Convex Optimization Overview

Duke ML Course Notes
Cynthia Rudin

Credit: Boyd, Ng and Knowles

We want to solve differentiable convex optimization problems of this form, which we call OPT:

$$\begin{aligned} \underset{\mathbf{x}\in\mathbb{R}^n}{\text{minimize}} \quad & f(\mathbf{x}) \\ \text{subject to} \quad & g_i(\mathbf{x}) \leq 0,\ i = 1, \ldots, m, \\ & h_i(\mathbf{x}) = 0,\ i = 1 \ldots, p, \end{aligned}$$

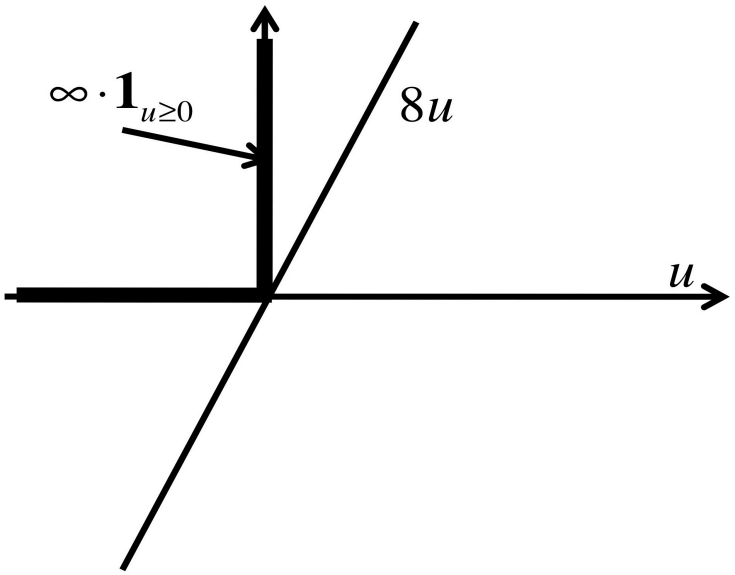
where $\mathbf{x} \in \mathbb{R}^n$ is the *optimization variable*, $f : \mathbb{R}^n \to \mathbb{R}$, $g_i : \mathbb{R}^n \to \mathbb{R}$ are *differentiable convex functions*, and $h_i : \mathbb{R}^n \to \mathbb{R}$ are *affine functions*.

Recall that a function $g : G \to \mathbb{R}$ is convex if G is a convex set, and for any $\mathbf{x}, \mathbf{z} \in G$ and $\theta \in [0, 1]$, we have $g(\theta\mathbf{x}+(1-\theta)\mathbf{z}) \leq \theta g(\mathbf{x})+(1-\theta)g(\mathbf{z})$. A function g is concave if $-g$ is convex. An affine function has the form $h(\mathbf{x}) = \mathbf{a}^T\mathbf{x} + b$ for some $\mathbf{a} \in \mathbb{R}^n, b \in \mathbb{R}$. (Affine functions are both convex and concave.)

We could rewrite the OPT with the constraints in the objective:

$$\min_{\mathbf{x}} \Theta_{\mathcal{P}}(\mathbf{x}) \quad \text{where}$$
$$\Theta_{\mathcal{P}}(\mathbf{x}) := f(\mathbf{x}) + \infty \sum_{i=1}^{m} \mathbb{1}_{[g_i(\mathbf{x})>0]} + \infty \sum_{i=1}^{p} \mathbb{1}_{[h_i(\mathbf{x})\neq 0]}. \tag{1}$$

But this is hard to optimize because it is non-differentiable and not even continuous. Why don't we replace $\infty \times \mathbb{1}_{[u\geq 0]}$ with something nicer? A line αu seems like a dumb choice...



but for $\alpha \geq 0$ the penalty is in the right direction (we are penalized for constraints being dissatisfied, and αu is a lower bound on $\infty \times \mathbb{1}_{[g_i(\mathbf{x})>0]}$).

Similarly, βu is a lower bound for $\infty \times \mathbb{1}_{[u \neq 0]}$ no matter what the sign of β is.

