

Separating Hyperplanes, Lagrange Multipliers, and Convex Duality

Rasmus Kyng, Scribe: Tim Taubner

Lecture 12 — Wednesday, May 13th

1 Overview

First part of this lecture introduces the concept of a separating hyperplane of two sets followed by a proof that for two closed, convex and disjoint sets a separating hyperplane always exists. This is a variant of the more general *separating hyperplane theorem*¹ due to Minkowski. Then *Lagrange multipliers* \mathbf{x} , \mathbf{s} of a convex optimization problem

$$\begin{aligned} \min_{\mathbf{y}} \mathcal{E}(\mathbf{y}) \\ \text{s.t. } \mathbf{A}\mathbf{y} = \mathbf{b} \\ c(\mathbf{y}) \leq 0 \end{aligned}$$

are introduced and with that, the *Lagrangian*

$$L(\mathbf{y}, \mathbf{x}, \mathbf{s}) = \mathcal{E}(\mathbf{y}) + \mathbf{x}^\top (\mathbf{b} - \mathbf{A}\mathbf{y}) + \mathbf{s}^\top c(\mathbf{y})$$

is defined. Finally, we deal with the dual problem

$$\max_{\mathbf{x}, \mathbf{s}, \mathbf{s} \geq 0} L(\mathbf{x}, \mathbf{s}),$$

where $L(\mathbf{x}, \mathbf{s}) = \min_{\mathbf{y}} L(\mathbf{y}, \mathbf{x}, \mathbf{s})$. We show *weak duality*, i.e. $L(\mathbf{y}, \mathbf{x}, \mathbf{s}) \leq \mathcal{E}(\mathbf{y})$ and that assuming *Slater's condition* the values of both the primal and dual is equal, which is referred to as *strong duality*.

2 Separating Hyperplane Theorem

Suppose we have two convex subsets $A, B \subseteq \mathbb{R}^n$ that are disjoint ($A \cap B = \emptyset$). We wish to show that there will always be a (hyper-)plane H that separates these two sets, i.e. A lies on one side, and B on the other side of H .

So what exactly do we mean by Hyperplane? Let's define it.

Definition 2.1 (Hyperplane). A *hyperplane* H of dimension n is the subset $H := \{\mathbf{x} \in \mathbb{R}^n : \langle \mathbf{n}, \mathbf{x} \rangle = \mu\}$. We say H has *normal* $\mathbf{n} \in \mathbb{R}^n$ and *threshold* μ . It is required that $\mathbf{n} \neq \mathbf{0}$.

Every hyperplane divides \mathbb{R}^n into two halfspaces $\{\mathbf{x} : \langle \mathbf{v}, \mathbf{x} \rangle \geq \mu\}$ and $\{\mathbf{x} : \langle \mathbf{v}, \mathbf{x} \rangle \leq \mu\}$. It separates two sets, if they lie in different halfspaces. We formally define separating hyperplane as follows.

¹Wikipedia is good on this: https://en.wikipedia.org/wiki/Hyperplane_separation_theorem

Definition 2.2 (Separating Hyperplane). We say a hyperplane H separates two sets A, B iff

$$\forall \mathbf{a} \in A : \langle \mathbf{n}, \mathbf{a} \rangle \geq \mu$$

$$\forall \mathbf{b} \in B : \langle \mathbf{n}, \mathbf{b} \rangle \leq \mu$$

If we replace \geq with $>$ and \leq with $<$ we say H strictly separates A and B .

It is easy to see that there exists disjoint non-convex sets that can not be separated by a hyperplane (e.g. a point cannot be separated from a ring around it). But can two disjoint convex sets always be strictly separated by a hyperplane? The answer is no: consider the two-dimensional case depicted in Figure 1 with $A = \{(x, y) : x \leq 0\}$ and $B = \{(x, y) : x > 0 \text{ and } y \geq \frac{1}{x}\}$. Clearly they are disjoint; however the only separating hyperplane is $H = \{(x, y) : x = 0\}$ but it intersects A .

