A second look at duality

Believe it or not, we now know enough to explore one of the key concepts in convex optimization: duality. In our first look at duality, we saw how we could recast a particular problem, find the closest point in a convex set to a fixed $\boldsymbol{x}_0$, as a maximization problem over the set of separating hyperplanes. Now we will show that the problem of minimizing a function over a convex set can be recast as a maximization of concave function created by taking the infimum of a different set of affine functions at every point.

This view of duality, called *Fenchel duality* is very general and geometric. Although mathematically equivalent to the tack we will take later using Lagrange multipliers, the geometric intuition required to grasp it is a lot different. There are pros and cons to thinking about duality this way:

Pros:

- Follows directly from basic definitions of convex sets and convex function.
- Often times can be used to reveal analytical conditions for optimality.
- Does not require any kind of smoothness of the function we are trying to minimize, nor the functions that define the constraints.

Cons:

- In general, it does not suggest too much in terms of algorithms for actually computing the solution to convex programs.
- It takes a couple hours of extreme concentration to get the geometric picture painted in your mind correctly.

Excellent supplementary reading material for this section is [BV04, Chapter 3.3] and [Lue69, Chapter 7].

The Fenchel conjugate

Let $f(\boldsymbol{x})$ be a function with domain[1] $\mathcal{C}$. The *Fenchel conjugate* of f is

$$f^*(\boldsymbol{a}) = \sup_{\boldsymbol{x}\in\mathcal{C}} \; [\langle \boldsymbol{x}, \boldsymbol{a}\rangle - f(\boldsymbol{x})] \,.$$

Notice that when f an indicator function for $\mathcal{C}$, i.e. $f(\boldsymbol{x}) = 0$ for $\boldsymbol{x} \in \mathcal{C}$ and $f(\boldsymbol{x}) = \infty$, $\boldsymbol{x} \notin \mathcal{C}$, then this is the same as the support functional $h_{\mathcal{C}}(\boldsymbol{a})$.

There are multiple ways we might interpret the Fenchel conjugate. The first follows directly from the definition above: $f^*(\boldsymbol{a})$ is simply the maximum amount that the linear functional $\langle \boldsymbol{x}, \boldsymbol{a}\rangle$ exceeds $f(\boldsymbol{x})$. The second is the intercept of the vertical axis of a supporting hyperplane of epi $f \subset \mathbb{R}^{N+1}$ with normal vector $\begin{bmatrix} \boldsymbol{a} \\ -1 \end{bmatrix}$.

Let's pause and take a look at a particular example in one dimension ($N = 1$). Suppose

$$f(x) = x^2 - 2x + 2 = (x-1)^2 + 1.$$

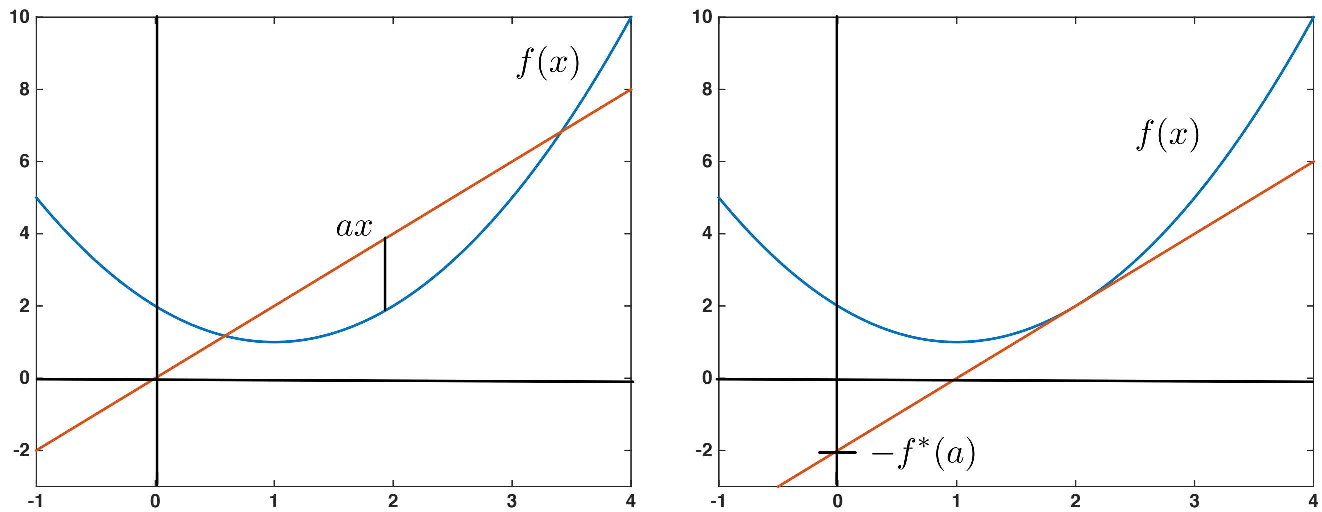
We have

$$\begin{aligned} f^*(a) &= \sup_{x\in\mathbb{R}} \left(ax - x^2 + 2x - 2\right) \\ &= \frac{a^2}{4} + a - 1. \end{aligned}$$

[1]The domain of f is simply where it is naturally defined, i.e. not blowing up to $+\infty$ or $-\infty$.

