Duality in convex analysis and optimization

We have started to throw around the word "dual" a lot without really explaining why. Gaining an appreciation for "duality" will require us to take a moment to reflect on the unifying ideas that tie together some of the things we have learned about convex sets, convex functions, and convex optimization.

First, about the word *dual*. This is a broad concept in mathematical analysis, but in the context of optimization it amounts to the following: there are two ways to think about a vector $\boldsymbol{w} \in \mathbb{R}^N$:

Primal : $\boldsymbol{w}$ corresponds to a point in $\mathbb{R}^N$; its entries are coordinates for its location.

Dual : $\boldsymbol{w}$ corresponds to a **linear function** on $\mathbb{R}^N$: $f_w(\boldsymbol{x}) = \langle \boldsymbol{x}, \boldsymbol{w} \rangle = \sum_{n=1}^N w_n x_n$. Every function that we can write as an inner product against a fixed vector is linear, and every linear function on $\mathbb{R}^N$ can be written as such an inner product.

This duality, that vectors are both points and linear functions, is one of the foundational concepts of the field of functional analysis, but plays an especially important (and illuminating) role in convex analysis and optimization. Linear functions play this role by defining **hyperplanes**

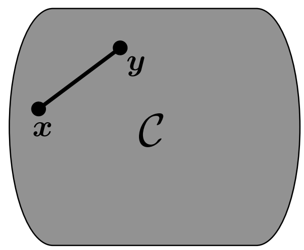
$$\mathcal{H}_{\omega,t} = \{\boldsymbol{x} \ : \ \langle \boldsymbol{x}, \boldsymbol{w} \rangle = t\},$$

and their corresponding **halfspaces** $\{\boldsymbol{x} \ : \ \langle \boldsymbol{x}, \boldsymbol{w} \rangle \le t\}$.

These notes will formalize these two points of view more explicitly for the three types of mathematical objects we have spent our time studying in this course: convex sets, convex functions, and convex optimization problems.

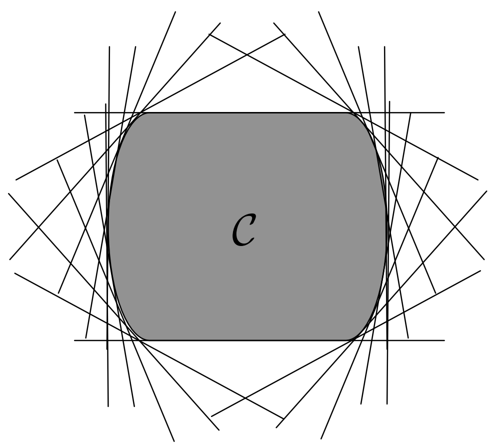
Convex Sets. There are two ways to describe a closed convex set $\mathcal{C} \subset \mathbb{R}^N$:

Primal : $\mathcal{C}$ is closed and obeys $\theta \boldsymbol{x} + (1-\theta)\boldsymbol{y} \in \mathcal{C}$ for all $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{C}$ and $\theta \in [0,1]$.



This way of thinking about $\mathcal{C}$ conceptualizes it as a collection of points.

Dual : $\mathcal{C}$ is an intersection of halfspaces.

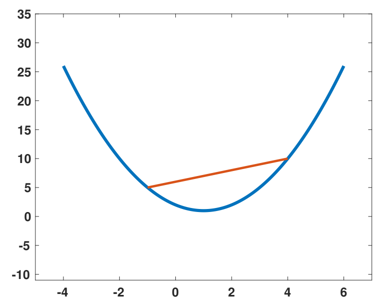


This way of thinking about $\mathcal{C}$ conceptualizes it via a collection of *functions*.

The concept, discussed much more below, that ties these two descriptions together is the **support function**. The "primal" way of describing a convex set $\mathcal{C}$ involves

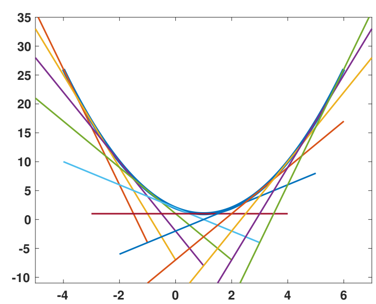
Convex functions. There are two ways to describe a closed[1] convex function $f : \mathbb{R}^N \to \mathbb{R}$:

Primal : f obeys $f(\theta\boldsymbol{x} + (1-\theta)\boldsymbol{y}) \leq \theta f(\boldsymbol{x}) + (1-\theta)f(\boldsymbol{y})$ for all $\boldsymbol{x}, \boldsymbol{y}$ and $\theta \in [0,1]$.



This way of thinking about f conceptualizes it as a collection of pairs $(\boldsymbol{x}, f(\boldsymbol{x}))$.

Dual : f is a pointwise supremum of affine functions.



This way of thinking about f conceptualizes it via a collection of (affine) functions.

