex hull of the integer points contained in the set. In particular, we'll be interested in classifying **integral polyhedra**, integral sets of the form $\{x \mid Ax \leq b\}$ that are defined by $A \in \mathbb{Z}^{n\times d}, b \in \mathbb{Z}^n$.
Notice that even if $A \in \mathbb{Z}^{n\times d}$ and $b \in \mathbb{Z}^d$, the set $\{x \mid Ax \leq b\}$ need not be an integral polyhedron! For instance, consider the set:

$$\{x \in \mathbb{R} \mid 0 \leq 2x \leq 1\} \tag{361}$$

Even though this set has integral parameters, it is *not* an integral polyhedron, as it is not equal to the convex hull of the integers contained inside it. This is because the right endpoint of this set is given by 1/2.

Proposition 7 ***Extreme points of Integral Polyhedra***
Consider a bounded polyhedron $P = \{x \in \mathbb{R}^d | Ax \leq b\}$ with $A \in \mathbb{Z}^{n\times d}, b \in \mathbb{Z}^n$. P is an integral polyhedron if and only if the extreme points of P are in $\mathbb{Z}^d$.

This result follows from the convex hull of a polyhedron returning the polyhedron in the case that P is bounded. Now, we state the following important result regarding extreme points of polyhedra.

Lemma 2 ***Extreme Points of Polyhedra***
Consider a polyhedron $P = \{x|Ax \leq b\} \subset \mathbb{R}^d$. A point $x \in \mathbb{R}^d$ is an extreme point of P if and only if it satisfies with equality a collection of d linearly independent inequalities of the system $Ax \leq b$.

Let's apply this lemma to the situation at hand. By this lemma, we know that an extreme point $\overline{x}$ of an integral polyhedron P must satisfy:

$$\overline{A}\overline{x} = \overline{b} \tag{362}$$

Where $\overline{A} \in \mathbb{Z}^{d\times d}$ is a nonsingular (due to linear independence from the lemma) $d \times d$ submatrix of A and $\overline{b} \in \mathbb{Z}^d$ is the corresponding set of entries of b. Since $\overline{A}$ is nonsingular, we may identify $\overline{x}$ using:

$$\overline{x} = \overline{A}^{-1}\overline{b} \tag{363}$$

Thus, if the inverse of $\overline{A}$ satisfies $\overline{A}^{-1} \in \mathbb{Z}^{d\times d}$, then $\overline{x} \in \mathbb{Z}^d$. How do we know if the inverse of a matrix will have integer entries? Consider the following result.

Lemma 3 ***Integral Matrix Inverse***
Consider a nonsingular matrix $M \in \mathbb{Z}^{d\times d}$. If $\det(M) = \pm 1$, then $M^{-1} \in \mathbb{Z}^{d\times d}$.

Let's informally discuss why this is true. For necessity, let's proceed by contradiction. If $\det M \neq \pm 1$, then $\det M^{-1} = 1/\det M$ won't be an integer. This implies that M^{-1} is not an integer matrix, which is a contradiction. For sufficiency, we may make an argument using the adjoint formula for a matrix inverse. We now define a class of matrices whose square submatrices have integral inverses when they exist.

Definition 23 ***Unimodular Matrix***
A matrix $A \in \mathbb{Z}^{n\times d}$ with $d \leq n$ is called unimodular if every $d \times d$ submatrix of A has a determinant equal to 0, 1, or -1.

Notice that if A has a block with determinant 0, it won't contain a set of linearly independent points. Thus, by the result above, it won't generate an extreme

point, as an extreme point requires d linearly independent inequalities.
Putting all of these developments together, we arrive at the following result:

Proposition 8 ***Classification of Integral Polyhedra***
Consider a polyhedron $P = \{x \in \mathbb{R}^d | Ax \leq b\}$ with $A \in \mathbb{Z}^{n \times d}$ and $b \in \mathbb{Z}^n$. Suppose P is bounded. If A is unimodular, then P is an integral polyhedron:

$$P = conv(P \cap \mathbb{Z}^d) \tag{364}$$

16 Lecture 16: Mixed Integer Programs II

Last lecture, we introduced integer programming. Recall that the class of mixed integer linear programs is of the the form:

$$\inf_{x \in \mathbb{Z}^d, y \in \mathbb{R}^k} \langle c, (x, y) \rangle \tag{365}$$

$$s.t\ A(x, y) \leq b \tag{366}$$

Let's briefly recap our development of the geometry of integral constrain sets. For the convex hull of the constraint set to be a polyhedron, we required that $A \in \mathbb{Z}^{n \times (d+k)}$ and $b \in \mathbb{Z}^n$ due to Meyer's theorem.
A polyhedron $P = \{x \in \mathbb{R}^d | Ax \leq b\}$ with $A \in \mathbb{Z}^{n \times d}, b \in \mathbb{Z}^n$ is called integral if $P = conv(\mathbb{Z}^d \cap P)$. Assuming P is bounded, this corresponds to P having integer extreme points.
We showed that if A is unimodular, then P is an integral polyhedron. Recall that unimodularity of $A \in \mathbb{Z}^{n \times d}$, for $n \geq d$, is defined as every $d \times d$ submatrix of A having determinant equal to 0, 1, or -1.
In this lecture, we'll consider inequalities of polyhedra described in several different ways that arise commonly in applications.

Proposition 9 ***Integral Polyhedra (Veinott & Dantzig)***
Consider a polyhedron P specified as:

$$P = \{x \in \mathbb{R}^d | Mx = b,\ x \geq 0\} \tag{367}$$

For $M \in \mathbb{Z}^{k \times d}$ and $b \in \mathbb{Z}^k$ with $k < d$. Suppose that P is also bounded. If $M^* \in \mathbb{Z}^{d \times k}$ is unimodular and of rank k, then P is integral for all $b \in \mathbb{Z}^k$.

Before we prove this statement, note that its converse is also true! We simply won't state or prove it here. Let's now prove the proposition stated above.

Proof: First, we rewrite the set P. An equivalent description of P is:

$$P = \left\{ x \in \mathbb{R}^d \middle| \begin{bmatrix} M \\ -M \\ -I \end{bmatrix} x \leq \begin{bmatrix} b \\ -b \\ 0 \end{bmatrix} \right\} \tag{368}$$

Here, the block matrix on the left hand side is in $\mathbb{R}^{(2k+d)\times d}$ and the vector on the right hand side is in $\mathbb{R}^{2k+d}$.
We know that any extreme point $\overline{x} \in \mathbb{R}^d$ of P satisfies a system of the form:

$$\overline{A}\overline{x} = \overline{b} \tag{369}$$

For $\overline{A} \in \mathbb{Z}^{d\times d}$ a nonsingular submatrix of:

$$\begin{bmatrix} M \\ -M \\ -I \end{bmatrix} \tag{370}$$

And $\overline{b} \in \mathbb{Z}^d$ the corresponding sub-vector of:

$$\begin{bmatrix} b \\ -b \\ 0 \end{bmatrix} \tag{371}$$

What structure does $\overline{A}$ have? We know $\overline{A}$ may be broken down as follows:

$$\overline{A} = \begin{bmatrix} k \text{ rows of } \begin{bmatrix} M \\ -M \end{bmatrix} \\ d-k \text{ rows of } -I \end{bmatrix} \tag{372}$$

Let's use this structure to evaluate the determinant of $\overline{A}$. The determinant of $\overline{A}$ is equivalent to the determinant of a $k \times k$ submatrix of the k rows of $(M; -M)$, as when we take the determinant of $\overline{A}$, the identity matrix "drops out" via cofactor expansion.
This $k \times k$ block matrix will look *something* like:

$$\begin{matrix} \pm \\ \pm \\ \pm \end{matrix} \begin{bmatrix} & & \\ & M & \\ & & \end{bmatrix} \tag{373}$$

Where each row is either a row of M or its negation. Each row in this submatrix is linearly independent. By this representation above, we see that taking the determinant of a submatrix is the same as taking the determinant of a submatrix of M up to a sign change.
So, if $\overline{A}$ has this structure, then the determinant of $\overline{A}$ is the the *same* as the determinant of a $k \times k$ submatrix of M (up to sign). Therefore, $\det(\overline{A}) = \pm 1$, as M^* is unimodular.
As every extreme point of P must satisfy the equations $Mx = b$, we have that $\overline{A}$ must have the preceding form, which implies every $\overline{x}$ must be integral. Thus, the analysis above is sufficient for all cases. □

