Lagrange duality

Another way to arrive at the KKT conditions, and one which gives us some insight on solving constrained optimization problems, is through the *Lagrange dual*. The dual is a maximization program in $\boldsymbol{\lambda}, \boldsymbol{\nu}$ — it is always concave (even when the original program is not convex), and gives us a systematic way to lower bound the optimal value.

The Lagrangian

We consider an optimization program of the form

$$\begin{aligned} \underset{\boldsymbol{x} \in \mathbb{R}^N}{\text{minimize}} \;\; f_0(\boldsymbol{x}) \quad & f_m(\boldsymbol{x}) \leq 0, \quad m = 1, \ldots, M \\ & h_p(\boldsymbol{x}) = 0, \quad p = 1, \ldots, P. \end{aligned} \tag{1}$$

Much of what we will say below applies equally well to nonconvex programs as well as convex programs, so we will make it clear when we are taking the f_m to be convex and the h_p to be affine. We will take the domain of all of the f_m and h_p to be all of $\mathbb{R}^N$ below; this just simplifies the exposition, we can easily replace this with the intersections of the $\operatorname{dom} f_m$ and $\operatorname{dom} h_p$. We will assume that the feasible set

$$\mathcal{C} = \{\boldsymbol{x} \;:\; f_m(\boldsymbol{x}) \leq 0,\; h_p(\boldsymbol{x}) = 0,\; m = 1, \ldots, M, \;\; p = 1, \ldots, P\}$$

is non-empty and a subset $\mathbb{R}^N$.

The **Lagrangian** takes the constraints in the program above and integrates them into the objective function. The Lagrangian $L : \mathbb{R}^N \times \mathbb{R}^M \times \mathbb{R}^P \to \mathbb{R}$ associated with this optimization program is

$$L(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) = f_0(\boldsymbol{x}) + \sum_{m=1}^{M} \lambda_m f_m(\boldsymbol{x}) + \sum_{p=1}^{P} \nu_p h_p(\boldsymbol{x})$$

The $\boldsymbol{x}$ above are referred to as **primal variables**, and the $\boldsymbol{\lambda}, \boldsymbol{\nu}$ as either **dual variables** or **Lagrange multipliers**.

The **Lagrange dual function** $g(\boldsymbol{\lambda}, \boldsymbol{\nu}) : \mathbb{R}^M \times \mathbb{R}^P \to \mathbb{R}$ is the minimum of the Lagrangian over all values of $\boldsymbol{x}$:

$$g(\boldsymbol{\lambda}, \boldsymbol{\nu}) = \inf_{\boldsymbol{x} \in \mathbb{R}^N} \left(f_0(\boldsymbol{x}) + \sum_{m=1}^{M} \lambda_m f_m(\boldsymbol{x}) + \sum_{p=1}^{P} \nu_p h_p(\boldsymbol{x}) \right).$$

Since the dual is a pointwise infimum of a family of affine functions in $\boldsymbol{\lambda}, \boldsymbol{\nu}$, g **is concave** regardless of whether or not the f_m, h_p are convex.

The key fact about the dual function is that is it is everywhere a lower bound on the optimal value of the original program. If $p^\star$ is the optimal value for (1), then

$$g(\boldsymbol{\lambda}, \boldsymbol{\nu}) \leq p^\star, \quad \text{for all } \boldsymbol{\lambda} \geq \mathbf{0},\ \boldsymbol{\nu} \in \mathbb{R}^P.$$

This is (almost too) easy to see. For any feasible point $\boldsymbol{x}_0$,

$$\sum_{m=1}^{M} \lambda_m f_m(\boldsymbol{x}_0) + \sum_{p=1}^{P} \nu_p h_p(\boldsymbol{x}_0) \leq 0,$$

and so

$$L(\boldsymbol{x}_0, \boldsymbol{\lambda}, \boldsymbol{\nu}) \leq f_0(\boldsymbol{x}_0), \quad \text{for all } \boldsymbol{\lambda} \geq \mathbf{0},\ \boldsymbol{\nu} \in \mathbb{R}^P,$$

meaning

$$g(\boldsymbol{\lambda}, \boldsymbol{\nu}) = \inf_{\boldsymbol{x} \in \mathbb{R}^N} L(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) \leq L(\boldsymbol{x}_0, \boldsymbol{\lambda}, \boldsymbol{\nu}) \leq f_0(\boldsymbol{x}_0).$$