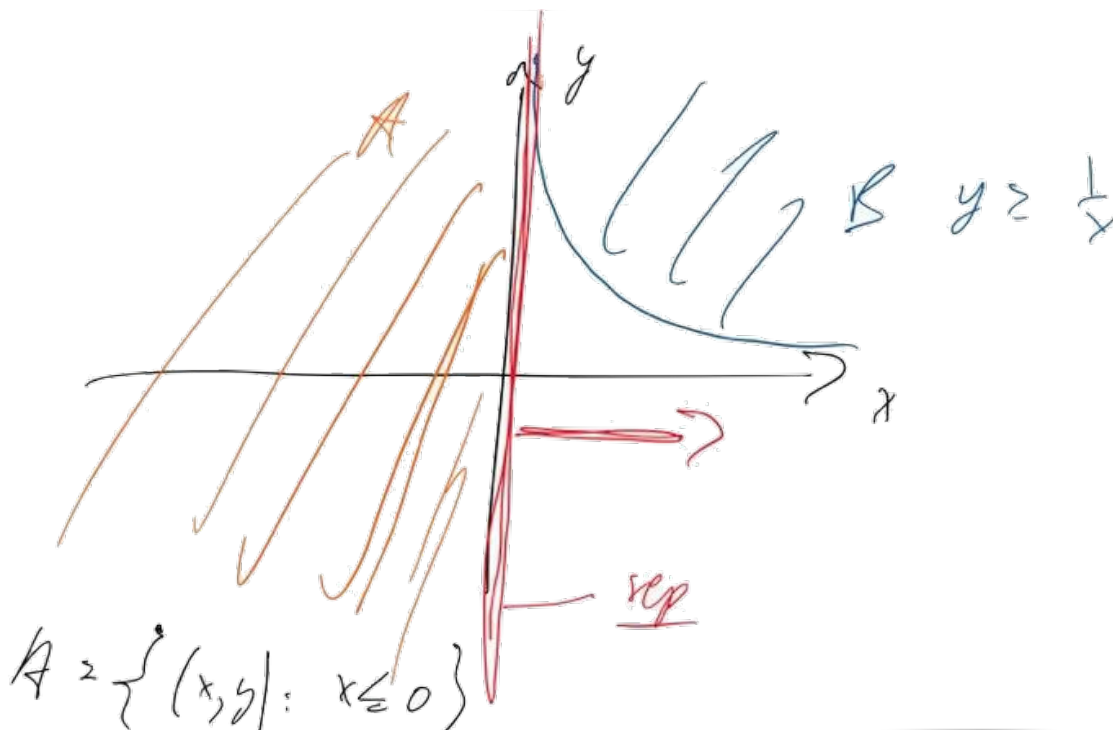


Figure 1: The sets $A = \{(x, y) : x \leq 0\}$ and $B = \{(x, y) : x > 0 \text{ and } y \geq \frac{1}{x}\}$ only permit a non-strictly separating hyperplane.

One can prove that there exists a non-strictly separating hyperplane for any two disjoint convex sets. We will prove that if we further require A, B to be closed and bounded, then a strictly separating hyperplane always exists. (Note in the example above how our choice of B is not bounded.)

Theorem 2.3 (Separating Hyperplane Theorem; closed, bounded sets). For two closed, bounded, and disjoint convex sets $A, B \in \mathbb{R}^n$ there exists a strictly separating hyperplane H . One such hyperplane is given by normal $\mathbf{n} = \mathbf{d} - \mathbf{c}$ and threshold $\mu = \frac{1}{2} (\|\mathbf{d}\|_2^2 - \|\mathbf{c}\|_2^2)$, where $\mathbf{c} \in A$, $\mathbf{d} \in B$ are the minimizers of the distance between A and B

$$\text{dist}(A, B) = \min_{\mathbf{a} \in A, \mathbf{b} \in B} \|\mathbf{a} - \mathbf{b}\|_2 > 0.$$

Proof. We omit the proof that $\text{dist}(A, B) = \min_{\mathbf{a} \in A, \mathbf{b} \in B} \|\mathbf{a} - \mathbf{b}\|_2 > 0$, which follows from A, B being disjoint, closed, and bounded. Now, we want to show that $\langle \mathbf{n}, \mathbf{b} \rangle > \mu$ for all $\mathbf{b} \in B$; then $\langle \mathbf{n}, \mathbf{a} \rangle < \mu$ for all $\mathbf{a} \in A$ follows by symmetry. Observe that

$$\begin{aligned} \langle \mathbf{n}, \mathbf{d} \rangle - \mu &= \langle \mathbf{d} - \mathbf{c}, \mathbf{d} \rangle - \frac{1}{2} \left(\|\mathbf{d}\|_2^2 - \|\mathbf{c}\|_2^2 \right) \\ &= \|\mathbf{d}\|_2^2 - \mathbf{d}^\top \mathbf{c} - \frac{1}{2} \|\mathbf{d}\|_2^2 + \frac{1}{2} \|\mathbf{c}\|_2^2 \\ &= \frac{1}{2} \|\mathbf{d} - \mathbf{c}\|_2^2 > 0. \end{aligned}$$

So suppose there exists $\mathbf{u} \in B$ such that $\langle \mathbf{n}, \mathbf{u} \rangle - \mu \leq 0$. We now look at the line defined by the distance minimizer \mathbf{d} and the point on the “wrong side” \mathbf{u} . Define $\mathbf{b}(\lambda) = \mathbf{d} + \lambda(\mathbf{u} - \mathbf{d})$, and take the derivative of the distance between $\mathbf{b}(\lambda)$ and \mathbf{c} . Evaluated at $\lambda = 0$ (which is when $\mathbf{b}(\lambda) = \mathbf{d}$), this yields

$$\left. \frac{d}{d\lambda} \|\mathbf{b}(\lambda) - \mathbf{c}\|_2^2 \right|_{\lambda=0} = 2 \langle \mathbf{d} - \lambda \mathbf{d} + \lambda \mathbf{u} - \mathbf{c}, \mathbf{u} - \mathbf{d} \rangle \Big|_{\lambda=0} = 2 \langle \mathbf{d} - \mathbf{c}, \mathbf{u} - \mathbf{d} \rangle.$$