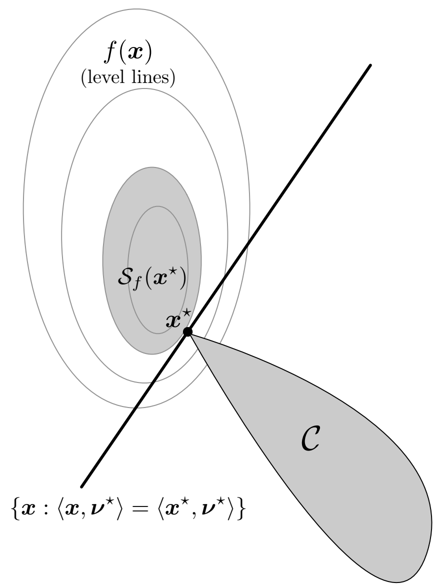
The concept, discussed much more below, that ties these two descriptions together is the **Fenchel conjugate**.

[1]We call of function *closed* if its sublevel sets are all closed. This is the same as the epigraph being a closed subset of $\mathbb{R}^{N+1}$.

Convex optimization.. There are two ways to describe the processes of minimizing convex $f(\boldsymbol{x})$ over a closed convex set $\mathcal{C}$:

Primal : We are looking for a $\boldsymbol{x}^\star \in \mathcal{C}$ such that $f(\boldsymbol{x}^\star) \leq f(\boldsymbol{y})$ for all $\boldsymbol{y} \in \mathcal{C}$.

Dual : We are looking for a hyperplane that separates $\mathcal{C}$ from a sublevel set of f while supporting both.



The concept, discussed much more below, that ties these two descriptions together is the **Fenchel dual**.

We will go through each of three concepts of duality in turn in these notes.

Duality for convex sets

In our discussion of constrained optimization for non-smooth functions, we introduced the notion of a **supporting hyperplane**. Specifically, if $\boldsymbol{a} \neq \mathbf{0}$ satisfies $\langle \boldsymbol{x}, \boldsymbol{a} \rangle \leq \langle \boldsymbol{x}_0, \boldsymbol{a} \rangle$ for all $\boldsymbol{x} \in \mathcal{C}$, then

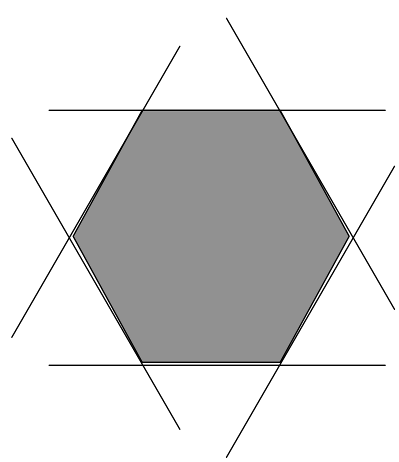
$$\mathcal{H} = \{\boldsymbol{x} : \langle \boldsymbol{x}, \boldsymbol{a} \rangle = \langle \boldsymbol{x}_0, \boldsymbol{a} \rangle\}$$

is a supporting hyperplane to $\mathcal{C}$ at $\boldsymbol{x}_0$. We introduced this in order to talk about *normal cones* and to provide a general set of optimality conditions for non-smooth optimization, but we will see here that the idea of supporting hyperplanes has a much deeper connection to convexity than you might initially expect.

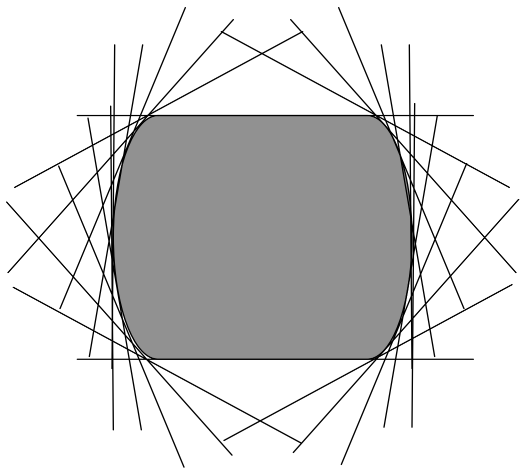
Specifically, let's start by making explicit something that is almost obvious given the definition of supporting hyperplanes:

A closed convex set is equal to the intersection of (halfspaces defined by) its supporting hyperplanes.