Definition 24 ***Totally Unimodular Matrix***
A matrix $M \in \mathbb{Z}^{p\times q}$ is called totally unimodular if every square submatrix of M has determinant 0*,* 1*, or* -1*.*

Note that this differs from a unimodular matrix in that we *no longer* restrict the dimension of our square submatrix - now, we consider *all possible* square submatrices. This means that all entries of a totally unimodular matrix must have value 0, 1, or -1, as each entry of a matrix is a 1×1 submatrix.
On the other hand, unimodular matrices may have values outside the set $\{0, 1, -1\}$. For instance, consider:

$$\begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix} \tag{374}$$

This matrix is unimodular despite having an entry equal to two. Let's see how we can augment the proposition above using the concept of a totally unimodular matrix.

Proposition 10 ***Integral Polyhedra (Hoffman & Kruskal)***
Consider a polyhedron P specified as:

$$P = \{x \in \mathbb{R}^d | Mx \leq b,\ x \geq 0\} \tag{375}$$

Here, $M \in \{0, 1, -1\}^{n\times d}$ and $b \in \mathbb{Z}^d$. Suppose P is bounded. If M is totally unimodular, then P is integral for all $b \in \mathbb{Z}^n$.

Note that as with the proposition above, the converse of this result is also true, but we won't state or prove it here.

Proof: We can reformulate the polyhedron P as:

$$P = \left\{ x \in \mathbb{R}^d \middle| \begin{bmatrix} M \\ -I \end{bmatrix} x \leq \begin{bmatrix} b \\ 0 \end{bmatrix} \right\} \tag{376}$$

Where the matrix on the left hand side is in $\mathbb{R}^{(d+n)\times d}$. Any $d \times d$ submatrix of this matrix has determinant equal to that of an $l \times l$ submatrix of M for $l \leq d$. As M is totally unimodular, this submatrix has a determinant of 0, 1, or -1. $\square$

Let's sketch out a second proof of this proposition that relies on the Veinott-Dantzig result stated above.

Proof Sketch: We can express P as the projection of the following polyhedron:

$$\left\{ \begin{bmatrix} x \\ y \end{bmatrix} \in \mathbb{R}^{d+n} \middle| Mx + y = b, \begin{bmatrix} x \\ y \end{bmatrix} \geq 0 \right\} \tag{377}$$

Here, we have used a *slack variable*, y, to rewrite the inequality constraint in the original definition of P as an equality constraint.
Now, we may appeal to the Veinott-Dantzig result and the assumption that M is totally unimodular to show that this "lifted" polyhedron is integral.
Finally, the elimination of the slack variable y by projection yields an integral polyhedron. □

17 Lecture 17: Numerical Methods I

This section was adapted from Yiheng Xie's notes. Thanks Yiheng!
Thus far in the course, we've been focused largely on developing the theory of convex optimization problems rather than the actual solution procedures. In this lecture, we discuss the basics of some numerical methods used to solve convex optimization problems.
Recall that the general formulation of a convex optimization problem is:

$$v = \inf_{x \in \mathbb{E}^d} f(x) \; s.t. \; x \in S \tag{378}$$

Where $f : \mathbb{E}^d \to \mathbb{R}$ is a convex function and $S \subseteq \mathbb{E}^d$ is a convex set.
Let's state a few assumptions about this convex optimization problem that we'll use to develop numerical solution procedures. First, we assume that we have access to the two attributes defined below:

Definition 25 ***Subgradient Oracle***
Given any $x \in int(dom(f))$, the subgradient oracle returns $g \in \partial f(x)$.

We can think of the subgradient oracle as "sampling" a subgradient from the subdifferential of the function f. Since f is convex, the subgradient oracle provides us with a global underestimate of the value of f at the point x.