However, this would imply that the gradient is strictly negative since

$$\begin{aligned} \langle \mathbf{n}, \mathbf{u} \rangle - \mu &= \langle \mathbf{d} - \mathbf{c}, \mathbf{u} \rangle - \langle \mathbf{d} - \mathbf{c}, \mathbf{d} \rangle + \langle \mathbf{d} - \mathbf{c}, \mathbf{d} \rangle - \mu \\ &= \langle \mathbf{d} - \mathbf{c}, \mathbf{u} - \mathbf{d} \rangle + \|\mathbf{d}\|_2^2 - \langle \mathbf{c}, \mathbf{d} \rangle - \frac{1}{2} \|\mathbf{d}\|_2^2 + \frac{1}{2} \|\mathbf{c}\|_2^2 \\ &= \langle \mathbf{d} - \mathbf{c}, \mathbf{u} - \mathbf{d} \rangle + \frac{1}{2} \|\mathbf{d} - \mathbf{c}\|_2^2 \leq 0. \end{aligned}$$

This contradicts the minimality of \mathbf{d} and thus concludes this proof. \square

A more general separating hyperplane theorem holds even when the sets are not closed and bounded:

Theorem 2.4 (Separating Hyperplane Theorem). *Given two disjoint convex sets $A, B \in \mathbb{R}^n$ there exists a hyperplane H separating them.*

3 Lagrange Multipliers and Duality of convex problems

In this Section, we’ll learn about *Lagrange Multipliers* and how they lead to convex duality. But first, let’s see an example to help illustrate where these ideas come from.

Imagine you were to prove that for all $\mathbf{x} \in \mathbb{R}^n$ we have $\|\mathbf{x}\|_p \leq n^{\frac{1}{2}-\frac{1}{p}} \|\mathbf{x}\|_2$ for some $1 \leq p \leq 2$. We can look at this as optimizing $\max_{\mathbf{x}} \|\mathbf{x}\|_p$ subject to $\|\mathbf{x}\|_2$ being constant, e.g. simply $\|\mathbf{x}\|_2 = 1$. Then the statement above follows from a scaling argument.

If we move from \mathbf{x} to $\mathbf{x} + \boldsymbol{\delta}$ with $\boldsymbol{\delta} \perp \nabla_{\mathbf{x}} \|\mathbf{x}\|_2$ and $\boldsymbol{\delta} \not\perp \nabla_{\mathbf{x}} \|\mathbf{x}\|_p$ means that for infinitesimally small $\boldsymbol{\delta}$ the 2-norm stays constant but the p -norm changes. That means for either $\mathbf{x} - \boldsymbol{\delta}$ or $\mathbf{x} + \boldsymbol{\delta}$ the p -norm while the 2-norm stays constant. Hence at the maximum of $\|\mathbf{x}\|_p$ the gradients of both norms have to be parallel, i.e.

$$\nabla_{\mathbf{x}} \left(\|\mathbf{x}\|_p - \lambda \|\mathbf{x}\|_2 \right) = 0.$$

This insight is the core idea of Lagrange multipliers (in this case λ).

Note that here the problem is not convex, as $\{\mathbf{x} : \|\mathbf{x}\|_2^2 = 1\}$ is not convex. In the following we will study Lagrange multipliers for general convex problems.

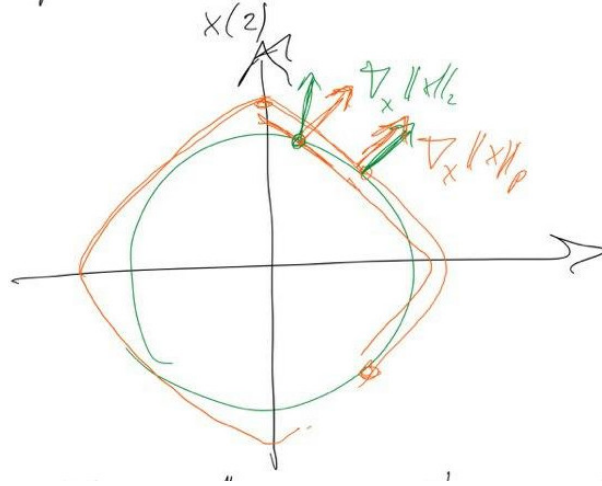


Figure 2: Looking at fixed $\|\mathbf{x}\|_p = \alpha$ and $\|\mathbf{x}\|_2 = 1$. (Here, $p = 1.5$.)