For example, this polygon is the intersection of a finite number of (six) halfspaces:



A general closed convex set is the intersection of a possibly infinite number of halfspaces:



That this is true can almost immediately be deduced from work we have done already. Let $\mathcal{C}_H$ be the intersection of the halfspaces defined by supporting hyperplanes at all points on the boundary of a closed convex set $\mathcal{C}$:

$$\mathcal{C}_H = \bigcap_{\boldsymbol{z}\in \operatorname{bd}\mathcal{C}} \ \bigcap_{(\boldsymbol{a},b)\in\mathcal{A}(\boldsymbol{z})} \{\boldsymbol{x} \ : \ \boldsymbol{a}^{\mathrm{T}}\boldsymbol{x} \leq b\}$$

$$\text{where } \ \mathcal{A}(\boldsymbol{z}) = \{\text{set of supporting hyperplanes at } \boldsymbol{z}\}$$
$$= \{(\boldsymbol{a}, b) \ : \ \{\boldsymbol{a}^{\mathrm{T}}\boldsymbol{x} = b\} \text{ is a supporting hyperplane at } \boldsymbol{z}\}.$$

If $\boldsymbol{x} \in \mathcal{C}$, then $\boldsymbol{x}$ is also in every halfspace in the intersection above, simply by the definition of "supporting hyperplane", thus $\boldsymbol{x} \in \mathcal{C} \Rightarrow \boldsymbol{x} \in \mathcal{C}_H$. If $\boldsymbol{x} \notin \mathcal{C}$, then there is at least one halfspace which does not contain $\boldsymbol{x}$; in particular, we can choose the hyperplane that supports $\mathcal{C}$ at $P_{\mathcal{C}}(\boldsymbol{x})$ with normal vector $\boldsymbol{x} - P_{\mathcal{C}}(\boldsymbol{x})$. Thus $\boldsymbol{x} \notin \mathcal{C} \Rightarrow \boldsymbol{x} \notin \mathcal{C}_H$.

Moral: *A (closed) convex set can be thought of either as a collection of points or via a collection of hyperplanes.*

What we would like now is a clean, mathematical way to specify the set of supporting hyperplanes for a convex set $\mathcal{C}$. This is done through the **support function** of $\mathcal{C}$

$$h_{\mathcal{C}}(\boldsymbol{\nu}) = \sup_{\boldsymbol{x} \in \mathcal{C}} \langle \boldsymbol{x}, \boldsymbol{\nu} \rangle. \tag{1}$$

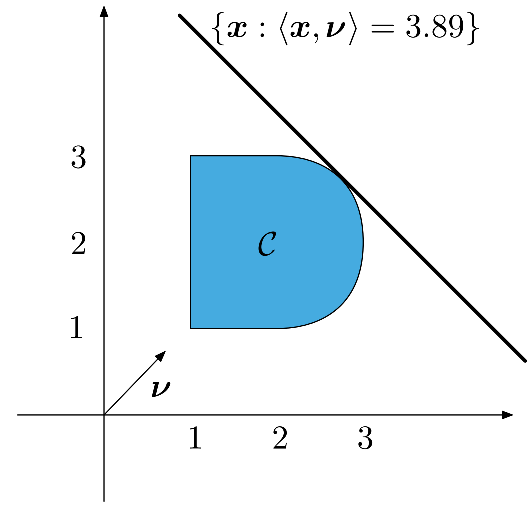
The support function takes linear functionals (which again are just vectors in $\mathbb{R}^N$) as an argument and returns the maximum[2] value of that linear functional over $\mathcal{C}$. Geometrically, we take take the halfspace $\{\boldsymbol{x} \ : \ \langle \boldsymbol{x}, \boldsymbol{\nu} \rangle \le b\}$ and see how large we need to make b so that it contains $\mathcal{C}$ (of course, b might be negative). If $\mathcal{C}$ is closed and $h_{\mathcal{C}}(\boldsymbol{\nu}) < +\infty$, then

$$\mathcal{H}_{\nu} = \{\boldsymbol{x} : \langle \boldsymbol{x}, \boldsymbol{\nu} \rangle = h_{\mathcal{C}}(\boldsymbol{\nu})\}$$

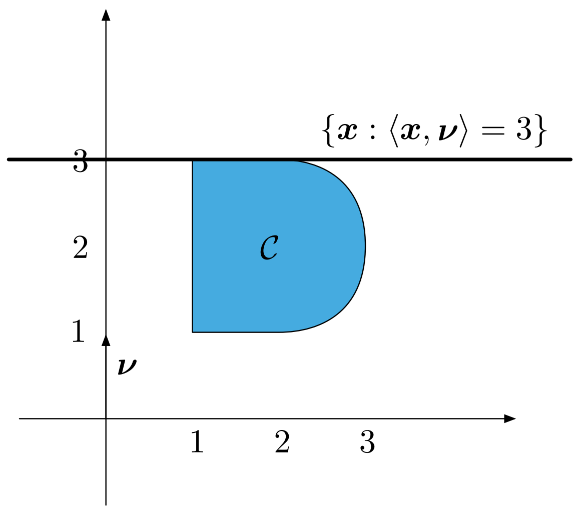
is a supporting hyperplane at all $\boldsymbol{x}^\star$ that maximize (1).

