III. Constrained Convex Optimization

Optimality conditions for constrained optimization

When we are solving an unconstrained optimization problem, the goal is clear: we want to find a point where the gradient vanishes. All of the algorithms we looked at over the last few lectures were in service of this condition. Once we add constraints, the optimality conditions are more complicated, and involve relationships between the gradient of the functional we are minimizing along with the gradients of the constraints — these are the so-called Karush-Kuhn-Tucker (KKT) conditions.

We will build up to the KKT conditions slowly. We will first derive a general (and very easy to prove) *geometric* necessary and sufficient condition for $\boldsymbol{x}^\star$ to be a minimizer of a constrained optimization program. We will then show how this simple result immediately yields the KKT conditions for certain kinds of constraints. In the next set of notes, we will derive the KKT conditions, show that they are always sufficient, and discuss conditions under which they are also necessary.

To keep things simpler, in our initial discussion of constrained optimization, we will restrict our focus to *smooth* optimization problems. As before, most of what we have to say can be extended to the non-smooth case by simply replacing gradients with subgradients, but we will assume that our objective function (and eventually, our constraints) are differentiable for the time being.

We start by considering the general constrained problem

$$\underset{\boldsymbol{x} \in \mathcal{C}}{\text{minimize}}\, f(\boldsymbol{x})$$

where $\mathcal{C}$ is a closed, convex set, and f is again a convex function.

We have the following fundamental result:

Let f be a differentiable convex function, and $\mathcal{C}$ be a closed convex set. Then $\boldsymbol{x}^\star$ is a minimizer of

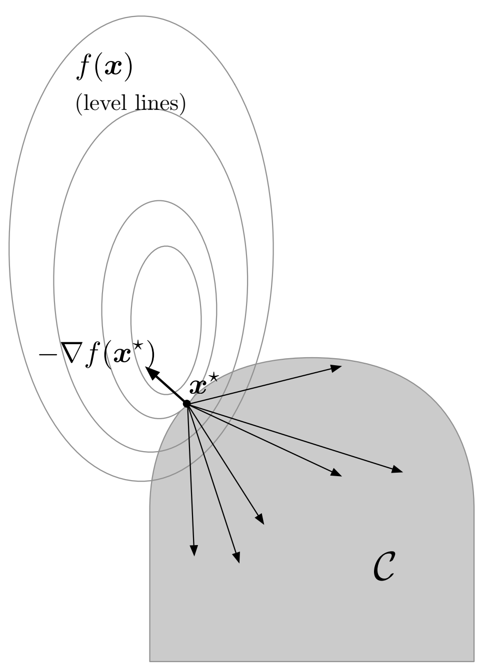
$$\underset{\boldsymbol{x}\in\mathcal{C}}{\text{minimize}}\, f(\boldsymbol{x})$$

if and only if $\boldsymbol{x}^\star \in \mathcal{C}$ and

$$\langle \boldsymbol{y} - \boldsymbol{x}^\star, \nabla f(\boldsymbol{x}^\star)\rangle \;\geq\; 0$$

for all $\boldsymbol{y} \in \mathcal{C}$.

This result is geometrically intuitive; it is saying that every vector from $\boldsymbol{x}^\star$ to another point $\boldsymbol{y}$ in $\mathcal{C}$ must make an **obtuse** angle with $-\nabla f(\boldsymbol{x}^\star)$. That is, there cannot be any descent directions from $\boldsymbol{x}^\star$ that lead to another point in $\mathcal{C}$. Here is a picture:



To prove this, we first argue that $\langle \boldsymbol{y} - \boldsymbol{x}^\star, \nabla f(\boldsymbol{x}^\star)\rangle \geq 0$ for all $\boldsymbol{y} \in \mathcal{C}$ implies that $\boldsymbol{x}^\star$ is optimal. Since f is convex, for any $\boldsymbol{y} \in \mathcal{C}$

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}^\star) + \langle \boldsymbol{y} - \boldsymbol{x}^\star, \nabla f(\boldsymbol{x}^\star)\rangle,$$

and so

$$f(\boldsymbol{y}) - f(\boldsymbol{x}^\star) \geq \langle \boldsymbol{y} - \boldsymbol{x}^\star, \nabla f(\boldsymbol{x}^\star)\rangle \;\geq\; 0,$$

Since this holds for every $\boldsymbol{y} \in \mathcal{C}$, $\boldsymbol{x}^\star$ is a minimizer.