For each constraint i, replacing $\infty \times \mathbb{1}_{[u \geq 0]}$ by αu, where $\alpha \geq 0$ in the objective, and similarly replacing the $\infty \times \mathbb{1}_{[u \neq 0]}$ term by βu, gives the *Lagrangian*

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = f(\mathbf{x}) + \sum_{i=1}^{m} \alpha_i g_i(\mathbf{x}) + \sum_{i=1}^{p} \beta_i h_i(\mathbf{x}). \tag{2}$$

We refer to $\mathbf{x} \in \mathbf{R}^n$ as the *primal variables* of the Lagrangian. The second argument of the Lagrangian is a vector $\boldsymbol{\alpha} \in \mathbf{R}^m$. The third is a vector $\boldsymbol{\beta} \in \mathbf{R}^p$. Elements of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are collectively known as the *dual variables* of the Lagrangian, or *Lagrange multipliers*.

If we take the maximum of $\mathcal{L}$ with respect to $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, where $\alpha_i \geq 0$, we recover OPT. Let's show this.

For a particular $\mathbf{x}$, let's say the constraints are satisfied. So $g_i(\mathbf{x}) \leq 0$, and to make the term $\alpha_i g_i(\mathbf{x})$ as high as it can be, we set $\alpha_i = 0 \ \forall i$. Also, since $h_i(\mathbf{x}) = 0 \ \forall i$, the β_i's can be anything and it won't change $\mathcal{L}$. To summarize, if the constraints are satisfied, the value of $\max_{\boldsymbol{\alpha},\boldsymbol{\beta}} \mathcal{L}$ is just $f(\mathbf{x})$, which is the same as the value of $\Theta_{\mathcal{P}}$.

If any constraint is not satisfied, then we can make $\mathcal{L}(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ infinite by fiddling with α_i (making it go to ∞) or β_i (making it go to $\pm\infty$), and then again the value of $\max_{\boldsymbol{\alpha},\boldsymbol{\beta}} \mathcal{L}$ is the same as the value of $\Theta_{\mathcal{P}}$.

Try to figure out how the fiddling works.

So we have formally:

$$\Theta_{\mathcal{P}}(\mathbf{x}) = \max_{\boldsymbol{\alpha},\boldsymbol{\beta};\alpha_i \geq 0, \forall i} \mathcal{L}(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}).$$

Remember that we want to minimize $\Theta_{\mathcal{P}}(\mathbf{x})$. This problem is the *primal problem*.

Specifically, the primal problem is:

$$\min_{\mathbf{x}} \left[\max_{\boldsymbol{\alpha},\boldsymbol{\beta}:\alpha_i \geq 0, \forall i} \mathcal{L}(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \right] = \min_{\mathbf{x}} \ \Theta_{\mathcal{P}}(\mathbf{x}). \tag{3}$$

In the equation above, the function $\Theta_{\mathcal{P}} : \mathbb{R}^n \to \mathbb{R}$ is called the *primal objective*.

We say that a point $\mathbf{x} \in \mathbb{R}^n$ is *primal feasible* if $g_i(\mathbf{x}) \leq 0, i = 1, \ldots, m$ and $h_i(\mathbf{x}) = 0, i = 1, \ldots, p$. The vector $\mathbf{x}^* \in \mathbb{R}^n$ denotes the solution of (3), and $p^* = \Theta_{\mathcal{P}}(\mathbf{x}^*)$ denotes the optimal value of the primal objective.

It turns out that $\Theta_{\mathcal{P}}(\mathbf{x})$ is a convex function of $\mathbf{x}$. Why is that? First, $f(\mathbf{x})$ is convex. Each of the $g_i(\mathbf{x})$'s are convex functions in x, and since the α_i's are constrained to be nonnegative, then $\alpha_i g_i(\mathbf{x})$ is convex in $\mathbf{x}$ for each i. (Of course if α_i were allowed to be negative, $\alpha_i g_i(x)$ would be concave instead, and we don't want that.) Similarly, each $\beta_i h_i(\mathbf{x})$ is convex in x (regardless of the sign of β_i) since $h_i(\mathbf{x})$ is linear. Since the sum of convex functions is always convex, $\mathcal{L}$ is convex for each $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. Finally, the maximum of a collection of convex functions is again a convex function, so we can conclude that $\Theta_{\mathcal{P}}(\mathbf{x}) = \max_{\boldsymbol{\alpha},\boldsymbol{\beta}} \mathcal{L}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{x})$ is a convex function of $\mathbf{x}$.

By switching the order of the minimization and maximization above, we obtain an entirely *different* optimization problem.

