Learning objectives. By the end of class, you should be able to

1. show that $\nabla f(x) = 0$ is the Farkas certificate that no descent direction exists;
2. define the tangent cone and the linearized cone, and derive the KKT conditions as the Farkas certificate that no linearized direction descends;
3. show that the tangent cone lies in the linearized cone, so that every KKT point is first-order optimal, and a global minimizer for a convex problem;
4. explain why the KKT conditions are necessary only under a constraint qualification, and check LICQ, MFCQ, and Slater's condition;
5. solve convex and nonconvex problems by computing their KKT points.

1 Recap: First-order optimality

Question 1. Let f be convex and differentiable. How do we minimize f without constraints?

Solution: Solve $\nabla f(x) = 0$. Every solution is a global minimizer.

If f is nonconvex, a solution of $\nabla f(x) = 0$ may be a saddle point or a local maximizer, but it is a good guess, and we can list all solutions and keep the best one.

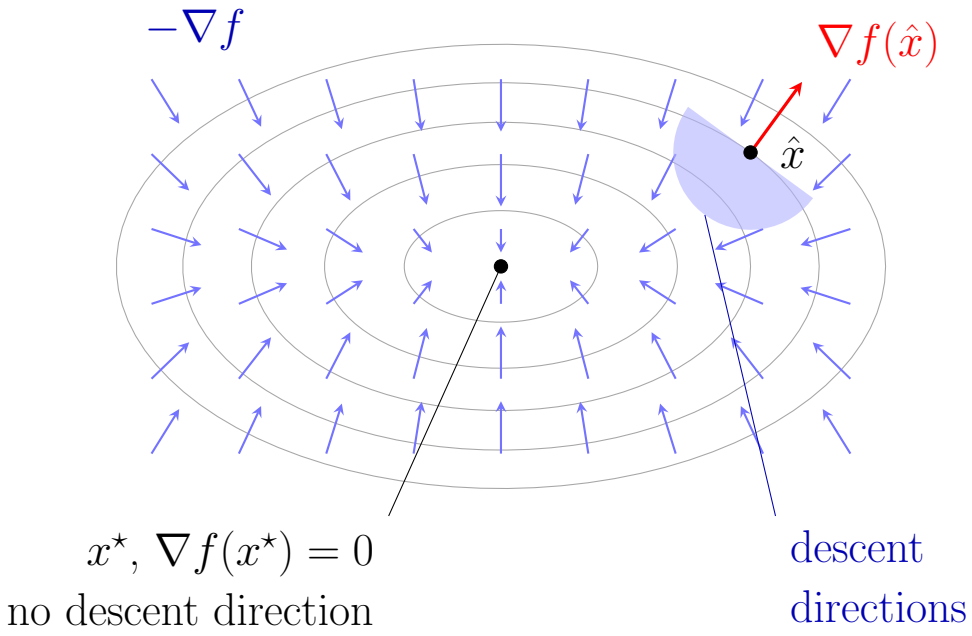
Constrained problems have exactly the same structure. The equations are called the **KKT conditions**, and they are the first-order optimality conditions for constrained problems.

First-order optimality means that no direction decreases f to first order. A direction d is a **descent direction** at x if

$$\nabla f(x)^T d < 0.$$

Then $f(x+td) = f(x)+t\nabla f(x)^T d+o(t) < f(x)$ for all small $t > 0$. So a local minimizer has no descent direction.

The gradient points uphill, across the level curve. Every direction on the downhill side of the level curve is a descent direction. Only where $\nabla f(x) = 0$ is there none.



Farkas' lemma. Exactly one of the following is feasible,

$$\begin{aligned} &\text{(P)} \quad \text{find } x \in \mathbb{R}^n \text{ s.t. } Ax \leq b, \\ &\text{(D)} \quad \text{find } y \in \mathbb{R}^m_+ \text{ s.t. } A^T y = 0,\ b^T y < 0. \end{aligned}$$

Question 2. Show that $\nabla f(x) = 0$ is precisely the Farkas certificate that no descent direction exists at x.

Solution: Scaling d, a descent direction exists if and only if (P) with $A = \nabla f(x)^T$ and $b = -1$ is feasible,

$$\nabla f(x)^T d \leq -1.$$

Its alternative (D) is

$$y \geq 0, \qquad y \nabla f(x) = 0, \qquad -y < 0,$$

which is feasible if and only if $\nabla f(x) = 0$.

