1 Introduction

In an earlier note, we found methods for finding a local minimum of some differentiable function \( f(u) : \mathbb{R}^m \to \mathbb{R} \). If \( f(u) \) is at least weakly convex, then the minimum value \( f^* \defeq f(u^*) \) is unique, and the region where this minimum value is achieved,

\[
M \defeq f^{-1}(f^*) = \{ u \in \mathbb{R}^m | f(u) = f^* \}
\]

is a single connected set.\(^1\) We wrote

\[
f^* \defeq \min_{u \in \mathbb{R}^m} f(u) \quad \text{and} \quad u^* \defeq \arg\min_{u \in \mathbb{R}^m} f(u).
\]

**Notation:** We used to denote the independent variable vector by \( z \). In the theory that follows, we will often talk about two such vectors at a time, and it is more convenient to call these \( u \) and \( v \) (what comes after \( z \) anyway?). Thus, in this note the independent variable is called \( u \) instead.

In some machine learning problems, and in particular in the theory of **Support Vector Machines** (SVMs), which we will study soon, we need to impose some constraints on the solution \( u^* \). These constraints will be in the form of a system of \( k \) affine inequalities,

\[
c(u) = Au + b \geq 0
\]

where \( A \) is a \( k \times m \) matrix and \( b \) is a column vector with \( k \) entries. This note introduces a methodology for solving minimization problems constrained in this way. The problem is called a **convex program**, and two notions of convexity are involved in its definition: set convexity and function convexity.

These two notions of convexity are reviewed in the next Section. Section 3 then defines a convex program, and Section 4 shows that the solutions to a convex program form a convex set. The remaining Sections characterize the solutions of a convex program, in the following sense. For an unconstrained convex optimization problem, we know that a point in the domain is a solution if and only if the gradient of the target function is zero at that point. For a convex program, the analogous condition is in the form of a system of necessary and sufficient equalities and inequalities called the **Karush-Kuhn-Tucker (KKT) conditions**. To establish these conditions, Sections 5 and 6 explore necessary preliminaries from the theory of convex cones. The KKT conditions are then derived in Section 7. Finally, Section 8 introduces the notion of Lagrangian duality, which allows transforming a convex program into an equivalent maximization problem. For now, the reasons for introducing Lagrangian duality will seem arbitrary. They will become clear in the context of Support Vector Machines, where duality leads to a far reaching generalization through what the literature often calls the **kernel trick**.

\(^1\)We will soon see that this region is actually a convex set.
2 Set and Function Convexity

Each row $c_i(u) \geq 0$ of the system of inequalities in expression 1 defines a half-space in $\mathbb{R}^m$, and therefore the set

$$C \overset{\text{def}}{=} \{ u \in \mathbb{R}^m : c(u) \geq 0 \}$$

of points that satisfy all inequalities is the intersection of $k$ half-spaces. Such a set is called a \textit{convex polyhedral set}. A set is \textit{polyhedral} when it is the intersection of half-spaces.

In one dimension, a convex polyhedral set is an interval; in two it is a convex polygon; in three it is a convex polyhedron. In general, a set $C \subseteq \mathbb{R}^m$ is \textit{convex} if it contains the line segment between any two of its points. Formally, $C$ is convex if for every $u, v \in C$ and $t \in [0, 1]$

$$t \, u + (1 - t) \, v \in C .$$

It is easy to check that $C$ is convex: If $u, v \in C$, then

$$Au + b \geq 0 \quad \text{and} \quad Av + b \geq 0 \quad (2)$$

and if $t \in [0, 1]$ then

$$c(t \, u + (1 - t) \, v) = A(t \, u + (1 - t) \, v) + b = t(Au + b) + (1 - t)(Av + b) \geq 0$$

because of the inequalities 2 and because both $t$ and $1 - t$ are nonnegative. This ends the proof.

Do not confuse the notion of a convex \textit{set} with that of a convex \textit{function}. Recall that a function $f : \mathbb{R}^m \to \mathbb{R}$ is \textit{convex} if the graph of $f$ is below the line through any two points in $\mathbb{R}^m$. Formally, $f$ is convex if for every $u, v \in \mathbb{R}^m$ and $t \in [0, 1]$ the following inequality holds

$$f(t \, u + (1 - t) \, v) \leq tf(u) + (1 - t)f(v) .$$

In particular, it follows from this definition that if $f$ is an affine function on $\mathbb{R}^m$ then both $f$ and $-f$ are convex. A function $f$ is said to be \textit{strictly convex} if the inequality above holds strictly whenever $u \neq v$:

$$u \neq v \quad \Rightarrow \quad f(t \, u + (1 - t) \, v) < tf(u) + (1 - t)f(v) .$$

It is easy to prove that the maximum of two (and therefore also countably many) convex functions is convex: Let

$$h(x) = \max\{h_1(x), h_2(x)\} \quad \text{for all } x \text{ in some domain}.$$ 

Then,

$$h((1 - t) \, x + ty) = \max\{h_1((1 - t) \, x + ty), h_2((1 - t) \, x + ty)\} \leq \max\{(1 - t) \, h_1(x) + t \, h_1(y), (1 - t) \, h_2(x) + t \, h_2(y)\} \leq (1 - t) \max\{h_1(x), h_2(x)\} + t \max\{h_1(y), h_2(y)\} = (1 - t) \, h(x) + th(y) ,$$

so that $h$ is convex. Similar reasoning shows that the sum of two (and therefore also countably many) convex functions is convex (try this).
3 Convex Programs