Since this holds for all feasible $\boldsymbol{x}_0$, $g(\boldsymbol{\lambda}, \boldsymbol{\nu}) \leq \inf_{\boldsymbol{x} \in \mathcal{C}} f_0(\boldsymbol{x}) = p^\star$.

The (Lagrange) dual to the optimization program (1) is

$$\underset{\boldsymbol{\lambda} \in \mathbb{R}^M, \boldsymbol{\nu} \in \mathbb{R}^P}{\text{maximize}} \ g(\boldsymbol{\lambda}, \boldsymbol{\nu}) \quad \text{subject to} \quad \boldsymbol{\lambda} \geq \mathbf{0}. \tag{2}$$

The dual optimal value $d^\star$ is

$$d^\star = \sup_{\boldsymbol{\lambda} \geq \mathbf{0}, \boldsymbol{\nu}} g(\boldsymbol{\lambda}, \boldsymbol{\nu}) = \sup_{\boldsymbol{\lambda} \geq \mathbf{0}, \boldsymbol{\nu}} \inf_{\boldsymbol{x} \in \mathbb{R}^N} L(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}).$$

Since $g(\boldsymbol{\lambda}, \boldsymbol{\nu}) \leq p^\star$, we know that

$$d^\star \leq p^\star.$$

The quantity $p^\star - d^\star$ is called the **duality gap**. If $p^\star = d^\star$, then we say that (1) and (2) exhibit **strong duality**.

Certificates of (sub)optimality

Any dual feasible[1] $(\boldsymbol{\lambda}, \boldsymbol{\nu})$ gives us a lower bound on $p^\star$, since $g(\boldsymbol{\lambda}, \boldsymbol{\nu}) \leq p^\star$. If we have a primal feasible $\boldsymbol{x}$, then we know that

$$f_0(\boldsymbol{x}) - p^\star \;\leq\; f_0(\boldsymbol{x}) - g(\boldsymbol{\lambda}, \boldsymbol{\nu}).$$

We will refer to $f_0(\boldsymbol{x}) - g(\boldsymbol{\lambda}, \boldsymbol{\nu})$ as the **duality gap** for the primal/dual (feasible) variables $\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}$. We know that

$$p^\star \in [g(\boldsymbol{\lambda}, \boldsymbol{\nu}), f_0(\boldsymbol{x})], \quad \text{and likewise} \quad d^\star \in [g(\boldsymbol{\lambda}, \boldsymbol{\nu}), f_0(\boldsymbol{x})].$$

If we are ever able to reduce this gap to zero, then we know that $\boldsymbol{x}$ is primal optimal, and $\boldsymbol{\lambda}, \boldsymbol{\nu}$ are dual optimal.

There are certain kinds of "primal-dual" algorithms that produce a series of (feasible) points $\boldsymbol{x}^{(k)}, \boldsymbol{\lambda}^{(k)}, \boldsymbol{\nu}^{(k)}$ at every iteration. We can then use

$$f_0(\boldsymbol{x}^{(k)}) - g(\boldsymbol{\lambda}^{(k)}, \boldsymbol{\nu}^{(k)}) \leq \epsilon,$$

as a stopping criteria, and know that our answer would yield an objective value no further than ϵ from optimal.

Strong duality and the KKT conditions

Suppose that for a convex program, the primal optimal value $p^\star$ and the dual optimal value $d^\star$ are equal

$$p^\star = d^\star.$$

[1] We simply need $\boldsymbol{\lambda} \geq \mathbf{0}$ for $(\boldsymbol{\lambda}, \boldsymbol{\nu})$ to be dual feasible.