2 Feasible directions

From now on, the problem is

$$\begin{aligned} \min_{x\in\mathbb{R}^n} \quad & f_0(x) \\ \text{s.t.} \quad & f_i(x) \le 0 \text{ for all } i \in \{1,\dots,m\}, \\ & h_j(x) = 0 \text{ for all } j \in \{1,\dots,p\}, \end{aligned}$$

with continuously differentiable functions and feasible set Ω. A point $x^\star \in \Omega$ is a **local minimizer** if $f_0(x^\star) \le f_0(x)$ for all $x \in \Omega$ near $x^\star$.

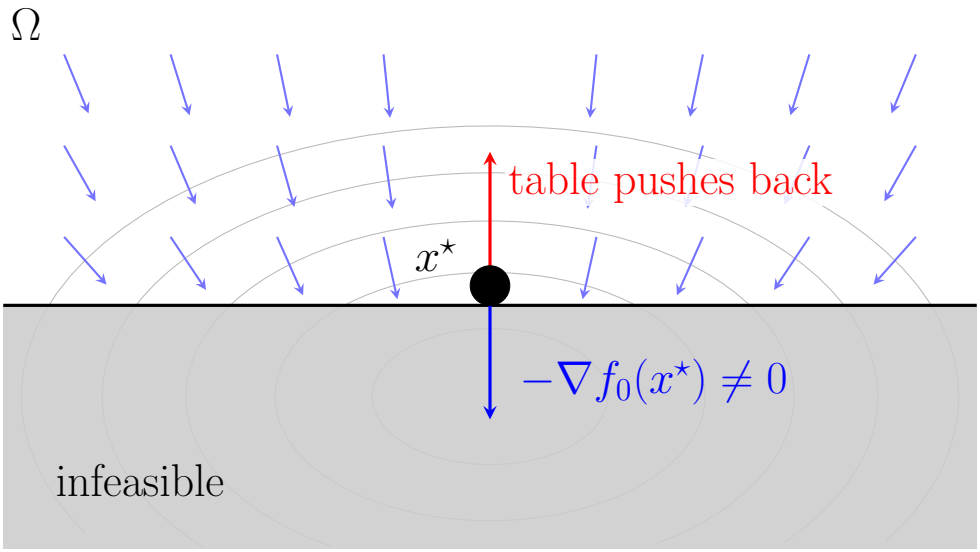
Question 3. With constraints, why is $\nabla f_0(x) = 0$ not the right condition?

Solution: For

$$\min_{x\in\mathbb{R}} \; x^2 \text{ s.t. } 1 - x \le 0,$$

$x^\star = 1$ but $\nabla f_0(1) = 2 \neq 0$. Only $d \ge 0$ stays feasible, and $\nabla f_0(1)d = 2d \ge 0$.

Think of f_0 as the height of a landscape, so that $-\nabla f_0$ is the force on a ball. The infeasible region is a table that the ball cannot enter.



The ball comes to rest on the table at the minimizer $x^\star$. There the force $-\nabla f_0(x^\star)$ is not zero. It points straight into the table, along the normal, and the table pushes back with an equal force. No direction that stays on or above the table goes downhill.

Only the directions in which we can move and stay feasible matter. We describe them with paths.

A **feasible path** from $x \in \Omega$ is a continuous map

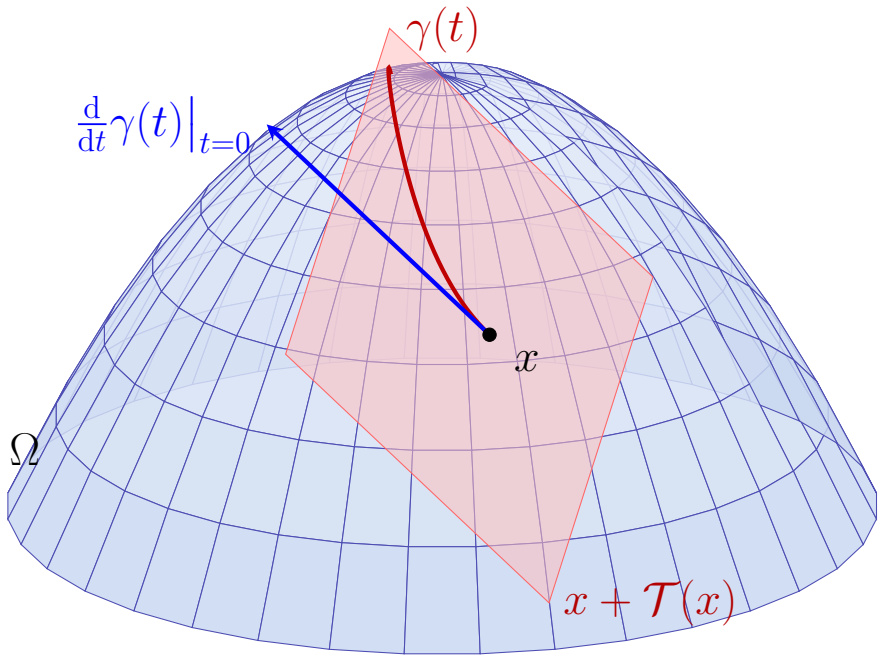
$$\gamma : [0, \epsilon) \to \Omega, \qquad \gamma(0) = x,$$