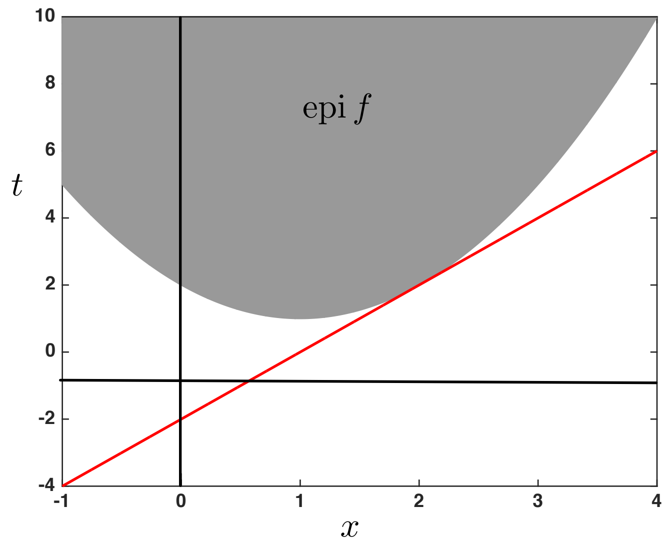
(You can verify the second equality by taking a derivative, finding the place where it equals zero, and then plugging it in.)

The figure below illustrates the situation for $a = 2$, where $f^*(2) = 2$.



On the left, we see $f(x)$ overlayed with ax for $a = 2$ (the horizontal axis is the x-axis). The *vertical* difference between the two functions is maximized at $x = 2$, and this distance is $f^*(a)$ which you can see is 2. On the write, we show $f(x)$ overlayed with ax shifted down by $f^*(a)$. By construction, this is line is now tangent to $f(x)$.

The second interpretation is illustrated below.



The epigraph of f,

$$\text{epi } f = \left\{ \begin{bmatrix} \boldsymbol{x} \\ t \end{bmatrix} \in \mathbb{R}^{N+1} \; : \; \boldsymbol{x} \in \mathcal{C}, \quad f(\boldsymbol{x}) \leq t \right\}.$$

is a subset of $\mathbb{R}^2$ — the horizontal axis above is index by x and the vertical axis by t. Then $\{\boldsymbol{x} \; : \; \langle \boldsymbol{x}, \tilde{\boldsymbol{a}} \rangle - f^*(a)\}$ is a supporting hyperplane of epi f, where $\tilde{\boldsymbol{a}} = \begin{bmatrix} a \\ -1 \end{bmatrix}$.

The conjugate function f^* is closely related to, but slightly different than, the support functional for epi f. For one thing, it is defined over $\boldsymbol{a} \in \mathbb{R}^N$, while epi $f \subset \mathbb{R}^{N+1}$. But it is the same as considering only normal vectors whose last component is -1 in the support functional for epi f:

$$f^*(\boldsymbol{a}) = \sup \left\{ \begin{bmatrix} \boldsymbol{a}^{\mathrm{T}} & -1 \end{bmatrix} \begin{bmatrix} \boldsymbol{x} \\ t \end{bmatrix}, \quad \begin{bmatrix} \boldsymbol{x} \\ t \end{bmatrix} \in \text{epi } f \right\}.$$

Similar definitions apply to concave functions $g(\boldsymbol{x}) : \mathbb{R}^N \rightarrow \mathbb{R}$ with domain $\mathcal{D}$. The conjugate function is now and infimum of linear functionals:

$$g^*(\boldsymbol{a}) = \inf_{\boldsymbol{x}\in\mathcal{D}} \left[\langle \boldsymbol{x}, \boldsymbol{a}\rangle - g(\boldsymbol{x})\right].$$

The interpretations are pretty much the same as above. The value of $g^*(\boldsymbol{a})$ is the minimum amount that the linear functional $\langle \boldsymbol{x}, \boldsymbol{a}\rangle$ exceed $g(\boldsymbol{x})$, and $-g^*(\boldsymbol{a})$ is the intercept along the vertical axis (the "t coordinate") of a supporting hyperplane of the hypograph of g:

$$\operatorname{hypo} g = \left\{ \begin{bmatrix} \boldsymbol{x} \\ t \end{bmatrix} \in \mathbb{R}^{N+1} \ : \ \boldsymbol{x} \in \mathcal{D},\ g(x) \geq t \right\}.$$

Let's look at a specific example of computing the conjugate of a concave function with $N = 1$. Take