3.1 General convex problems

A full formal treatment of convex duality would require us to be more careful about using inf and sup in place of min and max, as well as considering problems that have no feasible solutions. Today, we'll ignore these concerns.

Let us consider a general convex optimization problem with convex objective, linear equality constraints and convex inequality constraints

$$\begin{aligned} \min_{\mathbf{y} \in S} \mathcal{E}(\mathbf{y}) & \quad (1) \\ \text{s.t. } \mathbf{A}\mathbf{y} &= \mathbf{b} \\ \mathbf{c}(\mathbf{y}) &\leq \mathbf{0}, \end{aligned}$$

where $\mathcal{E}(\mathbf{y}) : S \rightarrow \mathbb{R}$ is defined on a convex subset $S \subseteq \mathbb{R}^n$, $\mathbf{A} \in \mathbb{R}^{m \times n}$ and $\mathbf{c}(\mathbf{y})$ is a vector of constraints $\mathbf{c}(\mathbf{y}) = (c_i(\mathbf{y}))_{i \in [k]}$. For every $i \in [k]$ the function $c_i : S \rightarrow \mathbb{R}$ is convex.

Definition 3.1 (Primal feasibility). We say that $\mathbf{y} \in S$ is *primal feasible* if all constraints are satisfied, i.e. $\mathbf{A}\mathbf{y} = \mathbf{b}$ and $\mathbf{c}(\mathbf{y}) \leq \mathbf{0}$.

In the following we will denote by $\alpha^* = \mathcal{E}(\mathbf{y}^*)$ the optimal value of the primal program where \mathbf{y}^* is an minimizer.

Definition 3.2. Next we introduce the *dual variables* $\mathbf{x} \in \mathbb{R}^m$, $\mathbf{s} \in \mathbb{R}^k$ and define the *Lagrangian* as

$$L(\mathbf{y}, \mathbf{x}, \mathbf{s}) = \mathcal{E}(\mathbf{y}) + \mathbf{x}^\top (\mathbf{b} - \mathbf{A}\mathbf{y}) + \mathbf{s}^\top \mathbf{c}(\mathbf{y}).$$

We also define a Lagrangian only in terms of the dual variables by minimizing over \mathbf{y} as

$$L(\mathbf{x}, \mathbf{s}) = \min_{\mathbf{y}} L(\mathbf{y}, \mathbf{x}, \mathbf{s}).$$

Definition 3.3 (Dual feasibility). We say (\mathbf{x}, \mathbf{s}) is dual feasible if $\mathbf{s} \geq \mathbf{0}$. If additionally \mathbf{y} is primal feasible, we say $(\mathbf{y}, \mathbf{x}, \mathbf{s})$ is primal-dual feasible.

Definition 3.4 (Dual problem). We define the *dual problem* as

$$\max_{\substack{\mathbf{x}, \mathbf{s} \\ \mathbf{s} \geq \mathbf{0}}} \min_{\mathbf{y}} L(\mathbf{y}, \mathbf{x}, \mathbf{s}) = \max_{\substack{\mathbf{x}, \mathbf{s} \\ \mathbf{s} \geq \mathbf{0}}} L(\mathbf{x}, \mathbf{s}) \quad (2)$$

and denote the optimal dual value by β^* .

For each \mathbf{y} , the Lagrangian $L(\mathbf{y}, \mathbf{x}, \mathbf{s})$ is linear in (\mathbf{x}, \mathbf{s}) and hence also concave in them. Hence $L(\mathbf{x}, \mathbf{s})$ is a concave function, because it is the pointwise minimum (over \mathbf{y}), of a collection of concave functions in (\mathbf{x}, \mathbf{s}) .

This also means that the dual problem is really a convex optimization problem in disguise, because we can flip the sign of $-L(\mathbf{x}, \mathbf{s})$ to get a convex function and minimizing this is equivalent to maximizing $L(\mathbf{x}, \mathbf{s})$.

$$\max_{\substack{\mathbf{x}, \mathbf{s} \\ \mathbf{s} \geq \mathbf{0}}} L(\mathbf{x}, \mathbf{s}) = - \min_{\substack{\mathbf{x}, \mathbf{s} \\ \mathbf{s} \geq \mathbf{0}}} -L(\mathbf{x}, \mathbf{s})$$

3.2 Weak Duality

First we see that the primal problem can be written in terms of the Lagrangian as

$$\alpha^* = \min_{\mathbf{y}} \max_{\mathbf{x}; \mathbf{s} \geq \mathbf{0}} L(\mathbf{y}, \mathbf{x}, \mathbf{s}) \quad (3)$$

This is because for a minimizing \mathbf{y} all constraints have to be satisfied and the Lagrangian simplifies to $L(\mathbf{y}, \mathbf{x}, \mathbf{s}) = \mathcal{E}(\mathbf{y})$. If $\mathbf{A}\mathbf{x} - \mathbf{b} = \mathbf{0}$ was violated, making \mathbf{x} large sends $L(\mathbf{y}, \mathbf{x}, \mathbf{s}) \rightarrow \infty$. And if $\mathbf{c}(\mathbf{y}) \leq \mathbf{0}$ is violated, we can make $L(\mathbf{y}, \mathbf{x}, \mathbf{s}) \rightarrow \infty$ by choosing large \mathbf{s} .