Our new minimization problem is to find the minimum value of a convex, smooth function over a convex set formed by the intersection of half-spaces. More formally:

A convex program is the following constrained optimization problem:

\[ f^* \overset{\text{def}}{=} \min_{u \in C} f(u) \]  \hspace{1cm} (3)

where the minimum \( f^* = f(u^*) \) is achieved at the point

\[ u^* \overset{\text{def}}{=} \arg \min_{u \in C} f(u) \]

and where

\[ C \overset{\text{def}}{=} \{ u \in \mathbb{R}^m : c(u) \geq 0 \} . \]

In these expressions, the function \( f \) is differentiable, has continuous first derivative, and is convex; \( m \) is a fixed positive integer; and the \( k \) inequalities that define the set \( C \) are affine:

\[ c(u) = Au + b \geq 0 . \]

In these definitions, \( A \) is a \( k \times m \) matrix and \( b \) is a column vector with \( k \) entries.

The function \( f \) is called the target of the problem, and a point \( u \) is said to be feasible if it is in \( C \).

The formulation above also accommodates affine equality constraints, because the equation

\[ c(u) = 0 \]

is equivalent to the system of the following two inequalities:

\[ c(u) \geq 0 \quad \text{and} \quad c(u) \leq 0 . \]

4 Convexity of the Solution Set of a Convex Program

A convex program does not necessarily have a solution. For instance, with

\[ m = 1 , \quad f(u) = e^{-u} \quad \text{and} \quad c(u) = u , \]

the convex, smooth, and decreasing function \( f \) is unbounded from below in the convex set

\[ C = \{ u : u \geq 0 \} , \]

so no minimum (local or global) exists (Figure 1). On the other hand,

**Theorem 4.1.** If a solution to a convex program exists, the set of all solution points \( u^* \) forms a convex set.

This theorem is proven in Appendix A. The proof requires only convexity of \( C \) and \( f \), and does not rely on differentiability. The result also applies to the unconstrained case, because \( \mathbb{R}^m \) is a convex set.

Thus, the solutions to a convex program form a single, convex set \( M \), on which \( f \) takes on its unique minimum value \( f^* \). In addition, \( M \) shrinks to a point if \( f \) is strictly convex there:
Figure 1: The function $e^{-u}$ is unbounded from below in the constraint set $C = \{ u : u \geq 0 \}$.

**Theorem 4.2.** If $u^*$ is a solution to a convex program and $f$ is strictly convex at $u^*$, then $u^*$ is the unique solution to the program.

Again, the proof, shown in Appendix B, requires no differentiability.

One could still conceive of local minima in addition to a single, convex region $M$ with all the points of global minimum. However, if $f$ is differentiable, this cannot happen, as a consequence of the striking fact that a convex differentiable function is always bounded from below by any of its tangent planes. More formally,

**Theorem 4.3.** Let $f$ be differentiable on the convex (not necessarily polyhedral) domain $\Omega$. Then, $f$ is convex if and only if

$$f(v) \geq f(u) + \nabla f(u)^T(v - u)$$

for all $u, v \in \Omega$.

This result is proven in Appendix C, and is striking because it shows that for a convex function, local information (the value and gradient of $f$ at $u$) yields a global underestimator of $f$ on $\Omega$.

If $f$ is differentiable on the convex set $\Omega$, then it is differentiable also on the (convex) region $M$ of its global minima, so that theorem 4.3 implies that

$$f(v) \geq f(u^*) \quad \text{for all} \quad u^* \in M \quad \text{and} \quad v \in \Omega.$$ 

The discussion in this Section can be summarized as follows.

**Corollary 4.4.** All minima of a convex program are global and the points where the unique minimum $f^*$ is achieved form a convex set. This set is a single point $u^*$ if the program’s target $f$ is strictly convex there.

---

5 Closed Convex Polyhedral Cones

**Motivation** As we know, a convex differentiable function $f : \mathbb{R}^m \to \mathbb{R}$ without constraints on its domain has a zero gradient at its points of minimum. Because of convexity, this condition is also sufficient for a minimum. For a convex program, a zero gradient is obviously still a necessary and sufficient condition for
a minimum in the interior of the constraint set $C$. However, the additional possibility now arises of minima on the boundary of $C$. For instance, if we let $m = 1$, the function

$$f(u) = e^u$$

has no global minimum (Figure 2 (left)). However, if we add the constraint

$$c_1(u) = u \geq 0$$

a global minimum value $f(u^*) = 1$ arises at $u = 0$ (Figure 2 (center)).

Adding constraint

$$c_2(u) = 1 - u \geq 0$$

even generates a global maximum value $e$ at $u = 1$ (Figure 2 (right)). Maxima do not occur in an unconstrained convex function, so this is an especially striking difference between the unconstrained and the constrained case. Both extrema are produced by the “collision” of the graph of the function $f$ with the new constraints.

Saddle points, on the other hand, cannot arise from restricting a convex function on a convex set. This is because a saddle point at $u$ would require both positive and negative derivatives along different directions out of $u$. These do not occur for a convex function.