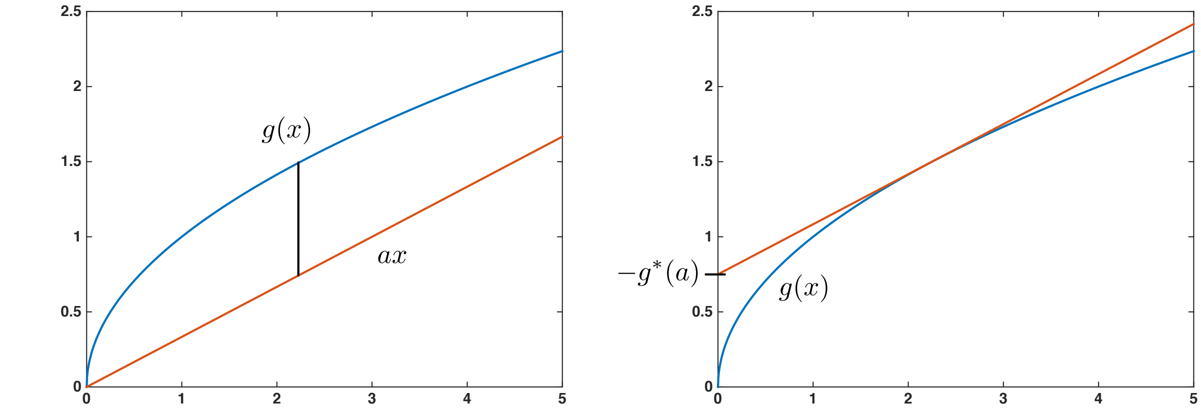
$$g(x) = \sqrt{x}, \quad \text{on} \ \ \{x \ : \ x \geq 0\} =: \mathcal{D}.$$

Then

$$\begin{aligned} g^*(\boldsymbol{a}) &= \inf_{x\geq 0} \left[ax - \sqrt{x}\right] \\ &= -\frac{1}{4a}. \end{aligned}$$

The last equality follows from taking a derivative with respect to x, and realizing it is zero when $x = 1/(4a^2)$, and then plugging this in to the expression in the brackets above.

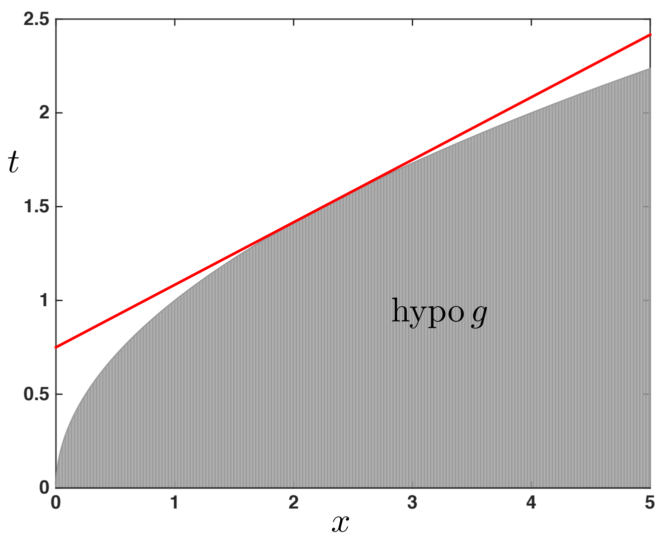
Here are plots for $a = 1/3$. In this case, $g^*(1/3) = -3/4$, and the maximal difference occurs at $x = 9/4$.



Again, the hyperplane

$$\{(x,t)\ :\ ax - t = g^*(a)\}$$

supports $\operatorname{hypo} g$:



Key examples

ℓ_1 **norm:** If $f(\boldsymbol{x}) = \|\boldsymbol{x}\|_1$, then

$$\begin{aligned} f^*(\boldsymbol{a}) &= \sup_{\boldsymbol{x}\in\mathbb{R}^N} \langle \boldsymbol{x}, \boldsymbol{a}\rangle - \|\boldsymbol{x}\|_1 \\ &= \sup_{\boldsymbol{x}} \sum_{n=1}^{N} a_n x_n - |x_n| \\ &= \begin{cases} 0, & \|\boldsymbol{a}\|_\infty \leq 1 \\ \infty, & \text{otherwise.} \end{cases} \end{aligned}$$

so $f^*(\boldsymbol{a}) = 0$ and $\operatorname{dom} f^* = \mathcal{C}^* = \{\boldsymbol{a} \in \mathbb{R}^N \ : \ \|\boldsymbol{a}\|_\infty \leq 1\}$.

ℓ_∞ **norm:** If $f(\boldsymbol{x}) = \|\boldsymbol{x}\|_\infty$, then

$$f^*(\boldsymbol{a}) = \sup_{\boldsymbol{x}\in\mathbb{R}^N} \sum_{n=1}^{N} a_n x_n - \max_n |x_n|.$$

We know that

$$\sum_{n=1}^{N} a_n x_n \leq \sum_{n=1}^{N} |a_n|\,|x_n| \ \leq \ \max_n |x_n|\, \|\boldsymbol{a}\|_1,$$

and so

$$\begin{aligned} f^*(\boldsymbol{a}) &\leq \sup_{\boldsymbol{x}} \|\boldsymbol{x}\|_\infty (\|\boldsymbol{a}\|_1 - 1) \\ &= \begin{cases} 0, & \|\boldsymbol{a}\|_1 \leq 1, \\ \infty, & \text{otherwise.} \end{cases} \end{aligned}$$