Note that we require $\mathbf{s} \geq \mathbf{0}$, as we only want to penalize the violation of the inequality constraints in one direction, i.e. when $\mathbf{c}(\mathbf{y}) > \mathbf{0}$.

For any primal-dual feasible $\mathbf{y}, \mathbf{x}, \mathbf{s}$ we have $L(\mathbf{y}, \mathbf{x}, \mathbf{s}) \leq \mathcal{E}(\mathbf{y})$ and hence also $L(\mathbf{x}, \mathbf{s}) = \min_{\mathbf{y}} L(\mathbf{y}, \mathbf{x}, \mathbf{s}) \leq \mathcal{E}(\mathbf{y})$.

In other words $\max_{\mathbf{x}; \mathbf{s} \geq \mathbf{0}} L(\mathbf{x}, \mathbf{s}) = \beta^* \leq \alpha^*$. This is referred to as *weak duality*.

Using the forms in Equations (2) and (3), we can also state this as

$$\alpha^* = \min_{\mathbf{y}} \max_{\mathbf{x}; \mathbf{s} \geq \mathbf{0}} L(\mathbf{y}, \mathbf{x}, \mathbf{s}) \geq \max_{\mathbf{x}; \mathbf{s} \geq \mathbf{0}} \min_{\mathbf{y}} L(\mathbf{y}, \mathbf{x}, \mathbf{s}) = \beta^*.$$

3.3 Strong Duality

So now that we have proved weak duality $\beta^* \leq \alpha^*$, what is strong duality? $\beta^* = \alpha^*$? The answer is yes, but strong duality only holds under some conditions.

One sufficient condition we look at today is a variant of *Slater's condition*².

Definition 3.5 (Slater's condition). A (primal) problem as defined in (1) fulfills Slater's condition if there exists a *strictly feasible* point, i.e. there exists $\tilde{\mathbf{y}}$ s.t. $\mathbf{A}\tilde{\mathbf{y}} = \mathbf{b}$ and $\mathbf{c}(\tilde{\mathbf{y}}) < \mathbf{0}$. This means that the strictly feasible point $\tilde{\mathbf{y}}$ lies strictly inside the set $\{\mathbf{y} : \mathbf{c}(\mathbf{y}) \leq \mathbf{0}\}$ defined by the inequality constraints.

²Again, Wikipedia is helpful here: https://en.wikipedia.org/wiki/Slater's_condition

Theorem 3.6. *For a problem satisfying Slater's condition, strong duality holds, i.e. $\alpha^* = \beta^*$. In other words, the optimal value of the primal problem α^* is equal to the optimal value of the dual.*

How are we going to prove this? Before we prove the theorem, let's make a few observations to get us warmed up. If you get bored, skip ahead to the proof.

It is sufficient to prove that $\alpha^* \leq \beta^*$, as the statement then follows in conjunction with weak duality. We define the set

$$G = \{(\mathcal{E}(\mathbf{y}), \mathbf{A}\mathbf{y} - \mathbf{b}, \mathbf{c}(\mathbf{y})) : \mathbf{y} \in S\},$$

where $S \subseteq \mathbb{R}^n$ is the domain of \mathcal{E} .

Immediately, we observe that we can write the optimal primal value as

$$\alpha^* = \min\{t : (t, \mathbf{v}, \mathbf{u}) \in G, \mathbf{v} = \mathbf{0}, \mathbf{u} \leq \mathbf{0}\}.$$

Similarly, we can write the Lagrangian (after minimizing over \mathbf{y})

$$L(\mathbf{x}, \mathbf{s}) = \min_{(t, \mathbf{v}, \mathbf{u}) \in G} (1, \mathbf{x}, \mathbf{s})^\top (t, \mathbf{v}, \mathbf{u}).$$

This is equivalent to the inequality, for $(t, \mathbf{v}, \mathbf{u}) \in G$,

$$(1, \mathbf{x}, \mathbf{s})^\top (t, \mathbf{v}, \mathbf{u}) \geq L(\mathbf{x}, \mathbf{s}).$$

which defines a hyperplane with $\mathbf{n} = (1, \mathbf{x}, \mathbf{s})$ and $\mu = L(\mathbf{x}, \mathbf{s})$ such that G is on one side.