The dual problem is:

$$\max_{\boldsymbol{\alpha},\boldsymbol{\beta}:\alpha_i\geq 0,\forall i}\left[\min_{\mathbf{x}}\mathcal{L}(\mathbf{x},\boldsymbol{\alpha},\boldsymbol{\beta})\right] = \max_{\boldsymbol{\alpha},\boldsymbol{\beta}:\alpha_i\geq 0,\forall i}\ \Theta_{\mathcal{D}}(\boldsymbol{\alpha},\boldsymbol{\beta}). \tag{4}$$

Function $\Theta_{\mathcal{D}} : \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$ is called the *dual objective*.

We say that $(\boldsymbol{\alpha},\boldsymbol{\beta})$ are *dual feasible* if $\alpha_i \geq 0, i = 1\ldots,m$.

Denote $(\boldsymbol{\alpha}^*,\boldsymbol{\beta}^*) \in \mathbb{R}^m \times \mathbb{R}^p$ as the solution of (4), and $d^* = \Theta_{\mathcal{D}}(\boldsymbol{\alpha}^*,\boldsymbol{\beta}^*)$ denote the optimal value of the dual objective.

The dual objective $\Theta_{\mathcal{D}}(\boldsymbol{\alpha},\boldsymbol{\beta})$, is a concave function of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. The dual objective is:

$$\Theta_{\mathcal{D}}(\boldsymbol{\alpha},\boldsymbol{\beta}) = \min_{\mathbf{x}}\mathcal{L}(\mathbf{x},\boldsymbol{\alpha},\boldsymbol{\beta}) = \min_{\mathbf{x}}\left[f(\mathbf{x}) + \sum_{i=1}^{m}\alpha_i g_i(\mathbf{x}) + \sum_{i=1}^{p}\beta_i h_i(\mathbf{x})\right].$$

For any fixed value of $\mathbf{x}$, the quantity inside the brackets is an affine function of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, and hence, concave. The $f(\mathbf{x})$ is just a constant as far as $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are concerned. Since the minimum of a collection of concave functions is also concave, we can conclude that $\Theta_{\mathcal{D}}(\boldsymbol{\alpha},\boldsymbol{\beta})$ is a concave function of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$.

Interpreting the Dual Problem

We make the following observation:

Lemma 1. *If $(\boldsymbol{\alpha},\boldsymbol{\beta})$ are dual feasible, then $\Theta_{\mathcal{D}}(\boldsymbol{\alpha},\boldsymbol{\beta}) \leq p^*$.*

Proof. Because of the lower bounds we made, namely

$$\alpha_i g_i(\mathbf{x}) \leq \infty \times \mathbb{1}_{[g_i(\mathbf{x})\geq 0]}$$

$$\beta_i h_i(\mathbf{x}) \leq \infty \times \mathbb{1}_{[h_i(\mathbf{x}) \neq 0]}$$

we have, when $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are dual feasible,

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \leq \Theta_{\mathcal{P}}(\mathbf{x}) \text{ for all } \mathbf{x}.$$

Taking the $\min_{\mathbf{x}}$ of both sides:

$$\underbrace{\min_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta})}_{\Theta_{\mathcal{D}}(\boldsymbol{\alpha}, \boldsymbol{\beta})} \leq \underbrace{\min_{\mathbf{x}} \Theta_{\mathcal{P}}(\mathbf{x})}_{p^*}.$$

■

The lemma shows that given any dual feasible $(\boldsymbol{\alpha}, \boldsymbol{\beta})$, the dual objective $\Theta_{\mathcal{D}}(\boldsymbol{\alpha}, \boldsymbol{\beta})$ provides a lower bound on the optimal value p^* of the primal problem.

Since the dual problem is $\max_{\boldsymbol{\alpha}, \boldsymbol{\beta}} \Theta_{\mathcal{D}}(\boldsymbol{\alpha}, \boldsymbol{\beta})$, the dual problem can be seen as a search for the tightest possible lower bound on p^*. This gives rise to a property of any primal and dual optimization problem pairs known as *weak duality*:

Lemma 2. *For any pair of primal and dual problems, $d^* \leq p^*$.*

Rewritten,

$$\max_{\boldsymbol{\alpha}, \boldsymbol{\beta}: \alpha_i \geq 0, \forall i} \min_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \leq \min_{\mathbf{x}} \max_{\boldsymbol{\alpha}, \boldsymbol{\beta}: \alpha_i \geq 0, \forall i} \mathcal{L}(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}).$$