Now suppose that $\boldsymbol{x}^\star$ is a minimizer. If there were a $\boldsymbol{y} \in \mathcal{C}$ such that $\langle \boldsymbol{y} - \boldsymbol{x}^\star, \nabla f(\boldsymbol{x}^\star)\rangle < 0$, then $\boldsymbol{d} = \boldsymbol{y} - \boldsymbol{x}^\star$ would be a descent direction, and there would exist a $0 < t < 1$ such that

$$f(\boldsymbol{x}^\star + t(\boldsymbol{y} - \boldsymbol{x}^\star)) < f(\boldsymbol{x}^\star).$$

Since $\mathcal{C}$ is convex and $\boldsymbol{x}^\star, \boldsymbol{y} \in \mathcal{C}$, we know $\boldsymbol{x}^\star + t(\boldsymbol{y} - \boldsymbol{x}^\star) \in \mathcal{C}$. But this contradicts the assertion that $\boldsymbol{x}^\star$ is a minimizer, and so no such $\boldsymbol{y}$ can exist.

Examples

The abstract geometrical result in the previous section will eventually lead us to the Karush-Kuhn-Tucker (KKT) conditions. But we will build up to this by looking at what it tells us in several important (and prevalent) cases.

We assume throughout this section that f is convex, differentiable, and defined on all of $\mathbb{R}^N$.

Linear constraints

Consider a convex optimization problem with linear[1] constraints,

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

where $\boldsymbol{A}$ is $M \times N$ and $\boldsymbol{b} \in \mathbb{R}^M$. At a solution $\boldsymbol{x}^\star$, we have

$$\langle \boldsymbol{y} - \boldsymbol{x}^\star, \nabla f(\boldsymbol{x}^\star)\rangle \ \geq \ 0,$$

for all $\boldsymbol{y}$ such that $\boldsymbol{A}\boldsymbol{y} = \boldsymbol{b}$. Since $\boldsymbol{A}\boldsymbol{x}^\star = \boldsymbol{b}$ as well, this is equivalent to

$$\langle \boldsymbol{h}, \nabla f(\boldsymbol{x}^\star)\rangle \geq 0, \quad \text{for all} \ \ \boldsymbol{h} \in \operatorname{Null}(\boldsymbol{A}).$$

Since $\boldsymbol{h} \in \operatorname{Null}(\boldsymbol{A}) \Leftrightarrow -\boldsymbol{h} \in \operatorname{Null}(\boldsymbol{A})$, we must have

$$\langle \boldsymbol{h}, \nabla f(\boldsymbol{x}^\star)\rangle = 0, \quad \text{for all} \ \ \boldsymbol{h} \in \operatorname{Null}(\boldsymbol{A}),$$

i.e. the gradient is **orthogonal** to the null space of $\boldsymbol{A}$. This means that it is in the row space,

$$\nabla f(\boldsymbol{x}^\star) \in \operatorname{Range}(\boldsymbol{A}^{\mathrm{T}}),$$

and so there is a $\boldsymbol{\nu} \in \mathbb{R}^M$ such that

$$\nabla f(\boldsymbol{x}^\star) + \boldsymbol{A}^{\mathrm{T}}\boldsymbol{\nu} = \boldsymbol{0}.$$

[1]We really should be saying *affine constraints*, but "linear constraints" is typical nomenclature for this type of problem.