Here are some examples that illustrate what the support function is capturing:

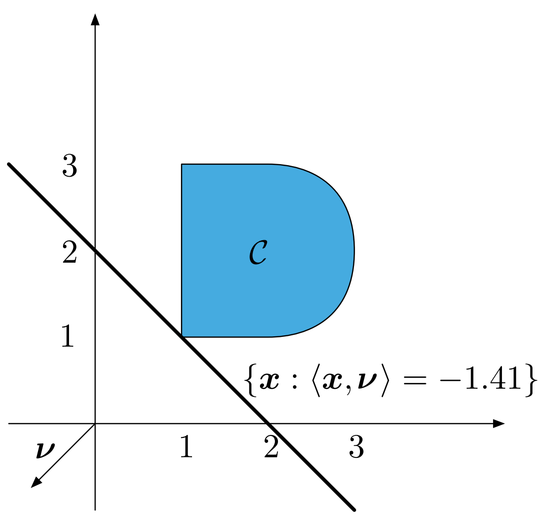
[2]We are writing sup instead of max here since if $\mathcal{C}$ is not bounded, it very well may be that $h_{\mathcal{C}}(\boldsymbol{\nu}) = \infty$, and if $\mathcal{C}$ is not closed, there very well may be that there is no $\boldsymbol{x} \in \mathcal{C}$ that maximizes $\langle \boldsymbol{x}, \boldsymbol{\nu} \rangle$. If $\mathcal{C}$ is closed and bounded (compact), then we can safely replace sup with max.



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



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



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

A closed convex shape is completely characterized by its support function. Given $h_{\mathcal{C}}(\boldsymbol{\nu})$, we can recover $\mathcal{C}$ by simply intersecting all the halfspaces defined by $h_{\mathcal{C}}(\boldsymbol{\nu})$,

$$\mathcal{C} = \bigcap_{\boldsymbol{\nu}\in\mathcal{C}^*} \{\boldsymbol{x} \ : \ \langle \boldsymbol{x}, \boldsymbol{\nu}\rangle \leq h_{\mathcal{C}}(\boldsymbol{\nu})\},$$

where $\mathcal{C}^*$ is the domain of $h_{\mathcal{C}}(\boldsymbol{\nu})$, the set of $\boldsymbol{\nu}$ such that $h_{\mathcal{C}}(\boldsymbol{\nu}) < \infty$.

Thus we can think of the support function $h_{\mathcal{C}}(\boldsymbol{\nu})$ as being the dual representation of a closed convex set $\mathcal{C}$.

We close this section with three examples, just to get a little more feel for how the support function characterizes shapes.

Example: The unit simplex. For

$$\Delta = \left\{\boldsymbol{x} \ : \ \boldsymbol{x} \geq \boldsymbol{0}, \ \sum_{n=1}^{N} x_n = 1\right\},$$

we have

$$h_{\Delta}(\boldsymbol{\nu}) = \max_{\boldsymbol{x}\in\Delta} \ \sum_{n=1}^{N} \nu_n x_n = \max(\nu_1, \ldots, \nu_N).$$

Example: Norm balls. Let $\mathcal{B}$ be the unit ball for an arbitrary norm $\|\cdot\|$,

$$\mathcal{B} = \{\boldsymbol{x} \ : \ \|\boldsymbol{x}\| \leq 1\}.$$

Then

$$h_{\mathcal{B}}(\boldsymbol{\nu}) = \max_{\|\boldsymbol{x}\|\leq 1} \langle \boldsymbol{x}, \boldsymbol{\nu}\rangle = \|\boldsymbol{\nu}\|_*,$$

where $\|\cdot\|_*$ is the corresponding dual norm.

Example: Affine sets. Let $\boldsymbol{A}$ be an $M \times N$ matrix with $\operatorname{rank}(\boldsymbol{A}) = M$, let $\boldsymbol{b}$ be a vector in $\mathbb{R}^M$ and let $\mathcal{C}$ be the set of all solutions to $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$,

$$\mathcal{C} = \{\boldsymbol{x} \ : \ \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}\}.$$

Recall that we can decompose any $\boldsymbol{\nu}$ into a component in $\operatorname{Null}(\boldsymbol{A})$ and a component in $\operatorname{Row}(\boldsymbol{A})$. If $\boldsymbol{\nu}$ has any component in $\operatorname{Null}(\boldsymbol{A})$ then we can make $\langle \boldsymbol{x}, \boldsymbol{\nu} \rangle$ arbitrarily large. Thus $\sup_{\boldsymbol{x}\in\mathcal{C}} \langle \boldsymbol{x}, \boldsymbol{\nu} \rangle = \infty$ unless $\boldsymbol{\nu} \in \operatorname{Row}(\boldsymbol{A})$. For $\boldsymbol{\nu} \in \operatorname{Row}(\boldsymbol{A})$, we have $\boldsymbol{\nu} = \boldsymbol{A}^{\mathrm{T}}(\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}})^{-1}\boldsymbol{A}\boldsymbol{\nu}$ and $h_{\mathcal{C}}(\boldsymbol{\nu})$ is

$$\sup_{\boldsymbol{x}\in\mathcal{C}} \langle \boldsymbol{x}, \boldsymbol{\nu} \rangle = \sup_{\boldsymbol{x}\in\mathcal{C}} \langle \boldsymbol{x}, \boldsymbol{A}^{\mathrm{T}}(\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}})^{-1}\boldsymbol{A}\boldsymbol{\nu} \rangle = \langle \boldsymbol{A}^{\mathrm{T}}(\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}})^{-1}\boldsymbol{b}, \boldsymbol{\nu} \rangle.$$