Because of the possible emergence of minima at the boundary of $C$, we need new necessary and sufficient conditions for minima of $f$ that take the boundary into account. These conditions, called the Karush-Kuhn-Tucker (KKT) conditions, will be derived in Section 7.

In the present Section, we introduce a general description of a local neighborhood at a point on the boundary of $C$. For the unconstrained case, a neighborhood of any point $u$ looks like all of $\mathbb{R}^m$. The neighborhood of a point on the boundary of a convex polyhedral set, on the other hand, looks like a closed convex polyhedral cone, introduced next.

**Intuition** A point $u_0$ on the boundary of a convex polygon $C$ in $\mathbb{R}^2$ is either on a side or on a vertex of $C$. If it is on a side, then the set $C$ looks locally like a half-plane $H$. “Locally” here means that we consider a small neighborhood of $u_0$, so small that other sides of $C$ don’t matter. Then only the constraint

$$c(u) = b + a^T u \geq 0$$
Figure 3: An angle $A$ at $u_0 = 0$, two generators $r_1$, $r_2$ along the rays that bound the angle, and two hyperplane normals $a_1$, $a_2$. The vector $r_1$ is orthogonal to $a_1$, and the same holds for $r_2$ and $a_2$. The hyperplane normals $a_1$, $a_2$ point towards the interior of $A$.

corresponding to that edge is active, that is, it actually constrains the position of $u_0$.

To simplify the representation when taking a local view of a boundary in a neighborhood of a point $u_0$, we translate the reference system by placing its origin at $u_0$. The inequality that defines the half-plane $H$ is then homogeneous in $u$:

$$ a^T u \geq 0. $$

When $u_0$ is a vertex of $C$, then $C$ looks locally like an angle $A$, that is, the part of the plane between two half-lines, or rays, that meet at the vertex (Figure 3). Since $C$ is convex, the angle between the two rays, measured inside $A$, is at most $\pi$ radians.

There are two simple ways to define the angle $A$. One is the original one, as an intersection of half-spaces (half-planes):

$$ a^T u \geq 0 $$

where the vectors $a_i$ are nonzero and, as we know, orthogonal to the line (hyperplane) that bounds each of the half-spaces (see Figure 3). Since these lines now pass through the origin, the equations are homogeneous in $u$. We call this an implicit representation.

The other representation of $A$ is as the conical linear combination of two vectors $r_1$, $r_2$ that point along the lines in the appropriate direction:

$$ A = \{ u \in \mathbb{R}^2 : u = \alpha_1 r_1 + \alpha_2 r_2 \text{ with } \alpha_1, \alpha_2 \geq 0 \}. $$

The inequalities on $\alpha$ are what makes this a conical combination. We call this representation of $A$ parametric, and the vectors $r_i$ are the generators of the angle. The half-lines along the generators are called rays.
Regardless of representation, the angle $A$ is an example of a Closed Convex Polyhedral cone (CCP cone). It is closed in that it contains its boundary (the two rays). It is convex because it is a convex set. It is polyhedral because it is bounded by hyperplanes. It is a cone in that if point $u$ is in the set, so is every point of the form $\alpha u$ for any nonnegative real number $\alpha$. Use either representation of $A$ to check that these properties hold.

In this simple, two-dimensional example, the relationship between the two representations of $A$ is easy to find: The vector $\mathbf{a}_i$ is perpendicular to the vector $\mathbf{r}_i$, and one just needs to pick the signs appropriately to make sure that the correct angle is chosen, out of the four possibilities. This simplicity, however, is misleading, as it does not hold in general. In $d$ dimensions for $d > 2$, many more situations can arise. In particular, the number of generators is not necessarily equal to the number of hyperplane normals and the problem of computing the hyperplane normals $\mathbf{a}_j$ from the generator vectors $\mathbf{r}_i$ is a difficult problem called the “representation conversion problem” in the literature. There are several algorithms for converting between these two representations, of varying degrees of efficiency [2]. We will not need to perform these conversions for our purposes.

We may get a faint glimpse of the fact that things are nontrivial even in the planar case. Specifically, the parametric representation of $A$ degenerates when the two rays point in exactly opposite directions,

$$\mathbf{r} \overset{\text{def}}{=} \mathbf{r}_1 = -\mathbf{r}_2$$

because then the two vectors merely span a line. First, note that even in this case the set $A$, which degenerates to a single line $L$ through the origin, is a CCP cone, formed by the intersection of the two half-planes

$$\mathbf{a}^T \mathbf{u} \geq 0$$

$$-\mathbf{a}^T \mathbf{u} \geq 0$$

where $\mathbf{a}$ is a nonzero vector orthogonal to $\mathbf{r}$. Thus, both an explicit and an implicit representation exist for a line $L$ through the origin.

