A second look at duality

Our first exposure to duality was in the context of constrained optimization: by introducing dual variables (Lagrange multipliers), we can combine the objective function and constraints into a single (Lagrangian) function which we can optimize either by minimizing it with respect to the primal variables or maximizing it with respect to the dual variables.

However, duality is a much broader concept than what we have seen so far, and can even be relevant in unconstrained problems.[1] As an example, suppose we wish to minimize the sum of two functions:

$$\underset{\boldsymbol{x}\in\mathbb{R}^N}{\text{minimize}} \;\; f(\boldsymbol{x}) + g(\boldsymbol{x}) \tag{1}$$

where f and g are both convex. For simplicity, we will assume that $\operatorname{dom} f = \operatorname{dom} g = \mathbb{R}^N$, although we could easily extend this to the case where the domain is a subset of $\mathbb{R}^N$ by replacing f and/or g with their extension to $\mathbb{R}^N$. This problem is unconstrained, but we can actually represent it as a constrained problem of the form:

$$\underset{\boldsymbol{x},\boldsymbol{z}\in\mathbb{R}^N}{\text{minimize}} \;\; f(\boldsymbol{x}) + g(\boldsymbol{z}) \quad \text{subject to} \quad \boldsymbol{x} = \boldsymbol{z}.$$

This formulation involves N affine equality constraints on the primal variables $\boldsymbol{x}$ and $\boldsymbol{z}$ of the form $h_n(\boldsymbol{x},\boldsymbol{z}) = z_n - x_n = 0$. The Lagrangian function for this problem is

$$\mathcal{L}(\boldsymbol{x},\boldsymbol{z},\boldsymbol{\nu}) = f(\boldsymbol{x}) + g(\boldsymbol{z}) + \langle \boldsymbol{z}-\boldsymbol{x}, \boldsymbol{\nu}\rangle.$$

[1] Moreover, as we will soon see, if you give up on the requirement that your objective function is differentiable, the distinction between constrained and unconstrained problems becomes a bit blurred.

To derive the dual problem, we first must compute the dual function:

$$\begin{aligned} d(\boldsymbol{\nu}) &= \inf_{\boldsymbol{x},\boldsymbol{z}\in\mathbb{R}^N} \ f(\boldsymbol{x}) + g(\boldsymbol{z}) + \langle \boldsymbol{z} - \boldsymbol{x}, \boldsymbol{\nu} \rangle \\ &= \inf_{\boldsymbol{x}\in\mathbb{R}^N} \ f(\boldsymbol{x}) - \langle \boldsymbol{x}, \boldsymbol{\nu} \rangle + \inf_{\boldsymbol{z}\in\mathbb{R}^N} \ g(\boldsymbol{z}) - \langle \boldsymbol{z}, -\boldsymbol{\nu} \rangle \\ &= -\underbrace{\sup_{\boldsymbol{x}\in\mathbb{R}^N} \ \langle \boldsymbol{x}, \boldsymbol{\nu} \rangle - f(\boldsymbol{x})}_{f^*(\boldsymbol{\nu})} - \underbrace{\sup_{\boldsymbol{z}\in\mathbb{R}^N} \ \langle \boldsymbol{z}, -\boldsymbol{\nu} \rangle - g(\boldsymbol{z})}_{g^*(-\boldsymbol{\nu})}, \end{aligned}$$

where

$$f^*(\boldsymbol{\nu}) := \sup_{\boldsymbol{x}\in\mathbb{R}^N} \ \langle \boldsymbol{x}, \boldsymbol{\nu} \rangle - f(\boldsymbol{x})$$

is the **convex conjugate** (or **Fenchel conjugate**) of f. We will try to give a bit more intuition into what the convex conjugate of a function represents below, but first we note that if we can calculate this function, then the resulting dual problem is

$$\underset{\boldsymbol{\nu}\in\mathbb{R}^N}{\text{maximize}} \ -f^*(\boldsymbol{\nu}) - g^*(-\boldsymbol{\nu})$$

or equivalently

$$\underset{\boldsymbol{\nu}\in\mathbb{R}^N}{\text{minimize}} \ f^*(\boldsymbol{\nu}) + g^*(-\boldsymbol{\nu}). \tag{2}$$

Before we can see this in action on some real optimization problems, however, we first need to understand what the convex conjugate is and, given a function f, how to actually compute f^*.