Definition 26 ***Projection Oracle***
Consider a closed, convex set $S \subseteq \mathbb{E}^d$. Let $P_S(x)$ be the operator that projects $x \in \mathbb{E}^d$ onto S. The projection oracle takes in $x \in \mathbb{E}^d$ and returns $P_S(x)$.

Recall from earlier in the course that $P_S(x)$, the projection of x onto S, is the unique solution to the optimization problem:

$$\inf_{z \in S} ||x - z||^2 \tag{379}$$

In order for this projection to be well-defined, we require that the convex set S is closed. If we didn't have this requirement, the boundary ∂S of the set may not be entirely contained in S. This would lead to ill-posed projections.
Notice that P_S also gives us a way to identify membership of an arbitrary $x \in S$ in S, as $x = P_S(x)$ if and only if $x \in S$.
Assuming that we have access to the subgradient and projection oracles for our given convex optimization problem, we may employ a numerical method called the **projected subgradient method**, which we now define.

Definition 27 ***Projected Subgradient Method***
Let $x^{(k)}$ denote an estimate of the optimal solution to a convex optimization problem at a time step $k \in \mathbb{Z}_+$. The projected subgradient method update rule is defined:

$$x^{(k+1)} = P_S(x^{(k)} - \eta_k g^{(k)}) \tag{380}$$

Where $g^{(k)} \in \partial f(x^{(k)})$ and η_k is the step size.

Notice that the projected subgradient method is not a "descent" method! For instance, consider the optimization problem:

$$v = \inf_{x \in \mathbb{R}} |x| \tag{381}$$

At the value $x^{(k)} = 0$, the subgradient method would send $x^{(k+1)}$ away from the optimal value, different to what a descent method would do.
Under what conditions will the projected subgradient method converge? Consider the following theorem.

Theorem 8 ***Projected Subgradient Convergence***
Consider a convex optimization problem:

$$v = \inf_{x \in \mathbb{E}^d} f(x) \ s.t. \ x \in S \tag{382}$$

Where $f : \mathbb{E}^d \to \mathbb{R}$ is a convex function and $S \subseteq int(dom(f))$ is a closed convex set. Further, suppose $||g|| \leq L$ for all $g \in \partial f(x)$ and all $x \in S$.

If x^* is the optimal solution of the optimization problem, the iterates of the projected subgradient method with step size $\eta_k \geq 0$ satisfy:

$$\min_{k \in 0,\ldots,T} f(x^{(k)}) - v \leq \frac{||x^{(0)} - x^*||^2 + L^2 \sum_{k=0}^{T} \eta_k^2}{2 \sum_{k=0}^{T} \eta_t} \tag{383}$$

Before we prove this theorem, we make several notes about its structure. First, note that we simply use $f(x^{(k)}) - v$ instead of $|f(x^{(k)}) - v|$, since v gives the global minimum of the function and $f(x) \geq v \; \forall x \in S$. This allows us to drop the absolute value.

Next, let's discuss the conditions under which convergence is guaranteed. Based on the theorem, we see that one way to guarantee convergence is to choose η_k such that $\sum_{k=0}^{T} \eta_k^2$ is finite and $\sum_{k=0}^{T} \eta_k \to \infty$ as $T \to \infty$. This will guarantee that $f(x^{(k)})$ approaches v by the inequality provided above. A choice $\eta_k = 1/k$ is one example of a sequence that gives this behavior.

Notice that this is a very general method with very few assumptions. Outside of the general convex optimization setup, we've assumed a closed constraint set S and that all subgradients are bounded for all $x \in S$. The generality of this approach comes with the drawback that this method has a relatively slow rate of convergence.

Let's now turn our attention back to proving the theorem above. First, we introduce the following lemma:

Lemma 4 ***Projections are Contractive***
Let $S \subset \mathbb{E}^d$ be a closed, convex set. Then, the projection map P_S onto the convex set is contractive. That is:

$$||P_S(x) - P_S(y)|| \leq ||x - y|| \; \forall x, y \in \mathbb{E}^d \tag{384}$$

Now, we return to the proof of the convergence theorem.