To establish strong duality, we would like to show the existence of a hyperplane such that for $(t, \mathbf{v}, \mathbf{u}) \in G$

$$\mathbf{n}^\top (t, \mathbf{v}, \mathbf{u}) \geq \alpha^* \text{ and } \mathbf{n} = (1, \hat{\mathbf{x}}, \hat{\mathbf{s}}) \text{ with } \hat{\mathbf{s}} \geq \mathbf{0}.$$

Then we would immediately get

$$\beta^* \geq L(\hat{\mathbf{x}}, \hat{\mathbf{s}}) = \min_{(t, \mathbf{v}, \mathbf{u}) \in G} (1, \mathbf{x}, \mathbf{s})^\top (t, \mathbf{v}, \mathbf{u}) \geq \alpha^*.$$

Perhaps not surprisingly, we will use the Separating Hyperplane Theorem. What are the challenges we need to deal with?

- We need to replace G with a convex set (which we will call A) and separate A from some other convex set (which we will call B).
- We need to make sure the hyperplane normal \mathbf{n} has 1 in the first coordinate and $\mathbf{s} \geq \mathbf{0}$, and the hyperplane threshold is α^* .

Proof of Theorem 3.6. Let's move to on finding two convex disjoint sets A, B to enable the use of the separating hyperplane Theorem 2.4.

First set we define A , roughly speaking, as a multi-dimensional epigraph of G . More precisely

$$A = \{(t, \mathbf{v}, \mathbf{u}) : \exists \mathbf{y} \in S, t \geq \mathcal{E}(\mathbf{y}), \mathbf{v} = \mathbf{A}\mathbf{y} - \mathbf{b}, \mathbf{u} \geq \mathbf{c}(\mathbf{y})\}.$$

Note that A is a convex set. The proof is similar to the proof that the epigraph of a convex function is a convex set. The optimal value of the primal program can be now written as

$$\alpha^* = \min_{(t, \mathbf{0}, \mathbf{0}) \in A} t.$$

And we define another set B of the same dimensionality as A by

$$B := \{(r \in \mathbb{R}, \mathbf{0} \in \mathbb{R}^m, \mathbf{0} \in \mathbb{R}^k) : r < \alpha^*\}.$$

This set B is convex, as it is a ray. An example of two such sets A, B is illustrated in Figure 3.

We show that $A \cap B = \emptyset$ by contradiction. Suppose A, B are not disjoint; then there exists \mathbf{y} such that

$$(\mathcal{E}(\mathbf{y}), \mathbf{A}\mathbf{y} - \mathbf{b}, \mathbf{c}(\mathbf{y})) = (r, \mathbf{0}, \mathbf{u})$$

with $\mathbf{u} \leq \mathbf{0}$. But this means that \mathbf{y} is feasible and $\mathcal{E}(\mathbf{y}) = r < \alpha^*$; contradicting the optimality of α^* .

To make things simpler, we assume that our linear constraint matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$, has full row rank and $m < n$ (but very little extra work is required to deal with the remaining cases, which we omit).

As we just proved, A and B are convex and disjoint sets and hence the separating hyperplane theorem (Theorem 2.4) we introduced earlier in this lecture implies the existence a separating hyperplane. This means there exists a normal $\mathbf{n} = (\tilde{\rho}, \tilde{\mathbf{x}}, \tilde{\mathbf{s}})$ and threshold μ and with A on one side, i.e.

$$(t, \mathbf{v}, \mathbf{u}) \in A \implies (t, \mathbf{v}, \mathbf{u})^\top (\tilde{\rho}, \tilde{\mathbf{x}}, \tilde{\mathbf{s}}) \geq \mu$$

and the set B on the other side:

$$(t, \mathbf{v}, \mathbf{u}) \in B \implies (t, \mathbf{v}, \mathbf{u})^\top (\tilde{\rho}, \tilde{\mathbf{x}}, \tilde{\mathbf{s}}) \leq \mu.$$

Now, we claim that $\tilde{\mathbf{s}} \geq \mathbf{0}$. Suppose $\tilde{\mathbf{s}}(i) < 0$, then for $\mathbf{u}(i) \rightarrow \infty$ the threshold would grow unbounded, i.e. $\mu \rightarrow -\infty$ contradicting that the threshold μ is finite by the separating hyperplane theorem. Similarly we claim $\tilde{\rho} \geq 0$, as if this were not the case, having $t \rightarrow \infty$ implies that $\mu \rightarrow -\infty$ again contradicting the finiteness of μ .