Thus

$$h_{\mathcal{C}}(\boldsymbol{\nu}) = \begin{cases} \langle \boldsymbol{A}^{\mathrm{T}}(\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}})^{-1}\boldsymbol{b}, \boldsymbol{\nu} \rangle & \boldsymbol{\nu} \in \operatorname{Row}(\boldsymbol{A}), \\ \infty & \boldsymbol{\nu} \notin \operatorname{Row}(\boldsymbol{A}). \end{cases}$$

The domain of the support function is the row space, $\mathcal{C}^* = \operatorname{Row}(\boldsymbol{A})$.

Duality for convex functions

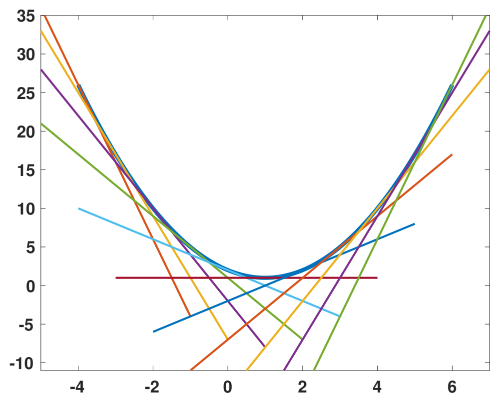
When dealing with convex *functions*, the dual perspective is to think of them as the envelope of a set of linear functionals To make this concrete, let's start by showing how we can write one particular convex function

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

as a (point-wise) maximum of linear functionals

$$f(x) = \max_{(\nu,b)\in\mathcal{H}} (\nu x - b)\,, \tag{2}$$

where $\mathcal{H}$ is a set of (slope,intercept) pairs that define all the possible lines to optimize over, i.e., the set of lines sketched in this figure:



We would like to have a mathematical description of this set of lines (the set of tangent lines to f). Of course, these are given by its derivative $f'(x)$; since f is convex, we know that

$$f(y) \geq f(x) + f'(x)(y-x),$$

with equality at $y = x$. So the range for ν in the optimization in (2) is

$$\{\nu \in \mathbb{R} \ : \ f'(x) = \nu \ \text{ for some } x\}.$$

In this case, ν will range over the entire real line, as for every ν, there is an x_ν such that $f'(x_\nu) = \nu$. More precisely, since $f'(x) = 2x - 2$, we have

$$f'(x_\nu) = \nu \quad \text{when} \quad x_\nu = \frac{\nu}{2} + 1.$$

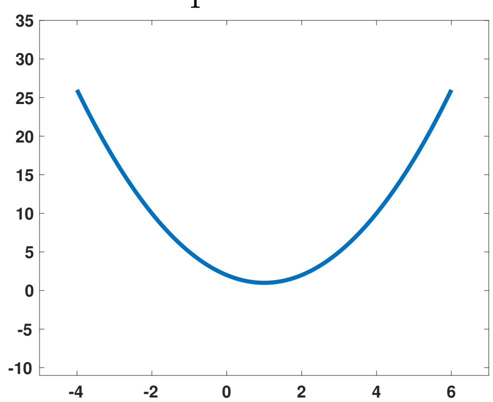
So we have a tangent (affine underestimator) of $f(x)$ at x_ν given by $f(x_\nu) + \nu(x - x_\nu)$. Rearranging this into the form $\nu x - b$ as in (2) we see that the intercept term b is exactly $\nu x_\nu - f(x_\nu)$. This quantity represents the optimal intercept that yields the tightest possible affine underestimator for a given slope, and we will denote it by

$$\begin{aligned} f^*(\nu) &= \nu \cdot x_\nu - f(x_\nu) \\ &= \frac{\nu^2}{4} + \nu - 1. \end{aligned}$$

Putting this all together, we can rewrite this $f(x) = (x-1)^2 + 1$ as a maximum of linear functions as

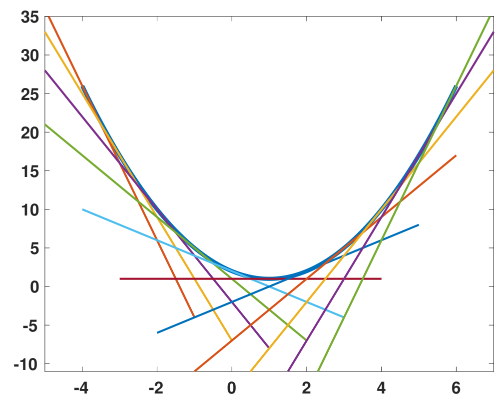
$$f(x) = \max_{\nu \in \mathbb{R}} (\nu x - f^*(\nu)) = \max_{\nu \in \mathbb{R}} \left(\nu x - \nu^2/4 + \nu - 1 \right).$$