for some $\epsilon > 0$, that is differentiable at $t = 0$.

The **tangent cone** at $x \in \Omega$ is the set of initial velocities of feasible paths,

$$\mathcal{T}(x) = \left\{ \frac{\mathrm{d}}{\mathrm{d}t}\gamma(t)\Big|_{t=0} : \ \gamma \text{ is a feasible path from } x \right\}.$$

Running a path faster or slower scales its velocity, so $\mathcal{T}(x)$ is a cone.



Example 1. Circle and disc. Minimize $x_1 + x_2$ over (a) the circle and (b) the disc,

$$\text{(a)} \quad h(x) = x_1^2 + x_2^2 - 2 = 0,$$
$$\text{(b)} \quad f_1(x) = x_1^2 + x_2^2 - 2 \leq 0.$$

Find $\mathcal{T}(\hat{x})$ at $\hat{x} = (-1, -1)$ in both cases.

Solution: Let $\gamma(t)$ denote any arbitrary feasible path starting from $\hat{x}$. Then,

$$\gamma(0) = \hat{x} = (-1, -1), \qquad \dot{\gamma}(0) = d = (d_1, d_2).$$

The initial velocity $\dot{\gamma}(0) = d$ is called a *tangent vector*.

Now,

$$\begin{aligned}\gamma_1(t)^2 + \gamma_2(t)^2 &= (-1 + td_1)^2 + (-1 + td_2)^2 + o(t) \\ &= 2 - 2t(d_1 + d_2) + o(t).\end{aligned}$$

In order for γ to be feasible, it is *necessary* for the tangent vector d to satisfy, respectively

(a) Set $\gamma_1(t)^2 + \gamma_2(t)^2 = 2$ yields $d_1 + d_2 = 0$.

(b) Set $\gamma_1(t)^2 + \gamma_2(t)^2 \leq 2$ yields $d_1 + d_2 \geq 0$.

Every valid tangent vector at $\hat{x}$ must satisfy these conditions, so we have respectively

$$\text{(a)} \quad \mathcal{T}(\hat{x}) \subseteq \{d : d_1 + d_2 = 0\},$$
$$\text{(b)} \quad \mathcal{T}(\hat{x}) \subseteq \{d : d_1 + d_2 \geq 0\}.$$

To show that these are in fact all the possible tangent vectors, we just need to use them to construct a feasible path. The disk (b) is easy; consider the straight line

$$\gamma(t) = \hat{x} + t(d_1, d_2).$$

If $d_1 + d_2 \geq 0$, then

$$\begin{aligned}\gamma_1(t)^2 + \gamma_2(t)^2 &= (-1 + td_1)^2 + (-1 + td_2)^2 \\ &= 2 - 2t(d_1 + d_2) + t^2(d_1^2 + d_2^2) \leq 2\end{aligned}$$

for sufficiently small t, so the straight line is a feasible path.

The circle (a) is more tricky; consider the curved line

$$\gamma(t) = \sqrt{2}\frac{\hat{x} + td}{\|\hat{x} + td\|}.$$

One can show that

$$\gamma(t) = \hat{x} + td + O(t^2),$$

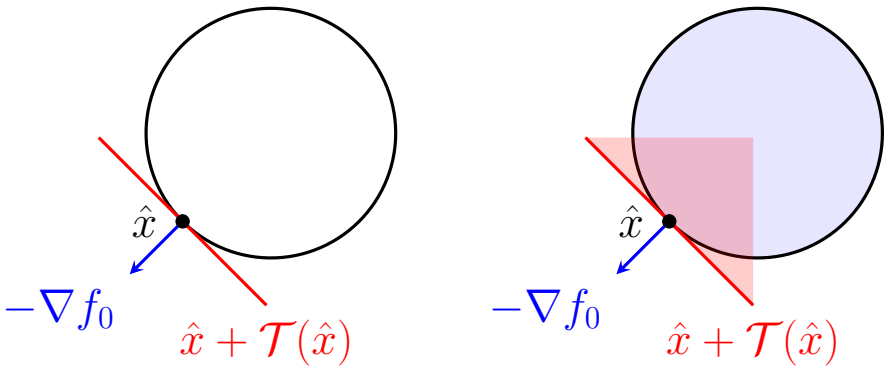
so d is indeed its tangent vector. Now, if $d_1 + d_2 = 0$, then