The convex conjugate

Before going any further, the first thing to say about the convex conjugate is that it is, as its name might suggest, convex. In fact, $f^*(\boldsymbol{\nu})$ is convex, *even if* $f(\boldsymbol{x})$ *is not*. We have seen the argument for this before: $f^*(\boldsymbol{\nu})$ is the pointwise supremum of convex functions since for any fixed $\boldsymbol{x}$, $\langle \boldsymbol{x}, \boldsymbol{\nu} \rangle - f(\boldsymbol{x})$ is a convex (affine) function).

The convex conjugate plays a fundamental role in duality. Before we assumed that $\operatorname{dom} f = \mathbb{R}^N$, but it will sometimes be useful to be explicit about the domain. If we let $\mathcal{D} = \operatorname{dom} f$, then the convex conjugate of f is

$$f^*(\boldsymbol{\nu}) = \sup_{\boldsymbol{x} \in \mathcal{D}} \ \langle \boldsymbol{x}, \boldsymbol{\nu} \rangle - f(\boldsymbol{x}).$$

There are multiple ways to think about the convex conjugate. Perhaps the most natural is simply that $f^*(\boldsymbol{\nu})$ is simply the maximum amount that the linear functional $\langle \boldsymbol{x}, \boldsymbol{\nu} \rangle$ exceeds $f(\boldsymbol{x})$.

Let's consider a particular example in one dimension ($N = 1$). Suppose

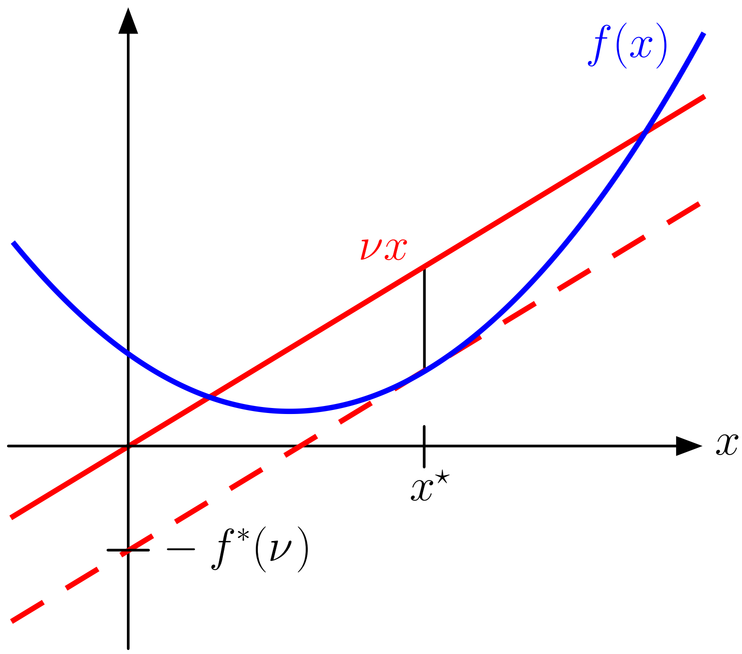
$$f(x) = x^2 - 2x + 2 = (x-1)^2 + 1.$$

We have

$$f^*(\nu) = \sup_{x \in \mathbb{R}} \left(\nu x - x^2 + 2x - 2 \right) = \frac{\nu^2}{4} + \nu - 1.$$

(You can verify the second equality by taking the derivative with respect to x, setting this to zero, and solving for x. This results in $x = \frac{\nu}{2} - 1$, and plugging this in yields the result above.)

Here is an example of what is happening:



Here we illustrate a function $f(x)$ overlayed with νx for an example value of ν. The *vertical* difference between the two functions is maximized at a particular value $x^\star$, and this distance is $f^*(\nu)$.