primal



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

dual



$f(x) = \max_\nu (\nu x - \nu^2/4 - \nu + 1)$

Notice that because the tangency point x_ν satisfies $f'(x_\nu) = \nu$, it is exactly the point that maximizes the concave function $x\nu - f(x)$ (as the gradient of this function is 0 at $x = x_\nu$). Thus we can write

$$x_\nu = \arg\max_{x\in\mathbb{R}} (x\nu - f(x)).$$

Since $f*(\nu) = \nu x_\nu - f(x_\nu)$, this implies that

$$f^*(\nu) = \max_{x\in\mathbb{R}} (x\nu - f(x)).$$

Contrast this with

$$f(x) = \max_{\nu\in\mathbb{R}} (\nu x - f^*(\nu))$$

The symmetry of the "max" expressions for $f(x)$ and $f^*(\nu)$ is not a coincidence — $f^*(\nu)$ is called the *Fenchel conjugate* of $f(x)$, and is the key quantity in the dual representation of convex functions (that is, representing a convex function as a maximum from a family of affine functions). We will develop this idea in full in the next section.

The Fenchel conjugate

We will start off with our main result. If $f(\boldsymbol{x})$ is a closed[3] convex function on domain $\mathcal{D}$, then we can write it as a supremum of affine functions as

$$f(\boldsymbol{x}) = \sup_{\boldsymbol{\nu}\in\mathcal{D}^*} (\langle \boldsymbol{x}, \boldsymbol{\nu}\rangle - f^*(\boldsymbol{\nu})), \tag{3}$$

where $f^*(\boldsymbol{\nu})$ is the **Fenchel conjugate**

$$f^*(\boldsymbol{\nu}) = \sup_{\boldsymbol{x}\in\mathcal{D}} (\langle \boldsymbol{x}, \boldsymbol{\nu}\rangle - f(\boldsymbol{x})), \tag{4}$$

[3] We call of function *closed* if its sublevel sets $\{\boldsymbol{x} \ : \ f(\boldsymbol{x}) \leq \beta\}$ are closed sets for all $\beta \in \mathbb{R}$.

and $\mathcal{D}^*$ is the domain of f^*, that is, the set of $\boldsymbol{\nu}$ such that the supremum above is $< \infty$.

Roughly speaking, the development in the general case tracks the development we gave for the simple quadratic function in the last section: given a normal vector $\boldsymbol{\nu}$ for a linear function, we find $\boldsymbol{x}_\nu$ such that $\nabla f(\boldsymbol{x}_\nu) = \boldsymbol{\nu}$ by solving

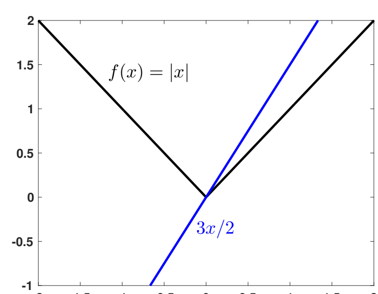
$$\boldsymbol{x}_\nu = \arg\max_{\boldsymbol{x}} \left(\langle \boldsymbol{x}, \boldsymbol{\nu} \rangle - f(\boldsymbol{x}) \right),$$

then calculate the intercept of the affine underestimator $\langle \boldsymbol{x} - \boldsymbol{x}_\nu, \boldsymbol{\nu} \rangle + f(\boldsymbol{x}_\nu)$ as

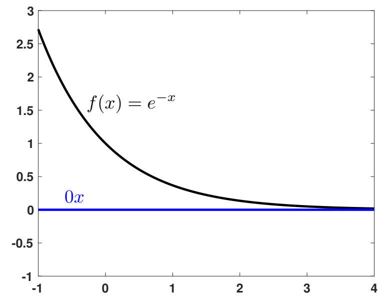
$$f^*(\boldsymbol{\nu}) = \max_{\boldsymbol{x}} \left(\langle \boldsymbol{x}, \boldsymbol{\nu} \rangle - f(\boldsymbol{x}) \right) = \langle \boldsymbol{x}_\nu, \boldsymbol{\nu} \rangle - f(\boldsymbol{x}_\nu).$$

But there are two things to look out for, which are the two reasons we are using sup in (3) and (4) instead of max. First, it might be that $\langle \boldsymbol{x}, \boldsymbol{\nu} \rangle - f(\boldsymbol{x})$ is unbounded above as a function of $\boldsymbol{x}$, so we have $f^*(\boldsymbol{\nu}) = \infty$. One example is $f(x) = |x|$, where

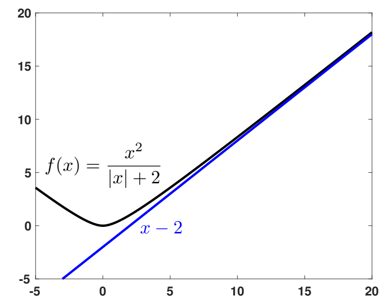
$$f^*(\nu) = \sup_{x \in \mathbb{R}} \nu x - |x| = \infty \quad \text{when} \quad |\nu| > 1.$$