$$\|\gamma(t)\|^2 = \underbrace{\|\hat{x}\|^2}_{=2} + 2t\,\underbrace{d^T\hat{x}}_{=0} + t^2\|d\|^2 + O(t^3)$$

so the curved path stays on the circle, and is therefore feasible. In summary, we derived *necessary* conditions for the tangent vectors, and showed that they are also *sufficient*

$$\text{(a)} \quad \mathcal{T}(\hat{x}) = \{d : d_1 + d_2 = 0\},$$
$$\text{(b)} \quad \mathcal{T}(\hat{x}) = \{d : d_1 + d_2 \geq 0\}.$$

(a) circle (b) disc



Theorem 2. No feasible descent direction. If $x^\star$ is a local minimizer, then

$$\nabla f_0(x^\star)^T d \geq 0 \quad \text{for all } d \in \mathcal{T}(x^\star).$$

Proof: Let $\gamma(t) = x^\star + td + o(t)$ be a feasible path. For small $t > 0$, by the chain rule,

$$f_0(x^\star) \leq f_0(\gamma(t)) = f_0(x^\star) + t\nabla f_0(x^\star)^T d + o(t).$$

Subtract $f_0(x^\star)$, divide by t, and let $t \downarrow 0$,

$$\nabla f_0(x^\star)^T d \geq 0.$$

□

A point $x \in \Omega$ is **first-order optimal** if no direction in $\mathcal{T}(x)$ descends,

$$\nabla f_0(x)^T d \geq 0 \quad \text{for all } d \in \mathcal{T}(x).$$

Every local minimizer is first-order optimal. The vast majority of constrained algorithms are ultimately designed to compute a first-order optimal point.

Question 4. Are first-order optimal points also locally optimal?

Solution: Not necessarily. Saddle points and local maxima are also first-order optimal, but they are not even locally minimal. So the implication doesn't go backwards.

3 The KKT conditions

Checking first-order optimality means verifying

$$\nabla f_0(x)^T d \geq 0 \quad \text{for all } d \in \mathcal{T}(x). \tag{FOC}$$

For unconstrained optimization, $\mathcal{T}(x) = \mathbb{R}^n$, and we have proved that (FOC) is true if and only if $\nabla f_0(x) = 0$.

For general constrained optimization, checking (FOC) is coNP-complete.

Even checking *first-order* optimality is NP-hard in general.

How do we proceed? Instead, we define an *outer approximation*

$$\mathcal{F}(x) \supseteq \mathcal{T}(x)$$

called the **linearized cone** and then verify

$$\nabla f_0(x)^T d \geq 0 \quad \text{for all } d \in \mathcal{F}(x). \tag{KKT}$$

Checking (KKT) can be solved in polynomial time via Farkas' lemma.

Question 5. What is the relationship between (KKT) and (FOC)?

Solution: The outer approximation (KKT) is a *sufficient* condition for *first-order* local optimality (FOC)

$$\text{(KKT)} \implies \text{(FOC)}.$$

Why? Because every $d \in \mathcal{T}(x)$ also satisfies $d \in \mathcal{F}(x)$, so $\nabla f_0(x)^T d \geq 0$ for every $d \in \mathcal{F}(x)$ implies the same for every $d \in \mathcal{T}(x)$. But if $\mathcal{F}(x) \supset \mathcal{T}(x)$ strictly, then the inverse is false.

3.1 Definition of linearized cone $\mathcal{F}(x)$

The **active set** at $x \in \Omega$ is

$$\mathcal{A}(x) = \{i \in \{1, \dots, m\} : f_i(x) = 0\}.$$

An inactive constraint has $f_i(x) < 0$, so it still holds after any small move.