If $\boldsymbol{x}^\star$ is a primal optimal point and $\boldsymbol{\lambda}^\star, \boldsymbol{\nu}^\star$ is a dual optimal point, then we must have

$$\begin{aligned} f_0(\boldsymbol{x}^\star) &= g(\boldsymbol{\lambda}^\star, \boldsymbol{\nu}^\star) \\ &= \inf_{\boldsymbol{x} \in \mathbb{R}^N} \left(f_0(\boldsymbol{x}) + \sum_{m=1}^M \lambda_m^\star f_m(\boldsymbol{x}) + \sum_{p=1}^P \nu_p^\star h_p(\boldsymbol{x}) \right) \\ &\leq f_0(\boldsymbol{x}^\star) + \sum_{m=1}^M \lambda_m^\star f_m(\boldsymbol{x}^\star) + \sum_{p=1}^P \nu_p^\star h_p(\boldsymbol{x}^\star) \\ &\leq f_0(\boldsymbol{x}^\star). \end{aligned}$$

The last inequality follows from the fact that $\lambda_m^\star \geq 0$ (dual feasibility), $f_m(\boldsymbol{x}^\star) \leq 0$, and $h_p(\boldsymbol{x}^\star) = 0$ (primal feasibility). Since we started out and ended up with the same thing, all of the things above must be equal, and so

$$\lambda_m^\star f_m(\boldsymbol{x}^\star) = 0, \quad m = 1, \ldots, M.$$

Also, since we know $\boldsymbol{x}^\star$ is a minimizer of $L(\boldsymbol{x}, \boldsymbol{\lambda}^\star, \boldsymbol{\nu}^\star)$ (second equality above), which is an unconstrained convex function (with $\boldsymbol{\lambda}, \boldsymbol{\nu}$ fixed), the gradient with respect to $\boldsymbol{x}$ must be zero:

$$\nabla_{\boldsymbol{x}} L(\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star, \boldsymbol{\nu}^\star) = \nabla f_0(\boldsymbol{x}^\star) + \sum_{m=1}^M \lambda_m^\star \nabla f_m(\boldsymbol{x}^\star) + \sum_{p=1}^P \nu_p^\star \nabla h_p(\boldsymbol{x}^\star) = \mathbf{0}.$$

Thus strong duality immediately leads to the KKT conditions holding at the solution.

Also, if you can find $\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star, \boldsymbol{\nu}^\star$ that obey the KKT conditions, not only do you know that you have a primal optimal point on your hands, but also we have strong duality (and $\boldsymbol{\lambda}^\star, \boldsymbol{\nu}^\star$ are dual optimal). For if KKT holds,

$$\nabla_{\boldsymbol{x}} L(\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star, \boldsymbol{\nu}^\star) = \mathbf{0},$$

meaning that $\boldsymbol{x}^\star$ is a minimizer of $L(\boldsymbol{x}, \boldsymbol{\lambda}^\star, \boldsymbol{\nu}^\star)$, i.e.

$$L(\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star, \boldsymbol{\nu}^\star) \leq L(\boldsymbol{x}, \boldsymbol{\lambda}^\star, \boldsymbol{\nu}^\star),$$

thus

$$\begin{aligned} g(\boldsymbol{\lambda}^\star, \boldsymbol{\nu}^\star) &= L(\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star, \boldsymbol{\nu}^\star) \\ &= f_0(\boldsymbol{x}^\star) + \sum_{m=1}^{M} \lambda_m^\star f_m(\boldsymbol{x}^\star) + \sum_{p=1}^{P} \nu_p^\star h_p(\boldsymbol{x}^\star) \\ &= f_0(\boldsymbol{x}^\star), \quad \text{(by KKT)}, \end{aligned}$$

and we have strong duality.