It also might be the case that we can form a "tight" underestimate from $\langle \boldsymbol{x}, \boldsymbol{\nu} \rangle - f^*(\boldsymbol{\nu})$ without there actually existing a point $\boldsymbol{x}_\nu$ with $\nabla f(\boldsymbol{x}_\nu) = \boldsymbol{\nu}$. Two examples are $f(x) = e^{-x}$ with $f^*(0) = 0$



and $f(x) = \frac{x^2}{|x|+2}$ with $f^*(1) = 2$



We will start by looking at a few more concrete examples just to get comfortable with the calculations involved, then we will talk about some general properties of the Fenchel conjugate. In the next lecture, we will see that the Fenchel conjugate also plays a central role (as expected) when looking at the dual interpretation of convex optimization problems.

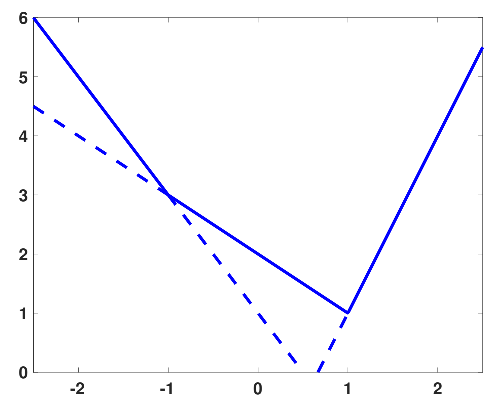
Moral: *A (closed) convex function can be thought of as a supremum of affine functions.*

Fenchel conjugate examples

Below are a few key examples for computing the Fenchel conjugate. For lots of other examples, see [BV04, Chapter 3.3] and [Bec17, Chapter 4.4].

Piecewise linear.

$$f(x) = \begin{cases} -2x+1 & x \le -1 \\ -x+2 & -1 \le x \le 1, \\ 3x-2 & x \ge 1. \end{cases}$$



Of course we can already write f as a max of three affine functions,

$$f(x) = \max(-2x+1, -x+2, 3x-2),$$

but it is instructive to compute the Fenchel conjugate anyway.

First, it should be clear that if $\nu < -2$ or $\nu > 3$, then $f^*(\nu) = \infty$, so the domain of f^* is $\mathcal{D}^* = [-2, 3]$.

For $-2 \le \nu \le -1$ we will have

$$x_\nu = \arg\max_x (\nu x - f(x)) = -1,$$

and so $f^*(\nu) = \nu x_\nu - f(x_\nu) = -\nu - 3$. For $-1 \le \nu \le 3$, $x_\nu = 1$, $f(x_\nu) = 1$ and so $f^*(\nu) = \nu - 1$. Thus

$$f^*(\nu) = \begin{cases} -\nu - 3 & -2 \le \nu \le -1, \\ \nu - 1 & -1 \le \nu \le 3, \\ \infty & \text{otherwise.} \end{cases}$$

Note that now when we write

$$f(x) = \max_{\nu \in [-2,3]} (\nu x - f^*(x))$$

we are taking a max over an infinity of lines with slopes on the continuum between -2 and 3; we are including all the affine under-estimators (subgradients) at $x = -1$ and $x = 1$.

Quadratic functions.

For symmetric positive definite $\boldsymbol{Q}$,

$$f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{x} \qquad \Leftrightarrow \qquad f^*(\boldsymbol{\nu}) = \frac{1}{2}\boldsymbol{\nu}^{\mathrm{T}}\boldsymbol{Q}^{-1}\boldsymbol{\nu}.$$

To show this, note that by taking a gradient and setting it equal to zero, the supremum

$$\sup_{\boldsymbol{x}} \left(\langle \boldsymbol{x}, \boldsymbol{\nu} \rangle - \frac{1}{2}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{x} \right), \quad \text{is achieved when} \quad \boldsymbol{\nu} - \boldsymbol{Q}\boldsymbol{x}_\nu = \boldsymbol{0}.$$

Since $\boldsymbol{Q}$ is invertible, we plug in $\boldsymbol{x}_\nu = \boldsymbol{Q}^{-1}\boldsymbol{\nu}$ to get

$$\langle \boldsymbol{x}_\nu, \boldsymbol{\nu} \rangle - \frac{1}{2}\boldsymbol{x}_\nu^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{x}_\nu = \boldsymbol{\nu}^{\mathrm{T}}\boldsymbol{Q}^{-1}\boldsymbol{\nu} - \frac{1}{2}\boldsymbol{\nu}^{\mathrm{T}}\boldsymbol{Q}^{-1}\boldsymbol{Q}\boldsymbol{Q}^{-1}\boldsymbol{\nu} = \frac{1}{2}\boldsymbol{\nu}^{\mathrm{T}}\boldsymbol{Q}^{-1}\boldsymbol{\nu}.$$

Monomials.