From Equation (3.3) it follows that $t\tilde{\rho} \leq \mu$ for all $t < \alpha^*$ which implies that $t\tilde{\rho} \leq \mu$ for $t = \alpha^*$ by taking the limit. Hence we have $\alpha^*\tilde{\rho} \leq \mu$. From $(t, \mathbf{v}, \mathbf{u}) \in A$ we get from Equation (3.3)

$$(\tilde{\rho}, \tilde{\mathbf{x}}, \tilde{\mathbf{s}})^\top (t, \mathbf{v}, \mathbf{u}) \geq \mu \geq \alpha^*\tilde{\rho}$$

and thus

$$(\tilde{\rho}, \tilde{\mathbf{x}}, \tilde{\mathbf{s}})^\top (\mathcal{E}(\mathbf{y}), \mathbf{A}\mathbf{y} - \mathbf{b}, \mathbf{c}(\mathbf{y})) \geq \alpha^*\tilde{\rho}.$$

Now we consider two cases; starting with the “good” case where $\tilde{\rho} > 0$. Dividing Equation (3.3) by $\tilde{\rho}$ gives

$$\mathcal{E}(\mathbf{y}) + \frac{\tilde{\mathbf{x}}^\top}{\tilde{\rho}} (\mathbf{A}\mathbf{y} - \mathbf{b}) + \frac{\tilde{\mathbf{s}}^\top}{\tilde{\rho}} \mathbf{c}(\mathbf{y}) \geq \alpha^*.$$

Noting that the left hand side above is $L(\mathbf{y}, \frac{\tilde{\mathbf{x}}}{\tilde{\rho}}, \frac{\tilde{\mathbf{s}}}{\tilde{\rho}})$ and that the equation holds for arbitrary \mathbf{y} ; therefore also for the minimum we get

$$\min_{\mathbf{y}} L\left(\mathbf{y}, \frac{\tilde{\mathbf{x}}}{\tilde{\rho}}, \frac{\tilde{\mathbf{s}}}{\tilde{\rho}}\right) \geq \alpha^*$$

and hence via definition of β^* finally

$$\beta^* \geq L \left(\frac{\tilde{\mathbf{x}}}{\tilde{\rho}}, \frac{\tilde{\mathbf{s}}}{\tilde{\rho}} \right) \geq \alpha^*.$$

Next consider the “bad” case $\tilde{\rho} = 0$. As $\alpha^* \tilde{\rho} \leq \mu$, we have $0 \leq \mu$. From Equation (3.3) we get

$$\mathbf{c}(\mathbf{y})^\top \mathbf{s} + \mathbf{x}^\top (\mathbf{b} - \mathbf{A}\mathbf{y}) \geq \mu \geq 0.$$

As Slater’s condition holds, there is an interior point $\tilde{\mathbf{y}}$, i.e. it satisfies $\mathbf{b} - \mathbf{A}\tilde{\mathbf{y}} = \mathbf{0}$ and $\mathbf{c}(\tilde{\mathbf{y}}) < \mathbf{0}$. Together with the equation above this yields

$$\mathbf{c}(\tilde{\mathbf{y}})^\top \tilde{\mathbf{s}} + \tilde{\mathbf{x}}^\top \mathbf{0} \geq 0$$

which implies $\mathbf{c}(\tilde{\mathbf{y}})^\top \tilde{\mathbf{s}} \geq 0$ and as $\mathbf{c}(\tilde{\mathbf{y}}) < \mathbf{0}$ this means $\tilde{\mathbf{s}} = \mathbf{0}$.

As the normal $(\tilde{\rho}, \tilde{\mathbf{s}}, \tilde{\mathbf{x}})$ of the hyperplane can not be all zeroes, this means the last “component” $\tilde{\mathbf{x}}$ must contain a non-zero entry, i.e. $\tilde{\mathbf{x}} \neq \mathbf{0}$. Furthermore $\tilde{\mathbf{x}}^\top (\mathbf{b} - \mathbf{A}\tilde{\mathbf{y}}) = \mathbf{0}$, $\mathbf{c}(\tilde{\mathbf{y}}) < \mathbf{0}$ and \mathbf{A} has full row rank, hence there exists $\boldsymbol{\delta}$ such that

$$\tilde{\mathbf{x}}^\top (\mathbf{b} - \mathbf{A}(\tilde{\mathbf{y}} + \boldsymbol{\delta})) < \mathbf{0} \text{ and } \mathbf{c}(\tilde{\mathbf{y}} + \boldsymbol{\delta}) < \mathbf{0}.$$

This, however, means that there is a point in A on the wrong side of the hyperplane, as

$$(\tilde{\rho}, \tilde{\mathbf{x}}, \tilde{\mathbf{s}})^\top (\mathcal{E}(\tilde{\mathbf{y}} + \boldsymbol{\delta}), \mathbf{b} - \mathbf{A}(\tilde{\mathbf{y}} + \boldsymbol{\delta}), \mathbf{c}(\tilde{\mathbf{y}} + \boldsymbol{\delta})) < 0$$

but the threshold is $\mu \geq 0$. □

Remark. Note that our reasoning about why $\mathbf{s} \geq \mathbf{0}$ in the proof above is very similar to our reasoning for why the primal program can be written as Problem (3).

Example. As an example of A and B as they appear in the above proof, consider

$$\min_{\substack{y \in (0, \infty) \\ 1/y - 1 \leq 0}} y^2$$

This leads to $\alpha^* = 1, y^* = 1$, and $A = \{(t, u) : y \in (0, \infty) \text{ and } t > y^2 \text{ and } u \geq 1/y - 1\}$, and $B = \{(t, 0) : t < 1\}$ and the separating hyperplane normal is $\mathbf{n} = (1, 2)$. These two sets A, B are illustrated in Figure 3.