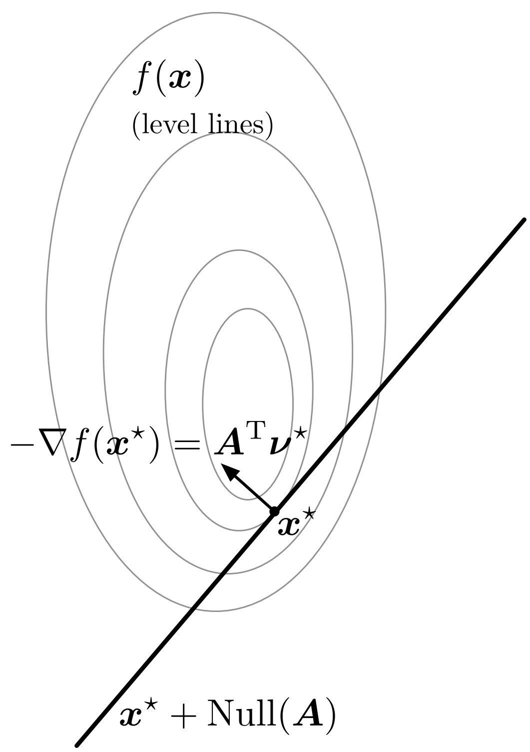
Summary:

$x^\star$ is a solution to

$$\underset{x\in\mathbb{R}^N}{\text{minimize}}\, f(x) \quad \text{subject to} \quad Ax = b,$$

if and only if

1. $Ax^\star = b$, and
2. there exists a $\nu^\star \in \mathbb{R}^M$ such that $\nabla f(x^\star) + A^{\mathrm{T}}\nu^\star = \mathbf{0}$.



Non-negativity constraints

Now consider the convex program

$$\underset{\boldsymbol{x}\in\mathbb{R}^N}{\text{minimize}} \;\; f(\boldsymbol{x}) \quad \text{subject to} \quad \boldsymbol{x} \geq \boldsymbol{0}.$$

At a solution $\boldsymbol{x}^\star$, we will have

$$\langle \boldsymbol{y} - \boldsymbol{x}^\star, \nabla f(\boldsymbol{x}^\star)\rangle \geq 0, \quad \text{for all} \;\; \boldsymbol{y} \in \mathbb{R}^N_+. \tag{1}$$

Since both $\boldsymbol{0} \in \mathbb{R}^N_+$ and $2\boldsymbol{x}^\star \in \mathbb{R}^N_+$, this means

$$\langle \boldsymbol{x}^\star, \nabla f(\boldsymbol{x}^\star)\rangle = 0, \tag{2}$$

and so

$$\langle \boldsymbol{y}, \nabla f(\boldsymbol{x}^\star)\rangle \geq 0, \quad \text{for all} \;\; \boldsymbol{y} \in \mathbb{R}^N_+,$$

meaning that the gradient has only non-negative values as well,

$$\nabla f(\boldsymbol{x}^\star) \geq \boldsymbol{0}. \tag{3}$$

The conditions (2) and (3) are sufficient as well, as together they imply (1).

Note that condition (3) is the same as saying there exists a $\boldsymbol{\lambda}^\star \geq \boldsymbol{0}$ such that

$$\nabla f(\boldsymbol{x}^\star) - \boldsymbol{\lambda}^\star = \boldsymbol{0}.$$

We can also see that (2) and (3), along with the fact that $\boldsymbol{x}^\star \in \mathbb{R}^N_+$, mean that $\nabla f(\boldsymbol{x}^\star)$ and $\boldsymbol{x}^\star$ can only be non-zero at different indices:

$$\begin{aligned} [\nabla f(\boldsymbol{x}^\star)]_n > 0 &\Rightarrow \;\; x_n = 0, \\ x_n > 0 &\Rightarrow [\nabla f(\boldsymbol{x}^\star)]_n = 0. \end{aligned}$$