Another way to think about $f^*(\nu)$ is that it tells us how far we need to shift νx down so that it will be tangent to $f(x)$ (or a subgradient of f at x if f is not differentiable), as illustrated by the dashed line. In the case where f is differentiable, this is easy to see: since f is convex, $-f$ is concave and $\nu x - f(x)$ will be maximized when $\nu - f'(x) = 0$.

Properties

- $f^*(\boldsymbol{\nu})$ is convex (even when $f(\boldsymbol{x})$ is not).
- Fenchel's inequality: For any $\boldsymbol{x}$ and $\boldsymbol{\nu}$ we have

$$f(\boldsymbol{x}) + f^*(\boldsymbol{\nu}) \geq \langle \boldsymbol{x}, \boldsymbol{\nu} \rangle.$$

- For any function $f(\boldsymbol{x})$, we can define the conjugate of $f^*(\boldsymbol{\nu})$ as

$$f^{**}(\boldsymbol{x}) = \sup_{\boldsymbol{\nu} \in \mathcal{D}^*} \langle \boldsymbol{\nu}, \boldsymbol{x} \rangle - f^*(\boldsymbol{\nu}),$$

  where $\mathcal{D}^*$ is the domain of f^*. For an arbitrary $f(\boldsymbol{x})$ we have

$$f^{**}(\boldsymbol{x}) \leq f(\boldsymbol{x})$$

- If $f(\boldsymbol{x})$ is convex and has a closed epigraph, then taking the conjugate of $f^*(\boldsymbol{\nu})$ recovers $f(\boldsymbol{x})$:

$$f^{**}(\boldsymbol{x}) = f(\boldsymbol{x}).$$

- 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 more properties, see [BV04, Chapter 3.3].

Examples

The Indicator Function

We define the **indicator function** or **characteristic function** for a convex set $\mathcal{C}$ is given by

$$I_{\mathcal{C}}(\boldsymbol{x}) = \begin{cases} 0, & \text{if } x \in \mathcal{C}, \\ +\infty, & \text{if } x \notin \mathcal{C}. \end{cases}$$

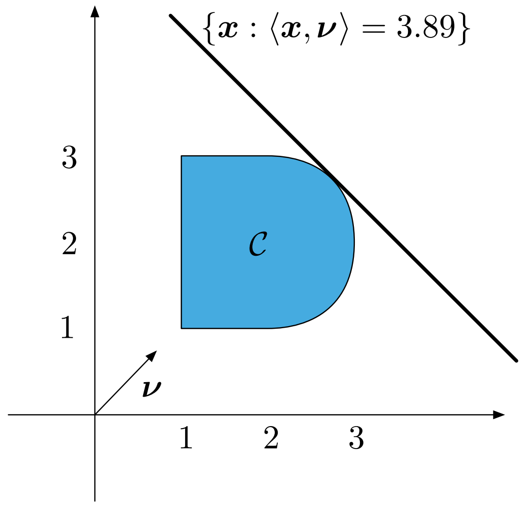
The convex conjugate of the indicator function is

$$\begin{aligned} h_{\mathcal{C}}(\boldsymbol{\nu}) = I_{\mathcal{C}}^*(\boldsymbol{\nu}) &= \sup_{\boldsymbol{x} \in \mathbb{R}^N} \ \langle \boldsymbol{x}, \boldsymbol{\nu} \rangle - I_{\mathcal{C}}(\boldsymbol{x}) \\ &= \sup_{\boldsymbol{x} \in \mathcal{C}} \ \langle \boldsymbol{x}, \boldsymbol{\nu} \rangle. \end{aligned}$$