Proof (Convergence Theorem): We begin by expanding the squared norm $||x^{(T+1)} - x^*||^2$. Using the definition of the projected subgradient method, we substitute for $x^{(T+1)}$ as follows:

$$||x^{(T+1)} - x^*||^2 = ||P_S(x^{(T)} - \eta_T g^{(T)}) - x^*||^2 \tag{385}$$

$$= ||P_S(x^{(T)} - \eta_T g^{(T)} - P_S(x^*)||^2 \tag{386}$$

Where we use the fact that $x^* \in S$ to conclude that $P(x^*) = x^*$. Now, using the contractive property of projections, we conclude:

$$||x^{(T+1)} - x^*||^2 \leq ||x^{(T)} - \eta_T g^{(T)} - x^*||^2 \tag{387}$$

Now, we expand the squared norm and apply the definition of the subgradient to form another inequality.

$$||x^{(T)} - \eta_T g^{(T)} - x^*||^2 \tag{388}$$
$$= ||x^{(T)} - x^*||^2 + \eta_T^2 ||g^{(T)}||^2 - 2\eta_T \left\langle g^{(T)}, x^{(T)} - x^* \right\rangle \tag{389}$$
$$\leq ||x^{(T)} - x^*||^2 + \eta_T^2 ||g^{(T)}||^2 - 2\eta_T (f(x^{(T)}) - f(x^*)) \tag{390}$$
$$\leq ||x^{(T)} - x^*||^2 + \eta_T^2 L^2 - 2\eta_T (f(x^{(T)}) - f(x^*)) \tag{391}$$

Where in the last step, we use that $||g|| \leq L \; \forall g \in \partial f(x)$. Next, we use the definition of the projected subgradient method to expand the first squared norm recursively until we work our way back to $x^{(0)}$. This yields the following inequality:

$$||x^{(T)} - x^*||^2 + \eta_T^2 L^2 - 2\eta_T (f(x^{(T)}) - f(x^*)) \tag{392}$$
$$\leq ||x^{(0)} - x^*||^2 + L^2 \sum_{k=0}^{T} \eta_k^2 - 2 \sum_{k=0}^{T} \eta_k (f(x^{(k)}) - f(x^*)) \tag{393}$$

Putting all of these inequalities together, we therefore have:

$$0 \leq ||x^{(T+1)} - x^*||^2 \tag{394}$$
$$\leq ||x^{(0)} - x^*||^2 + L^2 \sum_{k=0}^{T} \eta_k^2 - 2 \sum_{k=0}^{T} \eta_k (f(x^{(k)}) - f(x^*)) \tag{395}$$

Bringing the negative term to the left hand side of the first inequality, we then get:

$$2 \sum_{k=0}^{T} \eta_k (f(x^{(k)}) - f(x^*)) \leq ||x^{(0)} - x^*||^2 + L^2 \sum_{k=0}^{T} \eta_k^2 \tag{396}$$
$$2 \sum_{k=0}^{T} \eta_k \Big(\min_{k=\in 0,\ldots,T} f(x^{(k)}) - f(x^*) \Big) \leq ||x^{(0)} - x^*||^2 + L^2 \sum_{k=0}^{T} \eta_k^2 \tag{397}$$

Dividing both sides by $2 \sum_{k=0}^{T} \eta_k$, we arrive at the statement of the theorem. □

As we discussed above, choosing η_k such that $\sum_{k=0}^{T} \eta_k^2$ is finite and $\sum_{k=0}^{T} \eta_k^2 \to \infty$ as $T \to \infty$ yields convergence. Although this choice is sufficient for convergence, it's important to note that it is *not* necessary. The "best" choice of η_k is choosing $\eta_k \propto 1/\sqrt{k}$. This choice of step size can be shown to give the following relationship:

$$\frac{||x^{(0)} - x^*||^2 + L^2 \sum_{k=0}^{T} \eta_k^2}{2 \sum_{k=0}^{T} \eta_t} \propto \frac{||x^{(0)} - x^*||^2 + L^2 \log(T)}{\sqrt{T}} \tag{398}$$

If we recall that this fraction bounds the error between the iterative estimate of the optimal value and the true optimal value, we arrive at the following corollary of the theorem above:

Corollary 1 ***ϵ-Accurate Solutions with Projected Subgradient***
To obtain an ϵ-accurate solution of a convex optimization problem using the projected subgradient method, one may choose T according to the rule:

$$T \propto \frac{1}{\epsilon^2} \tag{399}$$

18 Lecture 18: Numerical Methods II

Last lecture, we discussed the projected subgradient method, which provided a very general framework for numerically solving constrained convex optimization problems. In this lecture, we'll focus on numerical methods for the unconstrained minimization of differentiable convex functions. Our focus will therefore be on the following optimization problem:

$$v = \inf_{x \in \mathbb{E}^d} f(x) \tag{400}$$

Where $f : \mathbb{E}^d \to \mathbb{R}$ is a differentiable convex function.
First, we'll discuss a numerical optimization technique called the method of steepest descent. This method involves finding the minimum of a function by following the function's direction of "steepest descent," a vector which depends on the gradient of the function and the geometry of a norm on $\mathbb{E}^d$.
To define the direction of steepest descent of a function f at a point $x \in \mathbb{E}^d$, we'll first need a vector called the **normalized steepest descent**. Fix a norm $||\cdot||$ on $\mathbb{E}^d$. The normalized steepest descent of f at a point $x \in \mathbb{E}^d$ is defined:

$$u_{nsd} = \text{arg min}\{\langle \nabla f(x), v \rangle \mid ||v|| \leq 1\} \tag{401}$$

Where "arg min" refers to the argument v that minimizes the quantity $\langle \nabla f(x), v \rangle$ over the search space $||v|| \leq 1$. The normalized steepest descent is the vector pointing in the "steepest downwards direction" of f at a point $x \in \mathbb{E}^d$.
We then define **steepest descent** of f at $x \in \mathbb{E}^d$ as the vector:

$$u = ||\nabla f(x)||_* u_{nsd} \tag{402}$$

Where $||\cdot||_*$ denotes the dual norm associated with $||\cdot||$.
Using these two vectors, we define the method of steepest descent as follows.

Definition 28 ***Method of Steepest Descent***
Let $||\cdot||$ be a fixed norm on $\mathbb{E}^d$ and $f : \mathbb{E}^d \to \mathbb{R}$ a differentiable convex function. The method of steepest descent proceeds according to the update rule:

$$x^{(k+1)} = x^{(k)} + s_k u^{(k)} \tag{403}$$

Where $u^{(k)}$ is defined as:

$$u^{(k)} = ||\nabla f(x^{(k)})||_* u_{nsd}^{(k)} \tag{404}$$

For $u_{nsd}^{(k)}$ the normalized steepest descent:

$$u_{nsd}^{(k)} = \arg\min\{\langle \nabla f(x^{(k)}, v\rangle \mid ||v|| \leq 1\} \tag{405}$$

Let's discuss some details of the setup of the method of steepest descent. First, we ask the question: why use the dual norm? When looking at the inner product:

$$\langle \nabla f(x^{(k)}), v\rangle \tag{406}$$

The vector v lives in the vector space $\mathbb{E}^d$, while the other lives in the dual space associated with $\mathbb{E}^d$. The representation of the inner product written above is actually an identification of the corresponding element of the dual space as a vector in $\mathbb{E}^d$. Thus, to be precise, we use the dual norm, $||\nabla f(x^{(k)}||_*$ when determining the direction of steepest descent.
Let's consider some examples of computing the direction of steepest descent $u^{(k)}$ under a few different norms.

1. $||\cdot||$ is the norm induced by the inner product:

$$||w|| = \sqrt{\langle w, w\rangle} \tag{407}$$

   In this case, $||\cdot||_* = ||\cdot||$, which means that the direction of steepest descent is given by:

$$u^{(k)} = -\nabla f(x^{(k)}) \tag{408}$$

   Thus, for this norm, the method of steepest descent specializes to **gradient descent**.