The upshot of this is that the conditions for strong duality are essentially the same as those under which KKT is necessary.

The program (1) and its dual (2) have strong duality if the f_m are affine inequality constraints, or there is an $\boldsymbol{x} \in \mathbb{R}^N$ such that for all the f_i which are not affine we have $f_i(\boldsymbol{x}) < 0$.

Examples

1. **Inequality LP**. Calculate the dual of

$$\underset{\boldsymbol{x}\in\mathbb{R}^N}{\text{minimize}} \ \langle \boldsymbol{x}, \boldsymbol{c}\rangle \quad \text{subject to} \quad \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}.$$

Answer: The Lagrangian is

$$\begin{aligned} L(\boldsymbol{x}, \boldsymbol{\lambda}) &= \langle \boldsymbol{x}, \boldsymbol{c}\rangle + \sum_{m=1}^{M} \lambda_m \left(\langle \boldsymbol{x}, \boldsymbol{a}_m\rangle - b_m\right) \\ &= \boldsymbol{c}^{\mathrm{T}}\boldsymbol{x} - \boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{b} + \boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x}. \end{aligned}$$

This is a linear functional in $\boldsymbol{x}$ — it is unbounded below unless

$$\boldsymbol{c} + \boldsymbol{A}^{\mathrm{T}}\boldsymbol{\lambda} = \boldsymbol{0}.$$

Thus

$$\begin{aligned} g(\boldsymbol{\lambda}) &= \inf_{\boldsymbol{x}} \left(\boldsymbol{c}^{\mathrm{T}}\boldsymbol{x} - \boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{b} + \boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x}\right) \\ &= \begin{cases} -\langle \boldsymbol{\lambda}, \boldsymbol{b}\rangle, & \boldsymbol{c} + \boldsymbol{A}^{\mathrm{T}}\boldsymbol{\lambda} = \boldsymbol{0} \\ -\infty, & \text{otherwise.} \end{cases} \end{aligned}$$

So the Lagrange dual program is

$$\begin{aligned} \underset{\boldsymbol{\lambda}\in\mathbb{R}^M}{\text{maximize}} \ -\langle \boldsymbol{\lambda}, \boldsymbol{b}\rangle \quad \text{subject to} \quad & \boldsymbol{A}^{\mathrm{T}}\boldsymbol{\lambda} = -\boldsymbol{c} \\ & \boldsymbol{\lambda} \geq \boldsymbol{0}. \end{aligned}$$

2. **Least-squares**. Calculate the dual of

$$\underset{\boldsymbol{x}\in\mathbb{R}^N}{\text{minimize}} \ \|\boldsymbol{x}\|_2^2 \quad \text{subject to} \quad \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}.$$

Check that the duality gap is zero.

Answer: The Lagrangian is

$$L(\boldsymbol{x}, \boldsymbol{\lambda}) = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{x} - \boldsymbol{\nu}^{\mathrm{T}}\boldsymbol{b} + \boldsymbol{\nu}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x}.$$

This is quadratic in $\boldsymbol{x}$ and will attain its minimum for

$$\boldsymbol{x} = -\frac{1}{2}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\nu}.$$

Thus

$$\begin{aligned} g(\boldsymbol{\nu}) &= \frac{1}{4}\boldsymbol{\nu}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\nu} - \boldsymbol{b}^{\mathrm{T}}\boldsymbol{\nu} - \frac{1}{2}\boldsymbol{\nu}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\nu} \\ &= -\frac{1}{4}\boldsymbol{\nu}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\nu} - \boldsymbol{b}^{\mathrm{T}}\boldsymbol{\nu}, \end{aligned}$$

and the Lagrange dual problem is

$$\underset{\boldsymbol{\nu}\in\mathbb{R}^M}{\text{maximize}} -\frac{1}{4}\boldsymbol{\nu}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\nu} - \boldsymbol{b}^{\mathrm{T}}\boldsymbol{\nu}.$$

Note that this will be maximized when $-\frac{1}{2}\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\nu} = \boldsymbol{b}$, which, when substituted into the dual problem yields

$$-\frac{1}{4}\boldsymbol{\nu}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\nu} + \frac{1}{2}\boldsymbol{\nu}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\nu} = \frac{1}{4}\boldsymbol{\nu}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\nu} = \left\| -\frac{1}{2}\boldsymbol{A}^{T}\boldsymbol{\nu} \right\|_2^2,$$

which shows that strong duality holds.