On the other hand, the limit of $A$ as the two rays approach a flat angle is a half-plane $H$. This is represented by just one of these inequalities, say,

$$\mathbf{a}^T \mathbf{u} \geq 0$$

where the sign of $\mathbf{a}$ must be chosen so that $\mathbf{a}$ points towards the interior of $H$. A half-plane is still a CCP cone, but two generators are not enough for it. To generate $H$, just think of it as formed by gluing together two non-degenerate cones, for instance, that generated by $\mathbf{r}$ and $\mathbf{a}$, and that generated by $-\mathbf{r}$ and $\mathbf{a}$, so that

$$H = \{ \mathbf{u} \in \mathbb{R}^2 : \mathbf{u} = \alpha_1 \mathbf{r} + \alpha_2 \mathbf{a} + \alpha_3 (-\mathbf{r}) \text{ with } \alpha_1, \alpha_2, \alpha_3 \geq 0 \}.$$  

This representation is obviously not unique, even if we were to restrict vectors to have unit length, because a half-plane can be formed in many ways by gluing two angles together. In addition, redundant parametric representations exist as well. For instance, one can think of gluing together two overlapping angles, which together require four generators. However, in all that follows, we will always think of minimal (that is, non-redundant) representations.

The fact that we were able to find both an implicit and a parametric representation even in this degenerate case (regardless of whether we are interested in either $L$ or $H$), is not a coincidence. The Farkas-Minkowski-Weyl theorem [3], whose proof is beyond the scope of these notes, states that every CCP cone has representations of both forms. Appendix D briefly elaborates on this theorem.
A single ray $R$ is also a CCP cone. Its parametric representation is obvious, and requires a single generator $r$. For an implicit representation, first represent the line $L$ that $R$ is on, as we did above, then cut off half of the line with the additional half-space $r^T u \geq 0$.

In three dimensions, the variety of CCP cones increases. It includes all the two-dimensional cones (half-lines, lines, half planes, and angles), plus pyramidal angles (think of the top of a pyramid) with any number of faces, and wedges. A wedge is the intersection of two half-spaces, and is the three-dimensional extrusion of an angle into the third dimension.

\textbf{Formalization}  A \textit{Closed Convex Polyhedral cone} (CCP cone) in $\mathbb{R}^m$ is an intersection of $k$ half-spaces through the origin, with $k \geq 1$:

$$P = \{ u \in \mathbb{R}^m : p_i^T u \geq 0 \text{ for } i = 1, \ldots, k \}$$

(4)

where the vectors $p_i$ are nonzero. This is the \textit{implicit} representation of $P$.

Redundant representations are possible. However, we henceforth assume that $k$ is the smallest number of half-spaces needed to represent the cone.

It is immediate to verify that $P$ is indeed a CCP cone, that is, that it has the following properties:

- It is closed, in that it contains its boundary (because of the weak inequalities).
- It is a cone, in that if $u \in P$ then $\alpha u \in P$ for any nonnegative $\alpha \in \mathbb{R}$ (because the inequalities are homogeneous in $u$).
- It is polyhedral, because it is the intersection of half-spaces.
- It is convex.

Convexity can be verified by applying the definition directly: If $u, v \in P$, then for all $i = 1, \ldots, k$ we have

$$p_i^T u \geq 0 \quad \text{and} \quad p_i^T v \geq 0$$

and therefore, given a real number $t \in [0, 1]$

$$p_i^T [tu + (1-t)v] = tp_i^T u + (1-t)p_i^T v \geq 0$$

because both $t$ and $1-t$ are nonnegative. Therefore, the point $tu + (1-t)v$, which is an arbitrary point on the line segment between $u$ and $v$, is also in $P$, that is, $P$ is a convex set.

\section{Cone Duality}

As we saw in some examples in the previous Section, the hyperplane normals $p_i$ and generators $r_j$ of a given CCP cone $P$ are two different sets of vectors, and they are not even equal in number. They are related in a complex way by the so-called “representation conversion problem” [2].

In this Section, we explore the notion of \textit{duality}, which looks at pairs of implicit and parametric representations of two different CCP cones, and such that the hyperplane normals of one are the generators of the other. As we will see in Section 7, duality will let us introduce necessary and sufficient conditions for a point $u \in \mathbb{R}^m$ to be a solution of a given convex program.
Figure 4: In both figures, the gray region delimited by black lines is the primal cone $P$ and the red region delimited by red lines is the corresponding dual cone $D$. All and only the points in a gray region have a nonnegative inner product with all and only the points of the corresponding red region. When the cone $P$ subtends an angle of $\phi$ radians, the cone $D$ subtends an angle of $\pi - \phi$ radians. When $\phi = \pi/2$, the two cones coincide, and $P$ is its own dual. Because of duality, the two figures are the same, but with the colors switched.

The dual of the cone $P$ with implicit representation

$$P = \{ u \in \mathbb{R}^m : p_i^T u \geq 0 \text{ for } i = 1, \ldots, k \}$$

is the CCP cone $D$ generated by the hyperplane normals of $P$:

$$D = \{ u \in \mathbb{R}^m : u = \sum_{i=1}^{k} \alpha_i p_i \text{ with } \alpha_1, \ldots, \alpha_k \geq 0 \} .$$

It is easy to verify that any set that has this representation is a convex, closed cone (try this). A theorem outlined in Section D implies that this cone is also polyhedral and, in fact, any CCP cone has representations of both forms. A representation in the form in which $D$ is given above is called parametric, and the vectors $p_i$ are the generators of the cone.

In a pair of CCP cones $(P, D)$ constructed in this way, the cone $P$ is called the primal cone. Note carefully that $P$ and $D$ are different CCP cones built from the same $k$ vectors $p_1, \ldots, p_k$ through two different representations (implicit and parametric).