2. $||\cdot||$ is the norm defined by:

$$||w|| = \sqrt{\langle w, Pw\rangle} \tag{409}$$

   For $P \succ 0$. In this case, the dual norm becomes:

$$||w||_* = \sqrt{\langle w, P^{-1}w\rangle} \tag{410}$$

   The steepest descent is then given by:

$$u^{(k)} = -P^{-1}\nabla f(x^{(k)}) \tag{411}$$

3. $||\cdot||$ is the l_1 norm, and $\mathbb{E}^d = \mathbb{R}^d$. Recall that the l_1 norm on $\mathbb{R}^d$ is defined:

$$||w||_{l_1} = \sum_{i=1}^{d} |w_i| \tag{412}$$

In this case, the dual norm is given by:

$$||w||_* = ||w||_{l_\infty} = \max_{i \in 1,\ldots,d} |w_i| \tag{413}$$

How do we compute the steepest descent direction in this case? Let $j \in \{1, ..., d\}$ be the coordinate at which $\nabla f(x^{(k)})$ has the largest magnitude. Then, we have that:

$$u^{(k)} = -[\nabla f(x^{(k)})]_j \cdot e^{(j)} \tag{414}$$

Where $e^{(j)}$ is the j'th standard basis vector on $\mathbb{R}^d$. This specialization of steepest descent is called **coordinate descent**, and is commonly used in large scale optimization problems where we only wish to update a few components of $x^{(k)}$ at a time.

Now that we've discussed the basic setup of the method of steepest descent and methods for computing the direction of steepest descent, we must consider the choice of step size. We'll discuss two methods for defining step size, the first of which is defined below:

Definition 29 ***Exact Linesearch***
The exact linesearch method solves the following 1-D optimization problem to identify the stepsize s_k:

$$s_k = \text{arg min } \{f(x^{(k)} + su^{(k)}) | s \geq 0\} \tag{415}$$

Based on the definition, we observe that the exact linesearch method of determining stepsize chooses a stepsize s_k that that minimizes f in the direction of steepest descent starting from the current iterate $x^{(k)}$. Note that $s \geq 0$ since the "sign" of the direction we wish to move in has already been chosen by the direction of steepest descent $u^{(k)}$.
The exact linesearch method is the "best" thing we can do at a given point $x^{(k)}$, as it finds the largest step we can make towards the minimum of the function. However, since it computes s_k exactly, it can be computationally expensive.
As a less computationally expensive alternative, consider the backtracking linesearch, defined as follows:

Definition 30 ***Backtracking Linesearch***
Fix a number $\beta \in (0,1)$ and a positive initial guess for s. Then, execute the following algorithm to identify s_k:

1. *While $f(x^{(k)} + su^{(k)}) \geq f(x^{(k)})$:*

$$s \leftarrow \beta s \tag{416}$$

2. *Set $s_k = s$.*

Starting from an initial guess, the backtracking linesearch looks for a smaller and smaller stepsize until a stepsize is found that yields a smaller value of f. This is the stepsize that is then defined to be s_k. Note that the initial guess for the stepsize is not of high importance, as the typical step size is small. 1 is a common initial guess for s.

Now that we've defined all of the components of the method of steepest descent, we may determine the conditions under which it converges. To state this theorem, we first state the following definition for strong convexity:

Definition 31 ***Strong Convexity***
A function $f : \mathbb{E}^d \to \mathbb{R}$ is strongly convex if there exists a $\mu > 0$ such that;

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2}\mu \langle y - x, y - x \rangle \ \forall x, y \in \mathbb{E}^d \tag{417}$$

Thus, a function is strongly convex if it is above a quadratic approximation of it at all points. We now state an informal theorem regarding the convergence of steepest descent.

Theorem 9 ***Convergence of Steepest Descent (Informal)***
Let $f : \mathbb{E}^d \to \mathbb{R}$ be a differentiable function such that:

1. *f is strongly convex.*
2. *f has Lipschitz gradients.*

For any choice of norm and either exact or backtracking linesearch for step size, the iterates of steepest descent satisfy:

$$f(x^{(k)}) - v \leq c^k (f^{(k)} - v) \tag{418}$$

For some $c \in (0,1)$ that depends on the strong convexity parameter μ and the Lipschitz constant for the gradients.