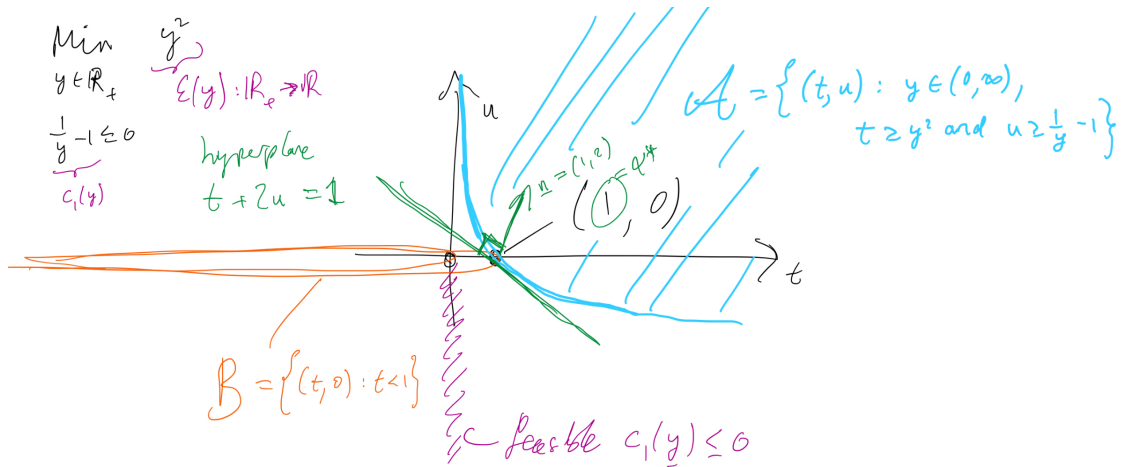


Figure 3: Example of the convex sets A and B we wish to separate by hyperplane.

3.4 The gradient perspective

Let's come back to what we said earlier about parallel gradient. Suppose \mathbf{y}^* is an optimizer of the primal problem and $\mathbf{x}^*, \mathbf{s}^*$ for the dual. We thus have

$$L(\mathbf{y}^*, \mathbf{x}^*, \mathbf{s}^*) = \alpha^* = \beta^*.$$

Because $L(\mathbf{y}, \mathbf{x}^*, \mathbf{s}^*)$ is a convex function in \mathbf{y} , it also follows that if $\mathcal{E} : S \rightarrow \mathbb{R}$ and \mathbf{c} are differentiable and the minimizer \mathbf{y}^* is not on the boundary of S , then we must have that the gradient w.r.t. \mathbf{y} is zero, i.e.

$$\nabla_{\mathbf{y}} L(\mathbf{y}, \mathbf{x}^*, \mathbf{s}^*)|_{\mathbf{y}=\mathbf{y}^*} = 0$$

and plugging in

$$L(\mathbf{y}, \mathbf{x}^*, \mathbf{s}^*) = \mathcal{E}(\mathbf{y}) + \mathbf{x}^{\top}(\mathbf{b} - \mathbf{A}\mathbf{y}) + \mathbf{s}^{\top}\mathbf{c}(\mathbf{y})$$

yields

$$\nabla \mathcal{E}(\mathbf{y}) + \mathbf{x}^{\top} \nabla_{\mathbf{y}}(\mathbf{b} - \mathbf{A}\mathbf{y}) + \mathbf{s}^{\top} \nabla \mathbf{c}(\mathbf{y}) = \mathbf{0}.$$

And this connects to our point of the parallel gradients from the beginning of this section.

3.5 Complementary Slackness

We will see more of this next time, but if we look at \mathbf{y}^* we see that

$$\mathcal{E}(\mathbf{y}^*) = \alpha^* = \mathcal{E}(\mathbf{y}^*) + \mathbf{x}^{\top}(\mathbf{b} - \mathbf{A}\mathbf{y}^*) + \mathbf{s}^{\top}\mathbf{c}(\mathbf{y}^*) = \mathcal{E}(\mathbf{y}^*) + \mathbf{s}^{\top}\mathbf{c}(\mathbf{y}^*)$$

and hence when the i -th convex constraint is not active, i.e. $c_i(\mathbf{y}^*) < 0$ the *slack* must be zero, i.e. $\mathbf{s}(i) = 0$. Conversely if the slack is non-zero, that is $\mathbf{s}(i) \neq 0$ implies that the constraint is active, i.e. $c_i(\mathbf{y}^*) = 0$.

A good reference for this is Boyd's free online book "Convex optimization" (linked to on the course website). It provides a number of different interpretations of duality. One particularly interesting one comes from economics: economists see the slack variables \mathbf{s} as prices for violating the constraints.