As an example, Figure 4 shows two different primal-dual pairs on the plane. Their primal and dual cones are all angles, with different apertures. The 90-degree angle is its own dual.

A key property of duality is its symmetry, which is proven in Appendix E together with the characterization of duality given in the following theorem. That Appendix also introduces an appealing geometric view of cone duality.

**Theorem 6.1.** If $P$ and $D$ are CCP cones, then all and only the points $u$ in $P$ have a nonnegative inner product with all and only the points $v$ in $D$. Since this property is symmetric in $P$ and $D$, it follows that if $D$ is the dual cone of CCP cone $P$, then $P$ is the dual cone of $D$. 
7 The Karush-Kuhn-Tucker Conditions

We now have the tools necessary to develop conditions for a point $u \in \mathbb{R}^m$ to be a solution to the convex program 3 defined on page 3. The point $u$ is a (weak) minimum if and only if, when moving away from $u$ in any direction while staying in the constraint set $C$, the value of the target function $f$ does not increase. Thus, if we pick a direction $s \in \mathbb{R}^m$ out of $u$, either we stay in $C$ and the value of $f$ increases (weakly),

$$s^T \nabla f(u) \geq 0$$

or at least some of the constraints are violated, that is, we leave $C$. To state this last condition, we define the index set $A(u)$ of the constraints that are active at $u$, that is,

$$A(u) = \{ i : c_i(u) = 0 \} .$$

If a constraint is not active at $u$, then it is not possible to violate it with a differentially small move away from $u$. If constraint number $i$ is active, on the other hand, it is violated as soon as $c_i$ decreases in some direction $s$. Thus, for a constraint violation to occur for a differentially small move away from $u$ along direction $s$, the following must hold:

$$\exists i \in A(u) : s^T \nabla c_i(u) < 0 .$$

In summary, $u$ is a minimum for the convex program 3 if an only if for all $s \in \mathbb{R}^m$ either $f$ does not decrease along $s$, or a move along $s$ violates some active constraint:

$$\forall s \in \mathbb{R}^m : [s^T \nabla f(u) \geq 0] \cup [\exists i \in A(u) : s^T \nabla c_i(u) < 0] .$$

Let now $P$ be the CCP cone defined by the gradients of the active constraints at $u$:

$$P = \{ s : s^T \nabla c_i(u) \geq 0 \text{ for } i \in A(u) \} .$$

This is the cone of directions along which one can make a differentially small move without violating any constraint, that is, while staying in the constraint set $C$.

For $u$ to be a minimum, we want the target function $f$ not to decrease if we move in any direction in $P$. This happens if and only if the gradient $g \overset{\text{def}}{=} \nabla f(u)$ of $f$ at $u$ has a nonnegative inner product with all the directions in $P$, that is, if $g$ is in the dual of $P$,

$$D = \{ g : g = \sum_{i \in A(u)} \alpha_i \nabla c_i(u) \text{ with } \alpha_i \geq 0 \} .$$

In summary:

**Theorem 7.1.** A point $u^* \in C$ is a solution to the convex program 3 if and only if there exist coefficients $\alpha_i^*$ for $i \in A(u)$ such that

$$\nabla f(u^*) = \sum_{i \in A(u^*)} \alpha_i^* \nabla c_i(u^*) \text{ with } \alpha_i^* \geq 0 .$$

The gradient computations in this equation can be “factored out” by writing

$$\frac{\partial}{\partial u} \mathcal{L}(u^*, \alpha^*) = 0 \text{ where } \mathcal{L}(u, \alpha) \overset{\text{def}}{=} f(u) - \sum_{i \in A(u)} \alpha_i c_i(u)$$

10
is a function of both \( u \) and \( \alpha \) called the the Lagrangian of the convex program.

Leibnitz notation (\( \partial / \partial u \)) rather than the nabla symbol (\( \nabla \)) was used here to denote the gradient, in order to emphasize that the gradient is with respect to \( u \) (as opposed to \( \alpha \)). The nonnegative numbers \( \alpha_i \) are called the Lagrange multipliers for the constraints.

Having a number of terms in the summation above that depends on \( u^* \) is inconvenient. Instead, we can include all the constraints, not just the active ones, and then add a condition that states that either a constraint \( c_i(u^*) \) is active \( (c_i(u^*) = 0) \) or the corresponding Lagrange multiplier \( \alpha^*_i \) is zero. Since both the \( \alpha^*_i \) and the \( c_i(u^*) \) are nonnegative in \( C \), we can write these \( k \) conditions as a single equation

\[
(\alpha^*)^T c(u^*) = 0.
\]

We can summarize this discussion as follows.

**Theorem 7.2. (Karush-Kuhn-Tucker)** A point \( u^* \) is the unique solution to the convex program defined in equation 3 on page 3 if and only if there exists a vector \( \alpha^* \in \mathbb{R}^k \) such that

\[
\begin{align*}
\frac{\partial \mathcal{L}(u^*, \alpha^*)}{\partial u} &= 0 \\
c(u^*) &\geq 0 \\
\alpha^* &\geq 0 \\
(\alpha^*)^T c(u^*) &= 0
\end{align*}
\]

where

\[
\mathcal{L}(u, \alpha) \overset{\text{def}}{=} f(u) - \alpha^T c(u)
\]

is the Lagrangian of the convex program.