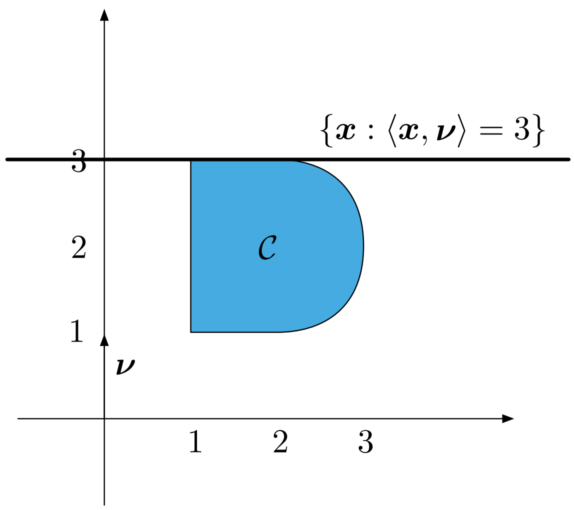
The function $h_{\mathcal{C}}(\boldsymbol{\nu})$ is also called the **support function** of $\mathcal{C}$. The support function defines a vector $\boldsymbol{\nu} \in \mathbb{R}^N$ that defines a linear function on $\mathcal{C}$ and then returns the maximum value of that linear function over $\mathcal{C}$.

Geometrically, this corresponds to taking the half-space $\{\boldsymbol{x} : \langle \boldsymbol{x}, \boldsymbol{\nu} \rangle \leq b\}$ and then determines how large b needs to be to ensure that the half-space contains all of $\mathcal{C}$ (since by definition we will have $\langle \boldsymbol{x}, \boldsymbol{\nu} \rangle \leq h_{\mathcal{C}}(\boldsymbol{\nu})$ for all $\boldsymbol{x} \in \mathcal{C}$).

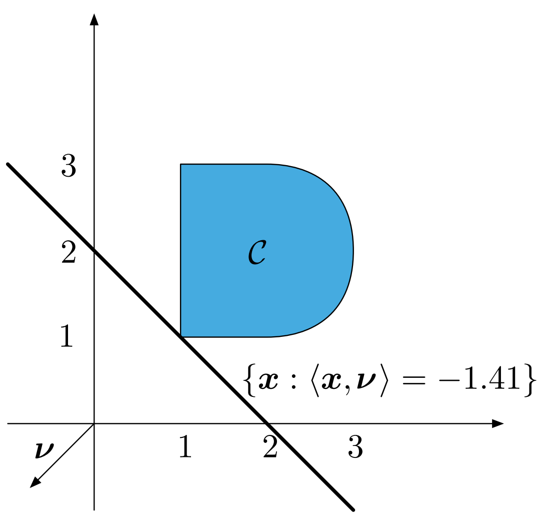
This is illustrated with an example on the following page.



$$\boldsymbol{\nu} = \frac{1}{\sqrt{2}}\begin{bmatrix}1\\1\end{bmatrix}$$
$$h_{\mathcal{C}}(\boldsymbol{\nu}) = 3.89$$



$$\boldsymbol{\nu} = \begin{bmatrix}0\\1\end{bmatrix}$$
$$h_{\mathcal{C}}(\boldsymbol{\nu}) = 3$$



$$\boldsymbol{\nu} = -\frac{1}{\sqrt{2}}\begin{bmatrix}1\\1\end{bmatrix}$$
$$h_{\mathcal{C}}(\boldsymbol{\nu}) = -1.41$$

Norms

As we have already seen on many occasions, norms frequently arise in optimization problems as either objective functions or constraints, and so their convex conjugates are important to be able to compute and work with. Before discussing the general case, let's look at what happens for $f(\boldsymbol{x}) = \|\boldsymbol{x}\|_2$ to get an idea of what we will need to do. In this case we have

$$\begin{aligned} f^*(\boldsymbol{\nu}) &= \sup_{\boldsymbol{x}\in\mathbb{R}^N} \; \langle \boldsymbol{x}, \boldsymbol{\nu} \rangle - \|\boldsymbol{x}\|_2 \\ &= \sup_{\boldsymbol{x}\in\mathbb{R}^N} \; \|\boldsymbol{x}\|_2 \|\boldsymbol{\nu}\|_2 - \|\boldsymbol{x}\|_2 \\ &= \sup_{\boldsymbol{x}\in\mathbb{R}^N} \; \|\boldsymbol{x}\|_2 (\|\boldsymbol{\nu}\|_2 - 1), \end{aligned}$$