Let's make a couple of comments about this theorem. First, note that as with projected gradient descent, we simply use $f(x^{(k)}) - v$ instead of $|f(x^{(k)}) - v|$, since v gives the global minimum of the function and $f(x) \geq v \; \forall x \in \mathbb{E}^d$.
Secondly, we make a brief note about the criterion for ϵ-accurate solutions. Using the method of steepest descent, we need $O(\log(1/\epsilon))$ iterations to obtain an ϵ-accurate solution. This type of convergence behavior is called **linear convergence**, where "linear" refers to some of the terms that appear in the proof of convergence.
Let's summarize what the method of steepest descent gives us. Using relatively minimal computation, the method of steepest descent allows us to find the minimum of a function using the functions *first order* information. What if we were to use higher order information (e.g. second derivatives) of the function? Could we use this information to reduce the number of iterations needed for an ϵ-accurate solution?
Newton's method is a numerical algorithm that uses a function's second derivative to obtain ϵ-accurate solutions to unconstrained minimization problems in fewer iterations than the method of steepest descent.

Definition 32 ***Newton's Method***
Let $f : \mathbb{E}^d \to \mathbb{R}$ be a convex, twice differentiable function. The iterates of Newton's method are given by:

$$x^{(k+1)} = x^{(k)} + s_k u^{(k)} \tag{419}$$

Where $s^{(k)} > 0$ and:

$$u^{(k)} = -[\nabla^2 f(x^{(k)})]^{-1} \nabla f(x^{(k)}) \tag{420}$$

Based on the definition of Newton's method, we notice that there are a few extra assumptions we require to apply Newton's method compared to the method of steepest descent. First, we assume that f is twice differentiable instead of once. Secondly, we require that the Hessian of f, $\nabla^2 f(x^{(k)})$, is invertible. Since the Hessian of a convex function is always at least PSD, we will require that the Hessian is PD in order to gain access to the inverse. Conveniently, if f is strongly convex, its Hessian will be positive definite. As such, we'll make the assumption of strong convexity when applying Newton's method.
What other differences are there compared to steepest descent? Newton's method is *not* a steepest descent method - there is no norm we can fix to

recover the method of steepest descent from Newton's method. Despite this, we can still interpret Newton's method using steepest descent!

Newton's method may be compared to a steepest method with a norm associated with iteration k given by:

$$||w|| = \sqrt{\langle w, \nabla^2 f(x^{(k)}) w\rangle} \tag{421}$$

This may be derived by looking at the direction of steepest descent under a norm defined by $||w|| = \langle w, Pw\rangle, P \succ 0$. Since this norm is *not* fixed, and changes with every k, Newton's method is not a steepest descent method.

At each iteration k, the method of steepest descent will minimize:

$$f(x^{(k)}) + \langle \nabla f(x^{(k)}), u\rangle + \frac{1}{2}\langle u, u\rangle \tag{422}$$

While Newton's method will minimize:

$$f(x^{(k)}) + \langle \nabla f(x^{(k)}), u\rangle + \frac{1}{2}\langle u, \nabla^2 f(x^{(k)} u\rangle \tag{423}$$

So, Newton's method tries to minimize a true quadratic approximation of the function f, while steepest descent instead minimizes a "cheap" quadratic approximation that doesn't actually use second order information.

What other properties does Newton's method have? Newton's method has the important property that it is **affinely invariant**. This means that minimizing $g(x) = f(Mx)$ using Newton's method for a non-singular linear map M yields iterates that are transformed by M. This is *not* a property shared by steepest descent!

Let's informally discuss the convergence properties of Newton's method.

Theorem 10 ***Convergence of Newton's Method (Informal)***
Let $f : \mathbb{E}^d \to \mathbb{R}$ be a twice-differentiable strongly convex function with Lipschitz Hessians. With either exact or backtracking linesearch for step size, Newton's method requires:

$$\frac{f(x^{(0)}) - v}{\gamma} + O(\log\log 1/\epsilon) \tag{424}$$

Iterations to converge to an ϵ-accurate solution.

Notice that the term on the left is simply a constant that depends on the initial condition x^0 and the optimal value v. The term on the right, $\log\log\epsilon$, is a very small number relative to $1/\epsilon$. Such convergence is called **quadratic convergence**.

Compared to steepest descent, Newton's method converges in fewer iterations but at the cost of more computationally expensive iterations. Especially for higher dimensional problems, where a large Hessian needs to be inverted, Newton's method can become prohibitively slow.