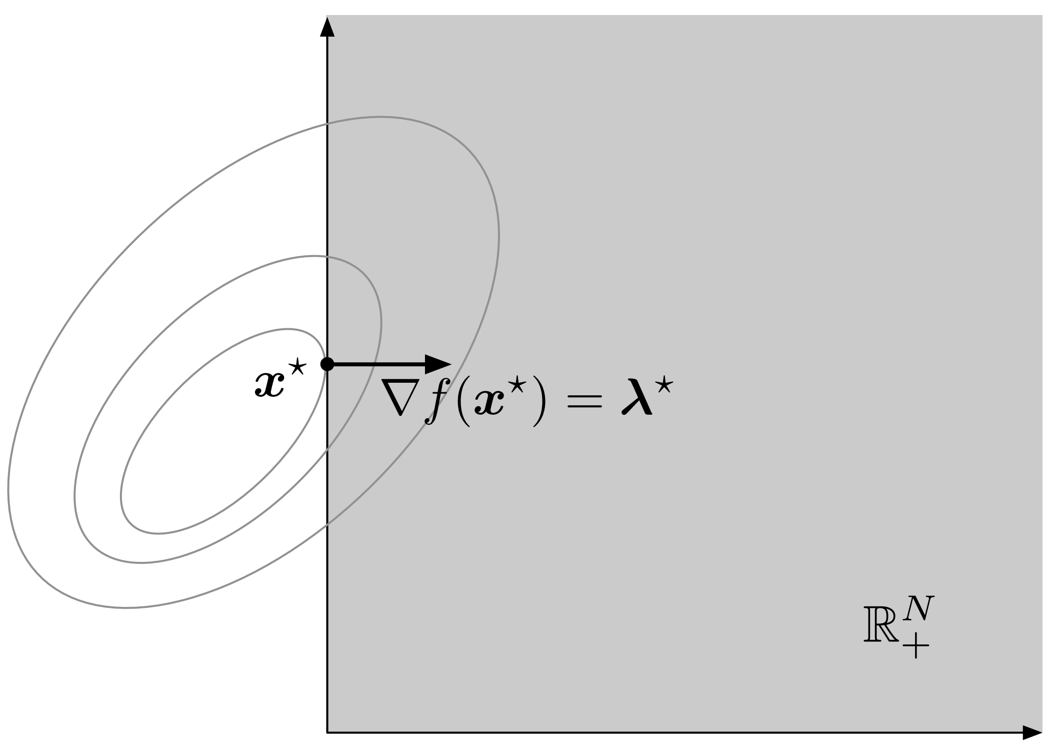
Summary:

$\boldsymbol{x}^\star$ is a solution to

$$\underset{\boldsymbol{x}\in\mathbb{R}^N}{\text{minimize}} \ f(\boldsymbol{x}) \quad \text{subject to} \quad \boldsymbol{x} \geq \mathbf{0},$$

if and only if

1. $\boldsymbol{x}^\star \geq \mathbf{0}$,
   and there exists a $\boldsymbol{\lambda}^\star \in \mathbb{R}^N$ such that
2. $\boldsymbol{\lambda}^\star \geq \mathbf{0}$, and
3. $\lambda_n x_n = 0$ for all $n = 1, \ldots, N$, and
4. $\nabla f(\boldsymbol{x}^\star) - \boldsymbol{\lambda}^\star = \mathbf{0}$.



A single convex inequality constraint

Now consider the convex program

$$\underset{\boldsymbol{x}\in\mathbb{R}^N}{\text{minimize}}\ f(\boldsymbol{x}) \quad \text{subject to} \quad g(\boldsymbol{x}) \leq 0,$$

where g is also a differentiable convex function. We will argue that in this case, the optimality conditions for $\boldsymbol{x}^\star$,

$$g(\boldsymbol{x}^\star) \leq 0, \quad \text{and} \quad \langle \boldsymbol{y} - \boldsymbol{x}^\star, \nabla f(\boldsymbol{x}^\star)\rangle \geq 0, \ \text{ for all } \boldsymbol{y} \text{ with } g(\boldsymbol{y}) \leq 0,$$

are equivalent to one of these two conditions holding,

1. $g(\boldsymbol{x}^\star) < 0$ and $\nabla f(\boldsymbol{x}^\star) = \mathbf{0}$, or
2. $g(\boldsymbol{x}^\star) = 0$ and the gradients of g and f are negatively aligned:

$$\nabla g(\boldsymbol{x}^\star) = -\lambda \nabla f(\boldsymbol{x}^\star), \quad \text{for some } \ \lambda > 0.$$