For $p > 1$,

$$f(x) = \frac{1}{p}|x|^p \qquad \Leftrightarrow \qquad f^*(\nu) = \frac{1}{q}|\nu|^q,$$

where $1/p + 1/q = 1$.

To show this, note that by setting the derivative of $\nu x - \frac{1}{p}|x|^p$ to zero, one can show that this quantity is maximized at the point $x^\star = \operatorname{sgn}(\nu)|\nu|^{1/p-1}$. From this one can derive that

$$f^*(\nu) = (1 - 1/p)|\nu|^{p/p-1} = \frac{1}{q}|\nu|^q$$

for $q = p/(p-1)$.

The ℓ_2 norm.

For the Euclidean norm on $\mathbb{R}^N$, i.e., $f(\boldsymbol{x}) = \|\boldsymbol{x}\|_2$, 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 $\|\cdot\|_2$.

General norms.

For an arbitrary norm on $\mathbb{R}^N$,

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

where $\|\boldsymbol{\nu}\|_*$ is the dual norm that we have encountered previously,

$$\|\boldsymbol{\nu}\|_* = \max_{\|\boldsymbol{x}\|\leq 1} \langle \boldsymbol{x}, \boldsymbol{\nu}\rangle.$$

This follows exactly the same line of reasoning as in the ℓ_2 norm using the generalized Cauchy-Schwarz inequality,

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

and the fact that we know that there is always a $\boldsymbol{x}$ that achieves equality above.

General squared norms.

If we square an arbitrary norm, we actually get a much "nicer" conjugate,

$$f(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{x}\|^2 \qquad \Leftrightarrow \qquad f^*(\boldsymbol{\nu}) = \frac{1}{2}\|\boldsymbol{\nu}\|_*^2.$$

To see this, first, we note that $\boldsymbol{\nu}^{\mathrm{T}}\boldsymbol{x} - \frac{1}{2}\|\boldsymbol{x}\|^2 \leq \|\boldsymbol{\nu}\|_*\|\boldsymbol{x}\| - \frac{1}{2}\|\boldsymbol{x}\|^2$. This second function is a quadratic function of $\|\boldsymbol{x}\|$ which has a maximum value of $(1/2)\|\boldsymbol{\nu}\|_*^2$. It is easy to show that it achieves this value for $\boldsymbol{x}^\star$ with $\langle \boldsymbol{x}^\star, \boldsymbol{\nu}\rangle = \|\boldsymbol{\nu}\|_*\|\boldsymbol{x}^\star\|$ and $\|\boldsymbol{x}^\star\| = \|\boldsymbol{\nu}\|_*$.

Indicator functions

Recall the indicator function for a convex set $\mathcal{C}$ is given by

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

The Fenchel conjugate of the indicator function is

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

So $\iota_{\mathcal{C}}^*(\boldsymbol{\nu})$ is the support function of the $\mathcal{C}$ that we encountered in our discussion of duality and convex sets above.

Fenchel conjugate properties

For the following properties of the Fenchel conjugate, we will make only very basic assumptions on f. First we assume that f is "proper": this means that $f(\boldsymbol{x})$ never attains the value $-\infty$ and that there is at least one $\boldsymbol{x}$ such that $f(\boldsymbol{x}) < \infty$. It is not hard to see that only the most pathological functions do not meet these conditions. We will also assume that f is closed. Note that we do not assume that f is convex unless explicitly specified.

We discuss the proofs of these statements in the Technical Details section. For more properties and discussion can be found in [BV04, Chapter 3.3] and [Bec17, Chapter 4].

- $f^*(\boldsymbol{\nu})$ is convex (and closed) even when $f(\boldsymbol{x})$ is not convex.
- Fenchel's inequality: For any $\boldsymbol{x}$ and $\boldsymbol{\nu}$ we have
$$f(\boldsymbol{x}) + f^*(\boldsymbol{\nu}) \geq \langle \boldsymbol{x}, \boldsymbol{\nu} \rangle.$$
We have equality above when $\boldsymbol{\nu} \in \partial f(\boldsymbol{x})$.
- 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
$$f^{**}(\boldsymbol{x}) \leq f(\boldsymbol{x})$$
- As we discussed above, if $f(\boldsymbol{x})$ is convex and closed, then taking the conjugate of $f^*(\boldsymbol{\nu})$ recovers $f(\boldsymbol{x})$:
$$f^{**}(\boldsymbol{x}) = f(\boldsymbol{x}).$$

- If f is convex, then the subdifferential of f^* is the inverse of the subdifferential of f in that

$$\boldsymbol{\nu} \in \partial f(\boldsymbol{x}) \;\Leftrightarrow\; \boldsymbol{x} \in \partial f^*(\boldsymbol{\nu}).$$

- 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

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

References

[Bec17] A. Beck. *First-order methods in optimization.* SIAM, 2017.

[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.