Moreover, it is easy to see that $f^*(\boldsymbol{a}) \geq 0$ when $\|\boldsymbol{a}\|_1 \leq 1$, as $\langle \mathbf{0}, \boldsymbol{a}\rangle - \|\mathbf{0}\|_\infty = 0$. Also, when $\|\boldsymbol{a}\|_1 > 1$, $\langle \boldsymbol{x}, \boldsymbol{a}\rangle - \|\boldsymbol{x}\|_\infty$ is unbounded, as we can see by taking $x_n = \alpha \operatorname{sgn} a_n$, and letting $\alpha \to \infty$. Thus

$$f^*(\boldsymbol{a}) = \begin{cases} 0, & \|\boldsymbol{a}\|_1 \le 1, \\ \infty, & \text{otherwise.} \end{cases}$$

For lots of other examples, see [BV04, Chapter 3.3].

Key properties

- $f^*(\boldsymbol{a})$ is convex even when $f(\boldsymbol{x})$ is not. This is because $f^*(\boldsymbol{a})$ is a pointwise supremum of affine functions (see the homework).
- If $f(\boldsymbol{x})$ is convex and has a closed epigraph, then we can recover $f(\boldsymbol{x})$ by again taking the conjugate of $f^*(\boldsymbol{a})$:

  $$f^{**}(\boldsymbol{x}) = \sup_{\boldsymbol{a} \in \mathcal{C}^*} \left[\langle \boldsymbol{a}, \boldsymbol{x} \rangle - f^*(\boldsymbol{a}) \right] = f(\boldsymbol{x}),$$

  where $\mathcal{C}^*$ is the domain of f^* (see the homework).
- If $f(\boldsymbol{x}_1, \boldsymbol{x}_2)$ can be written as the sum of two independent variables:

  $$f(\boldsymbol{x}_1, \boldsymbol{x}_2) = f_1(\boldsymbol{x}_1) + f_2(\boldsymbol{x}_2),$$

  then

  $$f^*(\boldsymbol{a}_1, \boldsymbol{a}_2) = f_1^*(\boldsymbol{a}_1) + f_2^*(\boldsymbol{a}_2).$$

For a few other properties, see [BV04, Chapter 3.3].

Fenchel duality

We consider now the general problem of computing the minimum of a convex function $f(\boldsymbol{x})$ with $\operatorname{dom} f = \mathcal{C}$ subject to the additional constraints that $\boldsymbol{x} \in \mathcal{D}$, where $\mathcal{D}$ is also convex[2]:

$$\inf_{\boldsymbol{x} \in \mathcal{C} \cap \mathcal{D}} f(\boldsymbol{x}).$$

Suppose the following technical conditions hold:

- the quantity $\inf_{\boldsymbol{x} \in \mathcal{C} \cap \mathcal{D}} f(\boldsymbol{x})$ is finite;
- either epi f or $\mathcal{D}$ has a non-empty interior;
- the intersection $\mathcal{C} \cap \mathcal{D}$ contains points in the relative interior[3] of both $\mathcal{C}$ and $\mathcal{D}$.

Then the **Fenchel Duality Theorem** states

$$\inf_{\boldsymbol{x} \in \mathcal{C} \cap \mathcal{D}} f(\boldsymbol{x}) = \max_{\boldsymbol{a} \in \mathcal{C}^* \cap \mathcal{D}^*} \left[h'_{\mathcal{D}}(\boldsymbol{a}) - f^*(\boldsymbol{a}) \right],$$

where $h'_{\mathcal{D}}(\boldsymbol{a})$ is related to the support function of $\mathcal{D}$, just with an infimum instead of a supremem:

$$h'_{\mathcal{D}}(\boldsymbol{a}) = \inf_{\boldsymbol{x} \in \mathcal{D}} \langle \boldsymbol{x}, \boldsymbol{a} \rangle = -\sup_{\boldsymbol{x} \in \mathcal{D}} \langle \boldsymbol{x}, -\boldsymbol{a} \rangle = -h_{\mathcal{D}}(-\boldsymbol{a}).$$

We write "inf" on the left and "max" on the right since we can also show that there is a $\boldsymbol{a}^\star$ that achieves this max, even when there is no

[2] A word on notation here. When we write "$\text{minimize}_{\boldsymbol{x} \in \mathcal{X}} f(\boldsymbol{x})$" we are referring to the computational task of finding an $\boldsymbol{x} \in \mathcal{X}$ that makes $f(\boldsymbol{x})$ as small as possible — this is an optimization program. When we write $\min_{\boldsymbol{x} \in \mathcal{X}} f(\boldsymbol{x})$ or $\inf_{\boldsymbol{x} \in \mathcal{X}} f(\boldsymbol{x})$, we are referring to the actual smallest value of $f(\boldsymbol{x})$ on $\mathcal{X}$ — this is a real number. We write inf instead of min (or sup instead of max) when we are unsure if there is a minimizer (maximizer).