Establishing this relies on the following geometric fact:[2]

Let $\boldsymbol{u}, \boldsymbol{v}$ be vectors in $\mathbb{R}^N$. If no $\boldsymbol{d}$ exists such that

$$\langle \boldsymbol{d}, \boldsymbol{u}\rangle < 0, \quad \text{and} \quad \langle \boldsymbol{d}, \boldsymbol{v}\rangle < 0 \quad \text{simultaneously,} \tag{4}$$

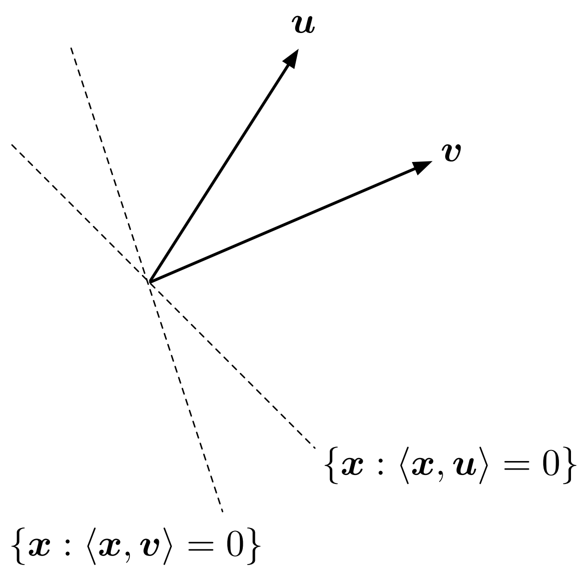
then $\boldsymbol{u}$ and $\boldsymbol{v}$ are negatively aligned,

$$\boldsymbol{u} = -\lambda \boldsymbol{v}, \quad \text{for some } \ \lambda > 0. \tag{5}$$

The converse also holds, as if (5) is true, there is no way (4) can be true.

[2]This is a special case of the famous Gordan Theorem.

The argument for this is simple. The sets $\{\boldsymbol{x} : \langle \boldsymbol{x}, \boldsymbol{u} \rangle < 0\}$ and $\{\boldsymbol{x} : \langle \boldsymbol{x}, \boldsymbol{v} \rangle < 0\}$ are open half spaces, and these half spaces are disjoint if and only if (5) holds.



Now suppose that $\boldsymbol{x}^\star$, with $g(\boldsymbol{x}^\star) \leq 0$, is a minimizer. We know that $\nabla g(\boldsymbol{x}^\star)$ and $\nabla f(\boldsymbol{x}^\star)$ must be negatively aligned, as otherwise our geometric fact dictates that there is a $\boldsymbol{d}$ that is a descent direction for both g and f, meaning there is a $0 < t < 1$ such that

$$\begin{aligned} f(\boldsymbol{x}^\star + t\boldsymbol{d}) &< f(\boldsymbol{x}^\star), \text{ and} \\ g(\boldsymbol{x}^\star + t\boldsymbol{d}) &< g(\boldsymbol{x}^\star) \leq 0. \end{aligned}$$

This would mean that there is a feasible point at which f is smaller than it is at $\boldsymbol{x}^\star$, directly contradicting the assertion that $\boldsymbol{x}^\star$ is a minimizer. Thus no such $\boldsymbol{d}$ can exist.