The conditions 5 are copied from the statement of the convex program, and mean that \( u^* \) is in the constraint set \( C \).

The condition 6 is called the complementarity condition, as it states that the sets of indices \( i \) on which \( \alpha^*_i \) and \( c_i(u^*) \) are nonzero are mutually complementary. A weaker version of equation 6 holds more generally. Specifically, away from the solution \( u^* \), a feasible point \( u \) is still required to be in \( C \), so that \( c(u) \geq 0 \). Also, the Lagrange multipliers are nonnegative everywhere, \( \alpha \geq 0 \), by the definition of a CCP cone. Therefore, we generally have

\[
\alpha^T c(u) \geq 0
\]

for all feasible points \( u \), with equality holding at the solution.

Together, the five conditions in the Karush-Kuhn-Tucker theorem are called the Karush-Kuhn-Tucker conditions, or KKT conditions for short. They fully characterize the solution to the convex program 3.

### 8 Lagrangian Duality

This Section defines what is called the dual problem of the convex program 3, which is then referred to as the primal problem. The dual problem is a convex maximization problem in the vector \( \alpha \) of the Lagrange multipliers of the primal problem. The solution of the dual is equivalent to that of the primal, in the sense that the optimal value of the primal target is equal to the optimal value of the dual target, and the optimal
vector $\alpha^*$ at which the dual problem achieves its maximum is equal to the vector of Lagrange multipliers that yield the optimal solution $u^*$ of the primal.

There are multiple reasons why this duality is useful. The reason that matters in this course is that the dual formulation of the optimization problem involved in the definition of Support Vector Machines (SVMs) leads to both insights into the meaning of the support vectors and a formulation of SVMs that can be generalized to a wide variety of classification problems through the so-called kernel trick.

At the minimum $u^*$ of the primal convex program, the Lagrangian

$$L(u, \alpha) \overset{\text{def}}{=} f(u) - \alpha^T c(u)$$

reduces to the target value,

$$L(u^*, \alpha^*) = f(u^*) ,$$

because of the complementarity condition 6. Also, since both $f(u)$ and $\alpha^T c(u)$ are convex, so is $L(u, \alpha)$, and therefore the first KKT condition implies that $(u^*, \alpha^*)$ is a global minimum of $L$.

For feasible points $u$ away from $u^*$, the Lagrangian is a lower bound for the target for any nonnegative $\alpha$, because $\alpha^T c(u)$ is nonnegative (equation 7) for all feasible points:

$$L(u, \alpha) \leq f(u) \quad \text{for all } u \in C \quad \text{and } \alpha \geq 0 .$$

This lower bound can be made independent of $u$ by introducing the (Lagrangian) dual function

$$D(\alpha) \overset{\text{def}}{=} \min_{u \in C} L(u, \alpha) .$$

For any nonnegative $\alpha$, the Lagrangian $L$ is a pointwise lower bound, while the dual $D$ is a lower bound (because of the min in its definition) that is constant with respect to $u$, and is therefore also a lower bound for the value $f^* = f(u^*)$ of the target at the minimum:

$$D(\alpha) \leq f^*$$

regardless of the (nonnegative) value of $\alpha$.

A natural question is how tight this bound can be made by varying $\alpha$, and the answer is striking: *The bound becomes an equality when and only when $\alpha = \alpha^*$, the vector of Lagrange multipliers that yield the minimum $u^*$ for the original convex program:*

$$\max_{\alpha \geq 0} D(\alpha) = f^* \quad \text{and} \quad \arg \max_{\alpha \geq 0} D(\alpha) = \alpha^* .$$

The proof is a one-liner, since the complementarity condition 6 yields

$$D(\alpha^*) = \min_{u \in C} L(u, \alpha^*) = L(u^*, \alpha^*) = f(u^*) - (\alpha^*)^T c(u^*) = f(u^*) = f^* .$$

Thus, since $D(\alpha) \leq f^*$ for all $\alpha \geq 0$ and $D(\alpha^*) = f^*$, the vector $\alpha^*$ is the solution of the following *dual problem:*

$$\alpha^* = \arg \max_{\alpha \geq 0} D(\alpha) .$$

The dual problem can be rewritten as a convex program by replacing $D$ with $-D$ and observing that $-D$ is convex in $\alpha$. To show convexity, note that the function

$$-L(u, \alpha) = \alpha^T c(u) - f(u)$$
is affine in $\alpha$ for each fixed value of $u$, so that $-D(\alpha)$ is the point-wise maximum of an infinite number of affine functions of $\alpha$, one function for each value of $u$. An affine function is both concave and convex and, as shown in Section 2, the maximum of two (and therefore arbitrarily many) convex functions is convex. To summarize:

Let

$$L(u, \alpha) \overset{\text{def}}{=} f(u) - \alpha^T c(u)$$

be the Lagrangian of the primal convex program

$$f^* \overset{\text{def}}{=} \min_{u \in C} f(u)$$

where

$$C \overset{\text{def}}{=} \{ u \in \mathbb{R}^m : c(u) \geq 0 \},$$

and define the dual problem to be

$$\max_{\alpha \geq 0} D(\alpha) \quad \text{for} \quad \alpha \geq 0$$

where

$$D(\alpha) \overset{\text{def}}{=} \min_{u \in C} L(u, \alpha)$$

is the Lagrangian dual of $L$. Then,

$$f^* = f(u^*) = \min_{u \in C} f(u) = L(u^*, \alpha^*) = \max_{\alpha \geq 0} D(\alpha) = D(\alpha^*) .$$

Thus, the Lagrangian has a minimum in $u$ and a maximum in $\alpha$ at the solution of both primal and dual problem. In other words, the point $(u^*, \alpha^*)$ is a saddle point for $L$.