[3] We didn't cover the concept of relative interior in the notes, but it basically means the interior of a set relative to an affine set which contains it. So, for example, the affine set $\{\boldsymbol{x} : \boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}\}$ does not have any interior points if $\boldsymbol{A}$ does not have full row rank, every point is in the relative interior. See [BV04, p. 23] for details.

$\boldsymbol{x}$ that achieves the infimum. The set $\mathcal{D}^*$ above refers to the domain of $h'_{\mathcal{D}}(\boldsymbol{a})$.

There is an analogous statement for maximizing concave functions. If $g(\boldsymbol{x})$ is concave with domain $\mathcal{C}$ and we want to maximize it subject to the constraint $\boldsymbol{x} \in \mathcal{D}$, we have

$$\sup_{\boldsymbol{x}\in\mathcal{C}\cap\mathcal{D}} g(\boldsymbol{x}) = \min_{\boldsymbol{a}\in\mathcal{C}^*\cap\mathcal{D}^*} \left[h_{\mathcal{D}}(\boldsymbol{a}) - g^*(\boldsymbol{a})\right].$$

The $h_{\mathcal{D}}$ above is the standard support function of $\mathcal{D}$:

$$h_{\mathcal{D}}(\boldsymbol{a}) = \sup_{\boldsymbol{x}\in\mathcal{D}} \langle \boldsymbol{x}, \boldsymbol{a}\rangle,$$

and $\mathcal{D}^*$ is the domain of this function.

This result has massive implications, but its proof is not terribly hard. We will sketch it here. In fact, we will consider the more general problem of minimizing the difference between a convex function $f(\boldsymbol{x})$ with domain $\mathcal{C}$ and and a concave function $g(\boldsymbol{x})$ with domain $\mathcal{D}$:

$$\inf_{\boldsymbol{x}\in\mathcal{C}\cap\mathcal{D}} \left[f(\boldsymbol{x}) - g(\boldsymbol{x})\right] = \max_{\boldsymbol{a}\in\mathcal{C}^*\cap\mathcal{D}^*} \left[g^*(\boldsymbol{a}) - f^*(\boldsymbol{a})\right].$$

The first result above simply took $g(\boldsymbol{x})$ to be the concave indicator function of $\mathcal{D}$:

$$\text{(special case above)} \quad g(\boldsymbol{x}) = \begin{cases} 0, & \boldsymbol{x} \in \mathcal{D}, \\ -\infty, & \boldsymbol{x} \notin \mathcal{D}. \end{cases}$$

That the inf expression above upper bounds the sup expression is straightforward. By definition for *all* $\boldsymbol{x} \in \mathcal{C}\cap\mathcal{D}$ and *all* $\boldsymbol{a} \in \mathcal{C}^*\cap\mathcal{D}^*$,

$$\begin{aligned} f^*(\boldsymbol{a}) &\geq \langle \boldsymbol{x}, \boldsymbol{a}\rangle - f(\boldsymbol{x}), \\ g^*(\boldsymbol{a}) &\leq \langle \boldsymbol{x}, \boldsymbol{a}\rangle - g(\boldsymbol{x}), \end{aligned}$$

and so

$$f(\boldsymbol{x}) - g(\boldsymbol{x}) \geq g^*(\boldsymbol{a}) - f^*(\boldsymbol{a}).$$

This this is true for all $\boldsymbol{x}$ and $\boldsymbol{a}$,

$$\inf_{\boldsymbol{x}\in\mathcal{C}\cap\mathcal{D}} \left[f(\boldsymbol{x}) - g(\boldsymbol{x})\right] \;\geq\; \sup_{\boldsymbol{a}\in\mathcal{C}^*\cap\mathcal{D}^*} \left[g^*(\boldsymbol{a}) - f^*(\boldsymbol{a})\right].$$

Therefore, equality will hold (and the theorem will be proven) if we can find a particular $\boldsymbol{a}^\star$ such that

$$\inf_{\boldsymbol{x}\in\mathcal{C}\cap\mathcal{D}} \left[f(\boldsymbol{x}) - g(\boldsymbol{x})\right] = g^*(\boldsymbol{a}^\star) - f^*(\boldsymbol{a}^\star).$$

We can find such a $\boldsymbol{a}^\star$ using the separating hyperplane theorem. For simplicity here, we will just sketch the rest of the proof[4]. Let $\mu = \inf_{\mathcal{C}\cap\mathcal{D}} f(\boldsymbol{x}) - g(\boldsymbol{x})$.[5] Consider the set

$$\operatorname{epi}_{-\mu} f = \left\{ \begin{bmatrix}\boldsymbol{x}\\ t\end{bmatrix} \in \mathbb{R}^{N+1} \;:\; \begin{bmatrix}\boldsymbol{x}\\ t+\mu\end{bmatrix} \in \operatorname{epi} f \right\},$$

which is the epigraph of f shifted down the vertical axis by μ. By construction, the vertical distance between $\operatorname{epi}_{-\mu} f$ and $\operatorname{hypo} g$ is zero, yet their relative interiors are disjoint. This means we can find a (nonvertical) hyperplane which separates[6] them, but both $\operatorname{epi}_{-\mu} f$ and $\operatorname{hypo} g$ will be arbitrarily close to this hyperplane. That is, there is a $\boldsymbol{a}^\star, b^\star$ such that

$$\begin{bmatrix}\boldsymbol{a}^\star & -1\end{bmatrix} \begin{bmatrix}\boldsymbol{x}\\ t\end{bmatrix} \;\leq\; b^\star, \quad \begin{bmatrix}\boldsymbol{x}\\ t\end{bmatrix} \in \operatorname{epi}_{-\mu} f,$$

[4]For the full argument, see [Lue69, Ch. 7.12] (or [Roc70] or any text on convex analysis).