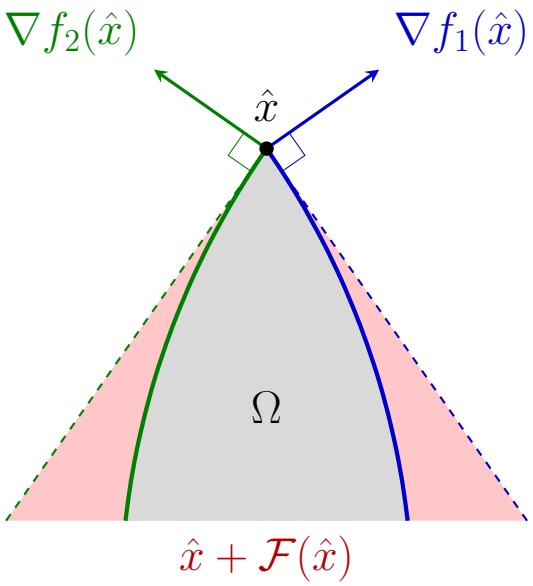
Define the **linearized cone** at $x \in \Omega$ as

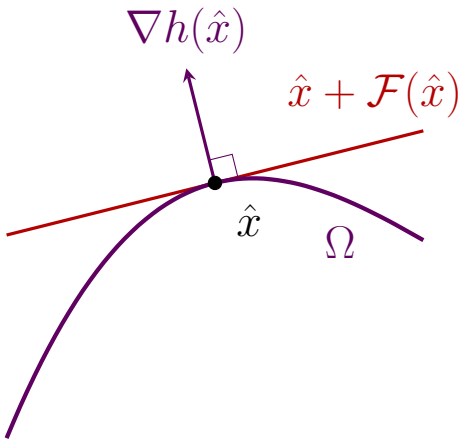
$$\mathcal{F}(x) = \left\{ d \in \mathbb{R}^n : \begin{array}{l} \nabla f_i(x)^T d \le 0 \text{ for all } i \in \mathcal{A}(x), \\ \nabla h_j(x)^T d = 0 \text{ for all } j \in \{1, \dots, p\} \end{array} \right\}.$$

What is the idea?

- Each **active** wall $f_i \le 0$ is replaced by the *half-space on the feasible side of its tangent line*, perpendicular to $\nabla f_i(x)$.
- Each **inactive** wall $f_i < 0$ is dropped.
- Each **equality** is replaced by its tangent line.

Observe that $\nabla f_i(x)^T$ and $\nabla h_j(x)^T$ are all fixed for fixed x, so these are all linear constraints; the resulting object $\mathcal{F}(x)$ is just some kind of polytope $\{d : Ad \le b, Fd = g\}$.





3.2 How to check (KKT)?

Since $\mathcal{F}(x)$ is a polytope, the existence of a descent direction is **completely defined by linear inequalities**. Farkas' lemma characterizes "if and only if" conditions when no such direction exists.

Theorem 3. KKT conditions. The following holds

$$\nabla f_0(x)^T d \geq 0 \quad \text{for all } d \in \mathcal{F}(x)$$

for a primal point x satisfying

$$\text{(primal feasibility) } f_i(x) \leq 0, \quad h_j(x) = 0 \text{ for all } i, j,$$

if and only if there exist dual multipliers $y \in \mathbb{R}^m$ and $z \in \mathbb{R}^p$ with

$$\text{(stationarity) } \nabla f_0(x) + \sum_{i=1}^{m} y_i \nabla f_i(x) + \sum_{j=1}^{p} z_j \nabla h_j(x) = 0,$$

$$\text{(dual feasibility) } y_i \geq 0 \text{ for all } i \in \{1, \ldots, m\},$$

$$\text{(complementarity) } y_i f_i(x) = 0 \text{ for all } i \in \{1, \ldots, m\}.$$

The KKT conditions are closely related to "taking the dual". Define the Lagrangian

$$\mathscr{L}(x, y, z) = f_0(x) + \sum_i y_i f_i(x) + \sum_j z_j h_j(x).$$

Then, stationarity reads $\nabla_x \mathscr{L}(x, y, z) = 0$, which *necessary* for x to solve $\inf_x \mathscr{L}(x, y, z)$, but not sufficient in general.

For convex problems, however, the link is stronger: KKT is *equivalent* to a strongly dual solution $f_0(x) = g(y, z)$, which might not exist.

Question 6. Why does z have no sign constraint?

Solution: An equality is two inequalities, $h_j(x) \leq 0$ and $-h_j(x) \leq$

0, with weights $z_j^+, z_j^- \geq 0$,

$$z_j^+ \nabla h_j(x) + z_j^- \big(-\nabla h_j(x)\big) = (z_j^+ - z_j^-)\nabla h_j(x),$$

and $z_j = z_j^+ - z_j^-$ has either sign.