Intuitively, this makes sense: on the left, we have a maximin value – this is the highest value that the $\boldsymbol{\alpha}, \boldsymbol{\beta}$ player can be sure to get without knowing the actions of the other player. Whatever the $\boldsymbol{\alpha}, \boldsymbol{\beta}$ player does, they do it thinking that the $\mathbf{x}$ player gets to react to bring the value down. On the right is the minimax value – this is the highest value the $\boldsymbol{\alpha}, \boldsymbol{\beta}$ player can get after observing what the $\mathbf{x}$ player does. The $\boldsymbol{\alpha}, \boldsymbol{\beta}$ player has the advantage in that case.

For some primal/dual optimization problems, an even stronger result holds, known as *strong duality*.

Lemma 3 *(Strong Duality). For any pair of primal and dual problems that satisfies certain technical conditions called constraint qualifications, then $d^* = p^*$.*

A number of different constraint qualifications exist, of which the most commonly invoked is *Slater's condition*: a primal/dual problem pair satisfy Slater's condition if there exists some feasible primal solution $\mathbf{x}$ for which all inequality constraints are strictly satisfied (i.e. $g_i(\mathbf{x}) < 0, i = 1 \ldots, m$). In practice, nearly all convex problems satisfy some type of constraint qualification, and hence the primal and dual problem have the same optimal value.

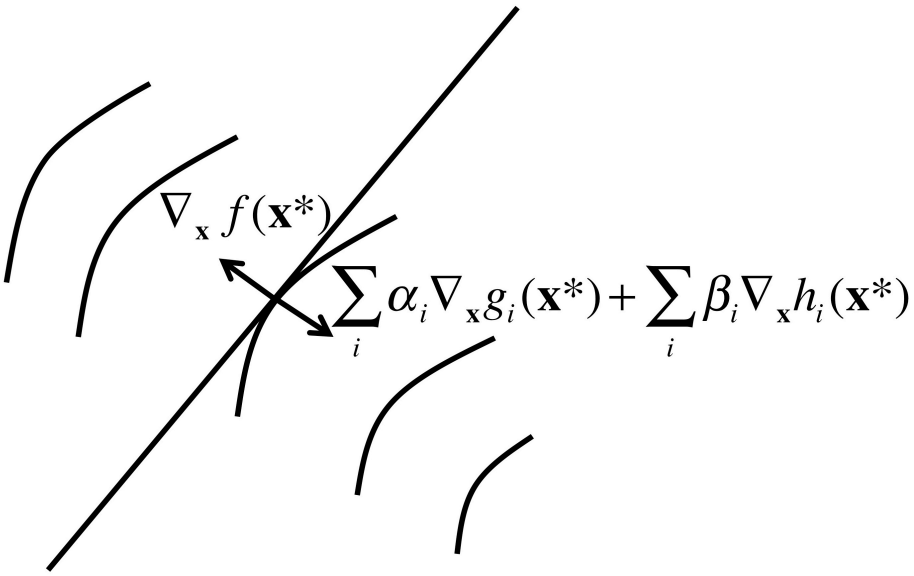
KKT Conditions

For an unconstrained convex optimization problem, we know we are at the global minimum if the gradient is zero. The KKT conditions are the equivalent conditions for the global minimum of a constrained convex optimization problem.

If strong duality holds and $(\mathbf{x}^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$ is optimal, then $\mathbf{x}^*$ minimizes $\mathcal{L}(\mathbf{x}, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$ giving us the first KKT condition, *Lagrangian stationarity*:

$$\nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)|_{\mathbf{x}^*} = \nabla_{\mathbf{x}} f(\mathbf{x})|_{\mathbf{x}^*} + \sum_i \alpha_i^* \nabla_{\mathbf{x}} g_i(\mathbf{x})|_{\mathbf{x}^*} + \sum_i \beta_i^* \nabla_{\mathbf{x}} h_i(\mathbf{x})|_{\mathbf{x}^*} = 0$$

We can interpret this condition by saying that the gradient of the objective function and constraint function must be parallel (and opposite). This concept is illustrated for a simple 2D optimization problem with one inequality constraint below.



The curves are contours of f, and the line is the constraint boundary. At $\mathbf{x}^*$, the gradient of f and gradient of the constraint must be parallel and opposing so that we couldn't move along the constraint boundary in order to get an improved objective value.