[5]This is where we need the first condition above.

[6]This is where we need the second condition above.

and

$$\begin{bmatrix}\boldsymbol{a}^\star & -1\end{bmatrix}\begin{bmatrix}\boldsymbol{x}\\ t\end{bmatrix} \;\geq\; b^\star, \quad \begin{bmatrix}\boldsymbol{x}\\ t\end{bmatrix} \in \operatorname{hypo} g,$$

and because both sets are arbitrarily close[7] to the hyperplane,

$$\begin{aligned} b^\star &= \sup_{(\boldsymbol{x},t)\in \operatorname{epi}_{-\mu} f} \begin{bmatrix}\boldsymbol{a}^\star & -1\end{bmatrix}\begin{bmatrix}\boldsymbol{x}\\ t\end{bmatrix} \\ &= \sup_{(\boldsymbol{x},t)\in \operatorname{epi}_{-\mu} f} \langle \boldsymbol{x}, \boldsymbol{a}^\star\rangle - t \\ &= \sup_{\boldsymbol{x}\in\mathcal{C}} \langle \boldsymbol{x}, \boldsymbol{a}^\star\rangle - f(\boldsymbol{x}) + \mu \\ &= f^*(\boldsymbol{a}^\star) + \mu, \end{aligned}$$

and similarly

$$\begin{aligned} b^\star &= \inf_{(\boldsymbol{x},t)\in \operatorname{hypo} g} \begin{bmatrix}\boldsymbol{a}^\star & -1\end{bmatrix}\begin{bmatrix}\boldsymbol{x}\\ t\end{bmatrix} \\ &= \inf_{\boldsymbol{x}\in\mathcal{D}} \langle \boldsymbol{x}, \boldsymbol{a}^\star\rangle - g(\boldsymbol{x}) \\ &= g^*(\boldsymbol{a}^\star). \end{aligned}$$

Thus

$$g^*(\boldsymbol{a}^\star) - f^*(\boldsymbol{a}^\star) = \mu = \inf_{\boldsymbol{x}\in\mathcal{C}\cap\mathcal{D}} f(\boldsymbol{x}) - g(\boldsymbol{x}).$$

Notice that we have found a $\boldsymbol{a}^\star$ that achieves the supremum; thus we can replace sup with max as in the theorem statement.

Picture:

[7] This is where we need the third condition about the relative interiors.

Super-Easy Example

Before we look at serious applications of this result, let's look at a very simple example just to get a feel for the computations involved. We will compute

$$\inf_{x \in [3,5]} x^3.$$

Of course, we know the answer already: it is 27, as the function above achieves its minimum value at $\hat{x} = 3$. But let's verify the Fenchel duality theorem for this case.

We will take $f(x) = x^3$ which is convex over the non-negative reals, so we take $\mathcal{C} = \{x : x \geq 0\}$. The constraint set is the interval $\mathcal{D} = [3, 5]$. First we compute the support functional,

$$h'_{\mathcal{D}}(a) = \inf_{x \in [3,5]} \; ax \quad = \quad \begin{cases} 3a, & a \geq 0, \\ 5a, & a < 0. \end{cases}$$

The conjugate of f is

$$f^*(a) = \sup_{x \geq 0} \; (ax - x^3).$$

For fixed $a \geq 0$, this expression is maximized at $x^* = \sqrt{a/3}$; for $a < 0$ it is maximized at $x^* = 0$. Thus

$$f^*(a) = \begin{cases} \frac{2}{3}\sqrt{\frac{a^3}{3}}, & a \geq 0 \\ 0, & a < 0. \end{cases}$$

Thus

$$\max_{a \in \mathbb{R}} \left[h_{\mathcal{D}}(a) - f^*(a)\right] \quad = \quad \max_{a \in \mathbb{R}} \begin{cases} 3a - \frac{2}{3}\sqrt{\frac{a^3}{3}}, & a \geq 0, \\ 5a, & a < 0. \end{cases}$$

It is easy to check that this expression is maximized at $a^* = 27$ (coincidence), and that

$$\left(3a - \frac{2}{3}\sqrt{\frac{a^3}{3}}\right)\Bigg|_{a=27} = 27.$$

Example: Resource allocation [Lue69]

The "law of diminishing returns" is a fundamental tenet of economics: as we put more and more resources into something, at some point, the incremental gains become less and less. You see this everywhere: what is the difference between spending \$5 on dinner, \$50 on dinner, \$500 on dinner? What are the differences between a \$50 bicycle, a \$500 bicycle, and a \$5000 bicycle?