The proof is a direct application of Farkas' lemma.

Proof: Scaling d, some $d \in \mathcal{F}(x)$ descends if and only if this system is feasible,

$$\begin{gathered} \nabla f_0(x)^T d \leq -1, \\ \nabla f_i(x)^T d \leq 0 \text{ for all } i \in \mathcal{A}(x), \\ \nabla h_j(x)^T d \leq 0, \quad -\nabla h_j(x)^T d \leq 0 \text{ for all } j \in \{1, \ldots, p\}. \end{gathered}$$

By Farkas' lemma, it is infeasible if and only if some $y_0, y_i, z_j^+, z_j^- \geq 0$ have $-y_0 < 0$ and

$$y_0 \nabla f_0(x) + \sum_{i \in \mathcal{A}(x)} y_i \nabla f_i(x) + \sum_{j=1}^{p} (z_j^+ - z_j^-) \nabla h_j(x) = 0.$$

Divide by $y_0 > 0$, and set $z_j = (z_j^+ - z_j^-)/y_0$ and $y_i = 0$ for $i \notin \mathcal{A}(x)$. Conversely, KKT multipliers give $y_0 = 1$. □

4 Comparing the two cones

The KKT conditions rule out descent directions in $\mathcal{F}(x)$. A local minimizer rules out descent directions in $\mathcal{T}(x)$.

Lemma 4. For every $x \in \Omega$,

$$\mathcal{T}(x) \subseteq \mathcal{F}(x).$$

Proof: Without loss of generality, suppose that every inequality f_i is active, $f_i(x) = 0$.

For a feasible path $\gamma(t) = x + td + o(t)$, the chain rule and $t \downarrow 0$ give

$$0 \geq f_i(\gamma(t)) = t\nabla f_i(x)^T d + o(t) \quad \Longrightarrow \quad \nabla f_i(x)^T d \leq 0,$$
$$0 = h_j(\gamma(t)) = t\nabla h_j(x)^T d + o(t) \quad \Longrightarrow \quad \nabla h_j(x)^T d = 0.$$

Therefore, these are *necessary* conditions for tangent vectors. □

Example 5. A cusp. Consider the feasible set

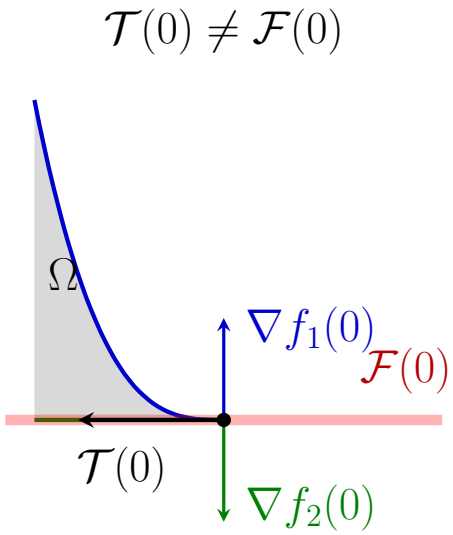
$$\Omega = \left\{x \in \mathbb{R}^2 : x_2 + x_1^3 \leq 0,\ -x_2 \leq 0\right\}.$$

One can show that $0 \in \Omega$ and that

$$\mathcal{T}(0) = \{d : d_1 \leq 0,\ d_2 = 0\},$$
$$\mathcal{F}(0) = \{d : d_2 \leq 0,\ -d_2 \leq 0\} = \{d : d_2 = 0\}.$$

Why do they differ?

Solution: At the tip of the cusp, the gradients $(0, 1)$ and $(0, -1)$ give the whole horizontal axis as $\mathcal{F}(0)$, but every feasible path moves to the left, so $\mathcal{T}(0)$ is only the left half.



The strict inclusion can cause the KKT conditions to fail altogether, even for convex problems.

Example 6. A minimizer that is not a KKT point. Consider the convex problem

$$\min_{x\in\mathbb{R}} \; x \text{ s.t. } x^2 \le 0.$$

Find the minimizer and all KKT points.