One interesting consequence of strong duality is next:

Lemma 4 *(Complementary Slackness). If strong duality holds, then* $\alpha_i^* g_i(\mathbf{x}^*) = 0$ *for each* $i = 1 \ldots, m$.

Proof. Suppose that strong duality holds.

$$\begin{aligned} p^* = d^* &= \Theta_{\mathcal{D}}(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*) \\ &= \min_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*) \\ &\leq \mathcal{L}(\mathbf{x}^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*) \\ &\leq \max_{\boldsymbol{\alpha}, \boldsymbol{\beta}; \alpha_i \geq 0, \forall i} \mathcal{L}(\mathbf{x}^*, \boldsymbol{\alpha}, \boldsymbol{\beta}) \\ &= \Theta_{\mathcal{P}}(\mathbf{x}^*) = f(\mathbf{x}^*) = p^*. \end{aligned}$$

The second line is the definition of $\Theta_{\mathcal{D}}$. The inequality in the third line says that $\min_{\mathbf{x}}$ is lower than choosing any $\mathbf{x}$, in particular $\mathbf{x}^*$. The inequality in the fourth line says that the choice of $\boldsymbol{\alpha}^*$ and $\boldsymbol{\beta}^*$ is not as large as the largest possible values for those variables. The first equality in the last line is from what we proved earlier. The second equality comes from the fact that the optimal solution is obtained when the constraints hold.
This means that all the inequalities are actually equalities. In particular,

$$\mathcal{L}(\mathbf{x}^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*) = f(\mathbf{x}^*) + \sum_i^m \alpha_i^* g_i(\mathbf{x}^*) + \sum_i^p \beta_i^* h_i(\mathbf{x}^*) = f(\mathbf{x}^*).$$

So,

$$\sum_i^m \alpha_i^* g_i(\mathbf{x}^*) + \sum_i^p \beta_i^* h_i(\mathbf{x}^*) = 0.$$

- Since $\mathbf{x}^*$ is primal feasible, each $h_i(\mathbf{x}^*) = 0$, so the second terms are all 0.
- Since α_i^*'s are dual feasible, $\alpha_i^* \geq 0$, and since $\mathbf{x}^*$ is primal feasible, $g_i(\mathbf{x}^*) \leq 0$.

So each $\alpha_i^* g_i(\mathbf{x}^*) \leq 0$, which means they are all 0.

$$\alpha_i^* g_i(\mathbf{x}^*) = 0 \ \ \forall i = 1, \ldots, m.$$

We can rewrite complementary slackness this way:

$$\begin{aligned} \alpha_i^* > 0 &\Longrightarrow \ \ g_i(\mathbf{x}^*) = 0 \ \ (\textit{“active” or “binding” constraints}) \\ g_i(\mathbf{x}^*) < 0 &\Longrightarrow \ \ \ \ \alpha_i^* = 0. \end{aligned}$$

In the case of support vector machines (SVMs), active constraints are known as *support vectors.*

We can now characterize the optimal conditions for a primal dual optimization pair:

Theorem 1 *Suppose that* $\mathbf{x}^* \in \mathbb{R}^n$, $\boldsymbol{\alpha}^* \in \mathbb{R}^m$, *and* $\boldsymbol{\beta}^* \in \mathbb{R}^p$ *satisfy the following conditions:*

- *(Primal feasibility)* $g_i(\mathbf{x}^*) \leq 0, i = 1, ..., m$ and $h_i(\mathbf{x}^*) = 0, i = 1, ..., p$.
- *(Dual feasibility)* $\alpha_i^* \geq 0, i = 1, \ldots, m$.
- *(Complementary Slackness)* $\alpha_i^* g_i(\mathbf{x}^*) = 0, i = 1, \ldots, m$.
- *(Lagrangian stationary)* $\nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*) = \mathbf{0}$.

Then $\mathbf{x}^*$ *is primal optimal and* $(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$ *are dual optimal. Furthermore, if strong duality holds, then any primal optimal* $\mathbf{x}^*$ *and dual optimal* $(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$ *must satisfy all these conditions.*

These conditions are known as the *Karush-Kuhn-Tucker (KKT) conditions.*[1]

[1]Incidentally, the KKT theorem has an interesting history. The result was originally derived by Karush in his 1939 master's thesis but did not catch any attention until it was rediscovered in 1950 by two mathematicians Kuhn and Tucker. A variant of essentially the same result was also derived by John in 1948.