**Summary** Let us review what we did in the last two Sections.

The primal convex program can be transformed into the problem of minimizing the Lagrangian $L$ with respect to the original unknown vector $u$ and the additional unknown vector $\alpha$ of Lagrange multipliers, and the KKT conditions are a system of equations and inequalities that are necessary and sufficient for the values $u^*, \alpha^*$ of $u, \alpha$ to yield the minimum.

The duality between minimizing $L$ and maximizing $D$ yields an alternative view of the same problem. Specifically, think of letting $\alpha$ vary, and ask for the value $u_0(\alpha)$ of $u$ that makes the Lagrangian $L$ as small as possible for each choice of $\alpha$. This value is the dual $D(\alpha) = L(u_0(\alpha), \alpha)$. Since the term $\alpha^T c(u)$ in $L$ is never negative in the feasible region for $u$, the value of $D(\alpha)$ is no greater than that of $f(u_0(\alpha))$:

$$f(u_0(\alpha)) - D(\alpha) = f(u_0(\alpha)) - [f(u_0(\alpha)) - \alpha^T c(u_0(\alpha))] = \alpha^T c(u_0(\alpha)) \geq 0,$$

and we say that there is a nonnegative duality gap between these two values.

The principle of Lagrangian duality then states that to make $f(u)$ as small as possible, the values in $\alpha$ are to be chosen to make $D(\alpha)$ as large as possible, and that for that choice the duality gap shrinks to zero.
Appendices

A  Proof of Theorem 4.1

If a solution to a convex program exists, the set of all solution points \( u^* \) forms a convex set.

Proof. Let \( M \) be the set of all global minima of \( f \) in \( C \), and let \( u \) and \( v \) be two points in \( M \), so that

\[
 f(u) = f(v). \tag{9}
\]

Since \( C \) is convex, for all \( t \in [0, 1] \) we have

\[
 (1 - t)u + tv \in C
\]

and since \( f \) is convex on \( C \)

\[
 f((1 - t)u + tv) \leq (1 - t)f(u) + tf(v) = (1 - t)f(u) + tf(u) = f(u)
\]

(using equation 9 for the first equality above), so that all points on the line segment between \( u \) and \( v \) are also global minima, and are therefore in \( M \). This shows that \( M \) is convex.

\( \Delta \)

B  Proof of Theorem 4.2

If \( u^* \) is a solution to a convex program and \( f \) is strictly convex at \( u^* \), then \( u^* \) is the unique solution to the program.

Proof. Assume, for a contradiction, that \( u \neq u^* \) is another solution, so that \( u \in C \) and \( f(u) = f(u^*) \).

Then since \( C \) is convex, the point

\[
 v = (1 - t)u + tu^*
\]

is in \( C \) for all \( t \in [0, 1] \). Furthermore, since \( f \) is strictly convex at \( u^* \),

\[
 f(v) = f((1 - t)u + tu^*) < (1 - t)f(u) + tf(u^*) = f(u^*) .
\]

Thus, the value of \( f \) at \( v \in C \) is smaller than that at \( u^* \), contradicting the assumption that \( u^* \) is a minimum, so that \( u \) cannot exist.

\( \Delta \)

C  Proof of Theorem 4.3

Let \( f \) be differentiable on the convex (not necessarily polyhedral) domain \( \Omega \). Then, \( f \) is convex if and only if

\[
 f(v) \geq f(u) + \nabla f(u)^T (v - u)
\]

for all \( u, v \in \Omega \).
Proof. Let \( u \) and \( v \) be two points in \( \Omega \). Any point on the line segment between \( u \) and \( v \) can be written as
\[
w = (1 - t)u + tv \quad \text{for} \quad t \in [0, 1]
\]
and any such point is in \( \Omega \) because of the convexity of \( \Omega \).

Assume that the inequality in the statement of the theorem holds, and rewrite it twice as follows:
\[
f(u) \geq f(w) + \nabla f(w)^T (u - w)
\]
\[
f(v) \geq f(w) + \nabla f(w)^T (v - w)
\]
Multiplying the first equation by \( 1 - t \), the second by \( t \), and using the definition of \( w \) in equation 10 yields
\[
(1 - t)f(u) + tf(v) \geq f(w)
\]
which means that \( f \) is convex.

For the converse, let
\[
g(t) = f((1 - t)u + tv)
\]
be the restriction of \( f \) on the segment in equation 10. The derivative of \( g \) at \( t \) is
\[
g'(t) = \nabla f((1 - t)u + tv)^T(v - u).
\]
In particular,
\[
g'(0) = \nabla f(u)^T(v - u),
\]
so that the inequality in the theorem statement becomes
\[
g(1) \geq g(0) + g'(0).
\]

Assume now that \( f \) is convex, and therefore so is \( g \). Then,
\[
g(t) \leq (1 - t)g(0) + tg(1).
\]
For any \( t \in (0, 1] \), this yields
\[
g(1) \geq g(0) + \frac{g(t) - g(0)}{t}
\]
and taking the limit for \( t \to 0 \) yields equation 11. This completes the proof.