Solution: $x^2 \le 0$ only at $x = 0$, so $\Omega = \{0\}$ and $x^\star = 0$. The KKT conditions are

$$1 + 2yx = 0, \qquad x^2 \le 0, \qquad y \ge 0, \qquad yx^2 = 0.$$

Feasibility forces $x = 0$, and then stationarity reads

$$1 + 2y \cdot 0 = 1 \neq 0.$$

So there is no KKT point. The inclusion is strict,

$$\mathcal{F}(0) = \{d : 0 \cdot d \le 0\} = \mathbb{R} \supsetneq \mathcal{T}(0) = \{0\},$$

and $d = -1 \in \mathcal{F}(0)$ descends.

Question 7. How are local minimizers, first-order optimal points, and KKT points related?

Solution: For $x \in \Omega$,

$$\begin{aligned}
\text{local minimizer} &\implies \text{first-order optimal} && \text{(paths)} \\
&\iff \text{no descent direction in } \mathcal{T}(x) && \text{(definition)} \\
&\impliedby \text{no descent direction in } \mathcal{F}(x) && (\mathcal{T} \subseteq \mathcal{F}) \\
&\iff \text{KKT holds at } x && \text{(Farkas).}
\end{aligned}$$

So every KKT point is first-order optimal. The third arrow cannot be reversed when some $d \in \mathcal{F}(x) \setminus \mathcal{T}(x)$ descends.

5 Constraint qualifications

The KKT conditions are sufficient for first-order optimality, because $\mathcal{T}(x) \subseteq \mathcal{F}(x)$. For them to be necessary, the two cones must be equal. A **constraint qualification** is a condition on the constraint functions that guarantees $\mathcal{T}(x) = \mathcal{F}(x)$. Each condition below says that the system

$$\sum_{i \in \mathcal{A}(x)} y_i \nabla f_i(x) + \sum_{j=1}^{p} z_j \nabla h_j(x) = 0$$

has only the trivial solution, for some class of weights.

Theorem 7. Constraint qualifications. Let $x \in \Omega$. Then $\mathcal{T}(x) = \mathcal{F}(x)$ if any one of the following holds.
(a) **LICQ.** The only solution of the system above is $y = 0$, $z = 0$.
(b) **MFCQ.** The only solution of the system above with $y \geq 0$ is $y = 0$, $z = 0$.
(c) **Slater.** Every f_i is convex, every h_j is affine, and some $\bar{x}$ has $f_i(\bar{x}) < 0$ for all $i \in \{1, \ldots, m\}$ and $h_j(\bar{x}) = 0$ for all $j \in \{1, \ldots, p\}$.

We state this without proof. You will check these conditions on problems in the homework. N&W Lemma 12.2 proves (a), and N&W Sec. 12.6 discusses (b). For (c), see Bazaraa, Sherali, and Shetty, *Nonlinear Programming*, Ch. 5.

LICQ allows weights of any sign. MFCQ allows only the signs of KKT multipliers, so LICQ implies MFCQ. If every f_i and h_j is affine, then $\mathcal{T}(x) = \mathcal{F}(x)$ as well, since every $d \in \mathcal{F}(x)$ gives the straight feasible path $x + td$.

Question 8. Why not write an equality $h(x) = 0$ as the two

inequalities $h(x) \le 0$ and $-h(x) \le 0$?

Solution: Both are active at every feasible point, and $y_1 = y_2 = 1$ gives

$$y_1 \nabla h(x) + y_2\big(-\nabla h(x)\big) = 0.$$

So MFCQ fails everywhere, and so does LICQ.

Question 9. Let $x^\star$ be a local minimizer at which a constraint qualification holds. Why does $x^\star$ satisfy the KKT conditions?

Solution: Follow the chain

$$\begin{aligned} \text{local minimizer} &\implies \text{no descent direction in } \mathcal{T}(x^\star) && \text{(paths)} \\ &\iff \text{no descent direction in } \mathcal{F}(x^\star) && (\mathcal{T} = \mathcal{F}) \\ &\iff \text{KKT holds at } x^\star && \text{(Farkas).} \end{aligned}$$

The middle step is the one that needs the constraint qualification.