Suppose now that there is an $\boldsymbol{x}^\star$ such that $g(\boldsymbol{x}^\star) = 0$ and a $\lambda > 0$ so that $\nabla g(\boldsymbol{x}^\star) = -\lambda \nabla f(\boldsymbol{x}^\star)$. Let $\boldsymbol{x}$ be any other feasible point; $g(\boldsymbol{x}) \leq 0$. Then, by the convexity of g,

$$g(\boldsymbol{x}^\star + \theta(\boldsymbol{x} - \boldsymbol{x}^\star)) \leq 0, \quad \text{for all } 0 \leq \theta \leq 1.$$

Since the above is true for *all* θ in this range, we know that $\boldsymbol{x} - \boldsymbol{x}^\star$ cannot be an ascent direction for g from $\boldsymbol{x}^\star$. Thus

$$\langle \boldsymbol{x} - \boldsymbol{x}^\star, \nabla g(\boldsymbol{x}^\star) \rangle \le 0.$$

Since $\nabla g(\boldsymbol{x}^\star) = -\lambda \nabla f(\boldsymbol{x}^\star)$, we now know

$$\langle \boldsymbol{x} - \boldsymbol{x}^\star, \nabla f(\boldsymbol{x}^\star) \rangle \ge 0.$$

Then by the convexity of f,

$$\begin{aligned} f(\boldsymbol{x}) &\ge f(\boldsymbol{x}^\star) + \langle \boldsymbol{x} - \boldsymbol{x}^\star, \nabla f(\boldsymbol{x}^\star) \rangle \\ &\ge f(\boldsymbol{x}^\star), \end{aligned}$$

and so $\boldsymbol{x}^\star$ is a minimizer.

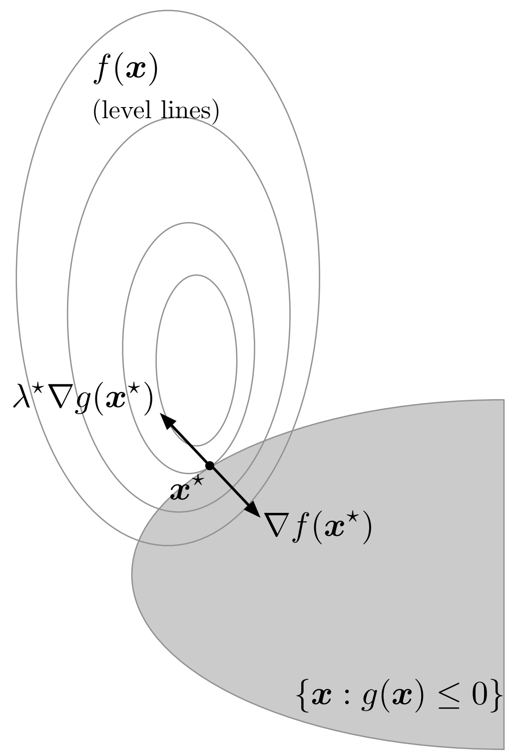
We can collect all of this into the following summary:

$\boldsymbol{x}^\star$ is a solution to

$$\underset{\boldsymbol{x} \in \mathbb{R}^N}{\text{minimize}} \; f(\boldsymbol{x}) \quad \text{subject to} \quad g(\boldsymbol{x}) \le 0,$$

if and only if

1. $g(\boldsymbol{x}^\star) \le 0$,
   and there exists a $\lambda^\star \in \mathbb{R}$ such that
2. $\lambda^\star \ge 0$, and
3. $\lambda^\star g(\boldsymbol{x}^\star) = 0$, and
4. $\nabla f(\boldsymbol{x}^\star) + \lambda^\star \nabla g(\boldsymbol{x}^\star) = \mathbf{0}$.


$f(\boldsymbol{x})$
(level lines)
$\lambda^\star \nabla g(\boldsymbol{x}^\star)$
$\boldsymbol{x}^\star$
$\nabla f(\boldsymbol{x}^\star)$
$\{\boldsymbol{x} : g(\boldsymbol{x}) \leq 0\}$