where the second inequality above follows from Cauchy-Schwarz (which can be made to hold with equality). There are two cases to consider here. If $\|\boldsymbol{\nu}\|_2 \leq 1$, then we are trying to maximize a non-positive quantity, which we can do simply by setting $\boldsymbol{x} = \boldsymbol{0}$, resulting in $f^*(\boldsymbol{\nu}) = 0$. However, if $\|\boldsymbol{\nu}\|_2 > 1$ then $\|\boldsymbol{x}\|_2(\|\boldsymbol{\nu}\|_2 - 1)$ can be made arbitrarily large. Thus

$$f^*(\boldsymbol{\nu}) = \begin{cases} 0, & \text{if } \|\boldsymbol{\nu}\|_2 \leq 1, \\ +\infty, & \text{if } \|\boldsymbol{\nu}\|_2 > 1. \end{cases}$$

In other words, $f^*(\boldsymbol{\nu})$ is the indicator function for the unit ball corresponding to $\|\cdot\|_2$.

Now suppose that $f(\boldsymbol{x}) = \|\boldsymbol{x}\|$ where $\|\cdot\|$ denotes an arbitrary norm. To extend the argument above, we will need to introduce a new quantity that may not seem intuitive at first but is quite natural once you see how it is used.

Specifically, for any norm $\|\cdot\|$, we can consider the **dual norm**, which is defined as

$$\|\boldsymbol{\nu}\|_* = \sup_{\boldsymbol{x}:\|\boldsymbol{x}\|\le 1} \langle \boldsymbol{x}, \boldsymbol{\nu}\rangle.$$

Note that this is the support function (the convex conjugate of the indicator function) of the unit ball corresponding to $\|\cdot\|$. Moreover, as a direct consequence of the definition of the dual norm we have

$$\langle \boldsymbol{x}, \boldsymbol{\nu}\rangle \le \|\boldsymbol{x}\|\|\boldsymbol{\nu}\|_*.$$

This is exactly what we need to extend our previous argument.

We again begin with setting $f(\boldsymbol{x}) = \|\boldsymbol{x}\|$ and wish to compute

$$f^*(\boldsymbol{\nu}) = \sup_{\boldsymbol{x}\in\mathbb{R}^N} \langle \boldsymbol{x}, \boldsymbol{\nu}\rangle - \|\boldsymbol{x}\|$$

As before, we can consider two cases. If $\|\boldsymbol{\nu}\|_* \le 1$ then

$$\langle \boldsymbol{x}, \boldsymbol{\nu}\rangle - \|\boldsymbol{x}\| \le \|\boldsymbol{x}\|\|\boldsymbol{\nu}\|_* - \|\boldsymbol{x}\| \le 0.$$

In this case the largest we can make $f^*(\boldsymbol{\nu})$ is zero (by setting $\boldsymbol{x} = 0$). On the other hand, if $\|\boldsymbol{\nu}\|_* > 1$ then there must exist an $\boldsymbol{x}$ such that $\langle \boldsymbol{x}, \boldsymbol{\nu}\rangle \ge \|\boldsymbol{x}\|$. If we replace this $\boldsymbol{x}$ by a rescaled version $t\boldsymbol{x}$, then we can make

$$\langle t\boldsymbol{x}, \boldsymbol{\nu}\rangle - \|t\boldsymbol{x}\| = t(\langle \boldsymbol{x}, \boldsymbol{\nu}\rangle - \|\boldsymbol{x}\|)$$

arbitrarily large. Thus we have

$$f^*(\boldsymbol{\nu}) = \begin{cases} 0, & \text{if } \|\boldsymbol{\nu}\|_* \le 1, \\ +\infty, & \text{if } \|\boldsymbol{\nu}\|_* > 1, \end{cases}$$

i.e., $f^*(\boldsymbol{\nu})$ is the indicator function for the unit ball corresponding to the dual norm $\|\cdot\|_*$.

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

References

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