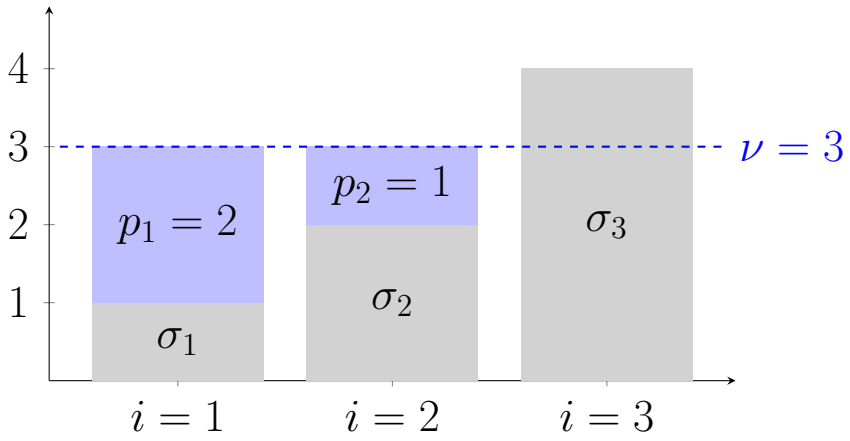
6 Worked Example

Power allocation. WiFi, 4G and 5G cellular, and DSL split one link into parallel subchannels. With power $p_i \geq 0$, subchannel i carries $\log(1+p_i/\sigma_i)$ units of information per use, where $\sigma_i > 0$ is its noise level. Since $\log(1+p_i/\sigma_i) = \log(\sigma_i + p_i) - \log \sigma_i$, maximizing the total rate with a power budget P is a convex problem.

Example 8. Power allocation. Solve the power allocation problem

$$\begin{aligned} \min_{p \in \mathbb{R}^n} \quad & -\sum_{i=1}^{n} \log(\sigma_i + p_i) \\ \text{s.t.} \quad & -p_i \leq 0 \text{ for all } i \in \{1, \ldots, n\}, \\ & p_1 + \cdots + p_n - P = 0 \end{aligned}$$

by KKT for $\sigma = (1, 2, 4)$ and $P = 3$.



Each subchannel in use has $\sigma_i + p_i = 1/z = 3$, a common water level ν. The unused subchannel has its floor $\sigma_3 = 4$ above the level. In general, $p_i = \max\{\nu - \sigma_i, 0\}$, with ν set by the budget. This is **water-filling**.

Solution: Introduce multipliers $y_i \geq 0$ for $p_i \geq 0$ and $z \in \mathbb{R}$ for the budget constraint. The KKT conditions are

$$-\frac{1}{\sigma_i + p_i} - y_i + z = 0,$$
$$y_i p_i = 0, \qquad y_i \geq 0, \qquad p_i \geq 0,$$
$$p_1 + p_2 + p_3 = 3.$$

Hence all indices with $p_i > 0$ satisfy $\sigma_i + p_i = 1/z$, while those with $p_i = 0$ satisfy $\sigma_i \geq 1/z$.

The largest σ_i should receive zero power first. Indeed, if all three $p_i > 0$, their common value would be

$$\frac{1}{z} = \frac{1+2+4+3}{3} = \frac{10}{3} < 4,$$

forcing $p_3 < 0$. We therefore try $p_3 = 0$ and $p_1, p_2 > 0$. The KKT conditions then reduce to

$$1 + p_1 = 2 + p_2, \qquad p_1 + p_2 = 3,$$

which immediately yields

$$p = (2, 1, 0).$$

The multipliers are $z = 1/3$, $y_1 = y_2 = 0$, and $y_3 = 1/3 - 1/4 = 1/12 \geq 0$, confirming that all KKT conditions hold.

7 Summary

First-order optimality. A local minimizer has no descent direction. Without constraints, $\nabla f(x) = 0$ is exactly the Farkas certificate of this fact.

Two cones. With constraints, a local minimizer has no descent direction in the tangent cone $\mathcal{T}(x^\star)$, the set of velocities of feasible paths. The linearized cone $\mathcal{F}(x)$ replaces each active constraint by the half-space behind its tangent line.

KKT conditions. By Farkas' lemma, multipliers $y \geq 0$ and z with

$$\nabla f_0 + \sum_i y_i \nabla f_i + \sum_j z_j \nabla h_j = 0, \qquad y_i f_i = 0$$

exist at x exactly when no direction in $\mathcal{F}(x)$ descends.

Sufficiency. Always $\mathcal{T}(x) \subseteq \mathcal{F}(x)$. So every KKT point is first-order optimal, and for a convex problem it is a global minimizer.

Necessity. The inclusion can be strict, and then a minimizer can fail to be a KKT point, even for the convex problem min x s.t. $x^2 \leq 0$. Where $\mathcal{T}(x^\star) = \mathcal{F}(x^\star)$, every local minimizer is a KKT point. LICQ, MFCQ, and Slater's condition each guarantee this.

Computing. Solvers compute KKT points. For a nonconvex problem, list the KKT points and compare their values, as for $\nabla f = 0$.