D The Farkas-Minkowski-Weyl Theorem

Theorem D.1. (Farkas-Minkowski-Weyl) Every CCP cone \( P \) has both an implicit representation
\[
P = \{ u \in \mathbb{R}^m : c_i^T u \geq 0 \quad \text{for} \quad i = 1, \ldots, k \}
\]
where the vectors \( c_i \) are nonzero and a parametric representation:
\[
P = \{ u \in \mathbb{R}^m : u = \sum_{j=1}^\ell \alpha_j r_j \quad \text{with} \quad \alpha_1, \ldots, \alpha_\ell \geq 0 \}
\]
where the vectors \( r_j \) are nonzero and are called the generators of \( P \). Conversely, every set with either representation is a CCP cone.

Proof. The proof can be found the literature [3] and is beyond the scope of these notes.
E  Proof of Theorem 6.1

If $P$ and $D$ are CCP cones, then all and only the points $u$ in $P$ have a nonnegative inner product with all and only the points $v$ in $D$. Since this property is symmetric in $P$ and $D$, it follows that if $D$ is the dual cone of CCP cone $P$, then $P$ is the dual cone of $D$.

To prove this symmetry property, we first state the following result, which is a special case of a theorem that is fundamental to constrained optimization. We will use this result in Section 6. The general result is for pairs of convex sets, rather than for a point and a CCP cone, and somewhat laborious to prove. However, the special result for CCP cones is rather intuitive, and we leave it without proof. The proof for the general case can be found in standard texts (for instance, Section 2.5.1 of a book by Boyd [1]).

**Theorem E.1. (Separating Hyperplane)** Let $P$ be a CCP cone and $v \notin P$. Then there exists a plane through the origin with equation
\[ a^T u = 0 \]
such that
\[ a^T v < 0 \quad \text{and} \quad a^T u \geq 0 \quad \text{for all} \quad u \in P. \]

This theorem states that it is always possible to place a hyperplane between a CCP cone and a point not in the convex set. The hyperplane touches the cone at the origin.

A corollary of this theorem states that no matter what the primal-dual pair of cones $(P, D)$ is, if we pick a point $v$ that is outside $P$ (either in $D$ or outside of it), then $D$ is always wide enough to contain a point $d$ that is more than 90 degrees away from $v$.

**Corollary E.2.** Let $P$ be a CCP cone, $D$ its dual cone, and $v \in \mathbb{R}^m$ a point not in $P$. Then, there exists a point $d \in D$ such that $d^T v < 0$.

**Proof.** By the separating hyperplane theorem, there exists some hyperplane with equation $d^T v = 0$ that separates $v$ and $P$, so that $d^T v < 0$ and $d^T p \geq 0$ for all $p \in P$. From the last inequality, $d \in D$ by definition of dual.

The symmetric nature of cone duality is in turn a corollary of what we just proved.

**Proof.** Let
\[ P = \{ u \in \mathbb{R}^m : p^T u \geq 0 \ \text{for} \ i = 1, \ldots, k \} \]  
(13)
so that the dual cone of $P$ is
\[ D = \{ v \in \mathbb{R}^m : v = \sum_{i=1}^k \alpha_i p_i ; \ \text{with} \ \alpha_1, \ldots, \alpha_k \geq 0 \}. \]  
(14)

Consider a point $u \in P$. From the definitions 13 of $P$ and 14 of $D$, we see that if $v \in D$ then
\[ v^T u = \sum_{i=1}^k \alpha_i p_i^T u \geq 0. \]

Thus, every $u \in P$ has nonnegative inner product with every $v \in D$.  

16
Figure 5: The Cranmer Park Sundial in Dever, Colorado. The stick of the sundial is called the gnomon, and its surface is called the dial.

Consider now a point $u'$ not in $P$. Then, by corollary E.2, there exists a point $v' \in D$ such that

$$(v')^T u' < 0.$$ 

Therefore, $u$ is in $P$ if and only if it has a nonnegative inner product with every $v$ in $D$. However, the inner product commutes, so that the equivalence also goes the other way: $v$ is in $D$ if and only if it has a nonnegative inner product with every $u$ in $P$. Therefore, $P$ is the dual of $D$. △

The Sundial The first sentence of Theorem 6.1 is a geometric characterization of cone duality, and is sometimes used as its definition. This characterization is a useful way to think about duality, and can be visualized as follows.

Pick a line through the origin and a point $u$ in the primal CCP cone $P$. This line can be thought as the gnomon of a sundial (Figure 5). The dial of the sundial is the (hyper)plane orthogonal to the gnomon through the origin. The half-space $H^+(u)$ on the same side of the dial as the gnomon is called the positive half-space of $u$. All and only the points in $H^+(u)$ have a nonnegative inner product with $u$.

The characterization of duality given in theorem 6.1 then states that if $P$ and $D$ are mutually dual, then a point $v$ is in $D$ if and only if it is in all the half-spaces $H^+(u)$ as the gnomon through $u$ sweeps $P$. In other words, $P$ is the union of all the gnomons and $D$ is the intersection of all the corresponding positive half-spaces.

References