What this means is that functions $g(x)$ that map resources to return are concave.

Suppose we have D dollars that we would like to allocate to N different activities in such a way that maximizes the return. The return of each activity is a (possibly different) concave function $g_n(x_n)$, where x_n is the amount of money invested. Our optimization problem is

$$\underset{\boldsymbol{x}\in\mathbb{R}^N}{\text{maximize}} \;\; g(\boldsymbol{x}) = \sum_{n=1}^{N} g_n(x_n) \quad \text{subject to} \quad \sum_{n=1}^{N} x_n = D, \quad \boldsymbol{x} \geq \boldsymbol{0}.$$

This is an optimization problem in N variables, and of course its solution depends on what we actually choose for the return functions $g_n(x_n)$. However, by using Fenchel duality, we can recast this problem as an optimization in a single variable.

Since the natural domain of the g_n is $x \geq 0$, let's take

$$\mathcal{C} = \{\boldsymbol{x} : \boldsymbol{x} \geq \mathbf{0}\}, \quad \mathcal{D} = \{\boldsymbol{x} : x_1 + \cdots + x_N = D\}.$$

We start by computing the support functional for $\mathcal{D}$:

$$h_{\mathcal{D}}(\boldsymbol{a}) = \sup_{\boldsymbol{x} \in \mathcal{D}} \langle \boldsymbol{x}, \boldsymbol{a} \rangle.$$

Since $\mathcal{D}$ is itself an affine set, $h_{\mathcal{D}}(\boldsymbol{a})$ is infinite for almost every $\boldsymbol{a}$ we plug in — the exception is if all of the entries of $\boldsymbol{a}$ are equal to one another. In this case,

$$\boldsymbol{a} = \lambda \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}, \quad h_{\mathcal{D}}(\boldsymbol{a}) = \lambda D.$$

Thus

$$\mathcal{D}^* = \text{Range}(\mathbf{1}),$$

where $\mathbf{1}$ is an N-vector of all ones.

Now we compute the conjugate $g^*(\boldsymbol{a})$. Since g is a sum of concave functions of independent variables,

$$g^*(\boldsymbol{a}) = \sum_{n=1}^{N} g_n^*(a_n),$$

where $g_n^*(a_n)$ is the conjugate of a function of a single variable:

$$g_n^*(a) = \inf_{x \geq 0} \left[ax - g_n(x)\right].$$

This means we can write the dual as

$$\min_{\boldsymbol{a}} \left[h_{\mathcal{D}}(\boldsymbol{a}) - g^*(\boldsymbol{a})\right] \quad = \quad \min_{\lambda \in \mathbb{R}} \left[\lambda D - \sum_{n=1}^{N} g_n^*(\lambda)\right].$$

That is, the expression to be minimized is a function **of a single variable** λ. All we need to know how to do is evaluate the conjugate functions g_n^*.

Example: ℓ_1 minimization

Let's look at an example of how we can apply the Fenchel duality theorem to learn something about the solution to a particular constrained optimization program. Consider the so-called "basis pursuit" program:

$$\min_{\boldsymbol{x}} \ \|\boldsymbol{x}\|_1 \quad \text{subject to} \quad \boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}, \tag{1}$$

where the vector $\boldsymbol{y} \in \mathbb{R}^M$ and matrix $\boldsymbol{A} \in \mathbb{R}^{M\times N}$ are given, $M < N$, and $\boldsymbol{A}$ has full row rank (so $\operatorname{rank}(\boldsymbol{A}) = M$). In this case, for any $\boldsymbol{y}$, there are many $\boldsymbol{x}$ with $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}$. We take $\mathcal{D}$ to be the vectors in $\mathbb{R}^N$ which obey these constraints:

$$\begin{aligned}\mathcal{D} &= \{\boldsymbol{x} \ : \ \boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}\} \\ &= \{\boldsymbol{x} \ : \ \boldsymbol{x} = \boldsymbol{v}_0 + \boldsymbol{v}, \quad \boldsymbol{v} \in \operatorname{Null}(\boldsymbol{A})\},\end{aligned}$$

where $\boldsymbol{v}_0$ is any feasible point (meaning $\boldsymbol{A}\boldsymbol{v}_0 = \boldsymbol{y}$); we might take

$$\boldsymbol{v}_0 = \boldsymbol{A}^{\mathrm{T}}(\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}})^{-1}\boldsymbol{y},$$

for instance. We have

$$\begin{aligned}h'_{\mathcal{D}}(\boldsymbol{a}) &= \inf_{\boldsymbol{x}\in\mathcal{D}} \langle \boldsymbol{x}, \boldsymbol{a}\rangle \\ &= \inf_{\boldsymbol{v}\in\operatorname{Null}(\boldsymbol{A})} \langle \boldsymbol{v}_0, \boldsymbol{a}\rangle + \langle \boldsymbol{v}, \boldsymbol{a}\rangle.\end{aligned}$$