3. **Minimum norm.** Calculate the dual of

$$\underset{\boldsymbol{x}\in\mathbb{R}^N}{\text{minimize}} \ \|\boldsymbol{x}\| \quad \text{subject to} \quad \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b},$$

where $\|\cdot\|$ is a general valid norm.

Answer: Use $f_0(\boldsymbol{x}) = \|\boldsymbol{x}\|$ to ease notation below. We start with the Lagrangian:

$$\begin{aligned} L(\boldsymbol{x}, \boldsymbol{\nu}) &= f_0(\boldsymbol{x}) + \sum_{p=1}^{P} \nu_p(\langle \boldsymbol{x}, \boldsymbol{a}_m \rangle - b_m) \\ &= f_0(\boldsymbol{x}) - \langle \boldsymbol{\nu}, \boldsymbol{b} \rangle + (\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\nu})^{\mathrm{T}}\boldsymbol{x} \end{aligned}$$

and so

$$\begin{aligned} g(\boldsymbol{\nu}) &= -\langle \boldsymbol{\nu}, \boldsymbol{b} \rangle + \inf_{\boldsymbol{x}} \left(f_0(\boldsymbol{x}) + (\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\nu})^{\mathrm{T}}\boldsymbol{x} \right) \\ &= -\langle \boldsymbol{\nu}, \boldsymbol{b} \rangle - \sup_{\boldsymbol{x}} \left(-f_0(\boldsymbol{x}) - (\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\nu})^{\mathrm{T}}\boldsymbol{x} \right) \\ &= -\langle \boldsymbol{\nu}, \boldsymbol{b} \rangle - f_0^*(-\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\nu}), \end{aligned}$$

where f_0^* is the Fenchel dual of f_0:

$$f_0^\star(\boldsymbol{y}) = \sup_{\boldsymbol{x}}(\langle \boldsymbol{x}, \boldsymbol{y} \rangle - f_0(\boldsymbol{x})).$$

With $f_0 = \|\cdot\|$, we know already that

$$f_0^\star(\boldsymbol{y}) = \begin{cases} 0, & \|\boldsymbol{y}\|_* \leq 1, \\ \infty, & \text{otherwise} \end{cases},$$

so

$$g(\nu) = \begin{cases} -\langle \boldsymbol{\nu}, \boldsymbol{b} \rangle, & \|\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\nu}\|_* \leq 1 \\ -\infty, & \text{otherwise} \end{cases}.$$

Thus the dual program is

$$\underset{\boldsymbol{\nu} \in \mathbb{R}^P}{\text{maximize}} \quad -\langle \boldsymbol{\nu}, \boldsymbol{b} \rangle \quad \text{subject to} \quad \|\boldsymbol{A}^{\mathrm{T}} \boldsymbol{\nu}\|_* \leq 1,$$

where $\|\cdot\|_*$ is the dual norm of $\|\cdot\|$. Note that since $\boldsymbol{\nu}$ is unconstrained and $\|-\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\nu}\|_* = \|\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\nu}\|_*$, we can equivalently express this as

$$\underset{\boldsymbol{\nu} \in \mathbb{R}^P}{\text{maximize}} \quad \langle \boldsymbol{\nu}, \boldsymbol{b} \rangle \quad \text{subject to} \quad \|\boldsymbol{A}^{\mathrm{T}} \boldsymbol{\nu}\|_* \leq 1,$$

which is identical to the Fenchel dual derived earlier in this course.