The only way the infimum above is not $-\infty$ is if $\langle \boldsymbol{v}, \boldsymbol{a} \rangle = 0$ for all $\boldsymbol{v} \in \text{Null}(\boldsymbol{A})$ — that is, $\boldsymbol{a} \in \text{Range}(\boldsymbol{A}^{\mathrm{T}})$, since the row space of a matrix is orthogonal to its column space. Thus

$$h'_{\mathcal{D}}(\boldsymbol{a}) = \langle \boldsymbol{v}_0, \boldsymbol{a} \rangle, \quad \mathcal{D}^* = \text{Range}(\boldsymbol{A}^{\mathrm{T}}).$$

We also take $f(\boldsymbol{x}) = \|\boldsymbol{x}\|_1$, $\mathcal{C} = \mathbb{R}^N$, and $g(\boldsymbol{x}) = 0$ for $\boldsymbol{x} \in \mathcal{D}$. The conjugate function is

$$f^*(\boldsymbol{a}) = 0, \quad \operatorname{dom} f^* = \mathcal{C}^* = \{\boldsymbol{a} \ : \ \|\boldsymbol{a}\|_\infty \leq 1\}.$$

The dual problem is then

$$\begin{aligned}
\max_{\boldsymbol{a} \in \mathcal{C}^* \cap \mathcal{D}^*} h'_{\mathcal{D}}(\boldsymbol{a}) - f^*(\boldsymbol{a}) &= \max_{\boldsymbol{a} \in \mathbb{R}^N} \ \langle \boldsymbol{v}_0, \boldsymbol{a} \rangle \quad \text{subject to} \ \ \boldsymbol{a} \in \text{Range}(\boldsymbol{A}^{\mathrm{T}}), \\
&\qquad\qquad\qquad\qquad\qquad\qquad\quad \|\boldsymbol{a}\|_\infty \leq 1, \\
&= \max_{\boldsymbol{w} \in \mathbb{R}^M} \ \langle \boldsymbol{v}_0, \boldsymbol{A}^{\mathrm{T}} \boldsymbol{w} \rangle \quad \text{subject to} \ \ \|\boldsymbol{A}^{\mathrm{T}} \boldsymbol{w}\|_\infty \leq 1 \\
&= \max_{\boldsymbol{w} \in \mathbb{R}^M} \ \langle \boldsymbol{y}, \boldsymbol{w} \rangle \quad \text{subject to} \ \ \|\boldsymbol{A}^{\mathrm{T}} \boldsymbol{w}\|_\infty \leq 1.
\end{aligned}$$

Now, suppose that $\boldsymbol{x}^\star$ is a minimizer of the primal (1). We know that the optimal values of the primal and dual programs are equal to one another, and so

$$\begin{aligned}
\|\boldsymbol{x}^\star\|_1 &= \max_{\boldsymbol{w}} \ \langle \boldsymbol{A}\boldsymbol{x}^\star, \boldsymbol{w} \rangle \quad \text{subject to} \ \ \|\boldsymbol{A}^{\mathrm{T}} \boldsymbol{w}\|_\infty \leq 1 \\
&= \max_{\boldsymbol{w}} \ \langle \boldsymbol{x}^\star, \boldsymbol{A}^{\mathrm{T}} \boldsymbol{w} \rangle \quad \text{subject to} \ \ \|\boldsymbol{A}^{\mathrm{T}} \boldsymbol{w}\|_\infty \leq 1.
\end{aligned}$$

This immediately tells us that if $\Gamma^\star \subset \{1, \ldots, N\}$ is the set of indexes where $\boldsymbol{x}^\star$ is nonzero,

$$\begin{aligned}
(\boldsymbol{A}^{\mathrm{T}} \boldsymbol{w})_n &= \operatorname{sign} x^\star[n], \quad && n \in \Gamma^\star, \\
|(\boldsymbol{A}^{\mathrm{T}} \boldsymbol{w})_n| &\leq 1, && n \notin \Gamma^\star.
\end{aligned}$$

This tells us two interesting things. First, whether or not $\boldsymbol{x}^\star$ is the solution to (1) depends only on the support of $\boldsymbol{x}^\star$ and the sign sequence on that support — if $\boldsymbol{x}^\star$ is a solution for $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}^\star$, then another vector $\boldsymbol{x}_1$ that is supported on any subset of $\Gamma^\star$ with the same sign sequence as $\boldsymbol{x}^\star$ will be a solution for $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{a}_1$.

Second, it tells us (after some thought) that there is always a solution to (1) with at most M non-zero components. Why? Well, if a solution $\boldsymbol{x}^\star$ is supported on a set $\Gamma^\star$ of size $|\Gamma^\star| > M$, then there is another solution $\boldsymbol{x}'$ supported on a proper subset of $\Gamma^\star$ ... (I will let you finish this argument at home).

References

[BV04] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[Lue69] D. G. Luenberger. *Optimization by Vector Space Methods*. Wiley, 1969.

[Roc70] R. T. Rockafellar. *Convex Analysis*. Princeton University Press, 1970.