Lecture 14. Lagrange Condition

14.1 Equality constrained optimization

We now consider equality constrained optimization problems.

$$\begin{aligned} \min_{\boldsymbol{x}\in\mathbb{R}^n} \quad & f(\boldsymbol{x}) && \text{(convex)} \\ \text{subject to} \quad & \boldsymbol{a}_i^\top \boldsymbol{x} - b_i = 0,\ \forall\, i && \text{(affine)} \end{aligned}$$

In a more compact form, it can be expressed as

$$\begin{aligned} \min_{\boldsymbol{x}\in\mathbb{R}^n} \quad & f(\boldsymbol{x}) && \text{(convex)} \\ \text{subject to} \quad & g(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x} - \boldsymbol{b} = \boldsymbol{0} && \text{(affine)} \end{aligned}$$

where $\boldsymbol{A} = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_m]^\top \in \mathbb{R}^{m\times n}$ and $\boldsymbol{b} = (b_1, \ldots, b_m) \in \mathbb{R}^m$.

The *feasible set* of this problem is

$$\Omega = \{\boldsymbol{x} \in \mathbb{R}^n \mid \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}\}\,.$$

Recall that, such a set is an affine set. Given any $\boldsymbol{x}_0 \in \Omega$, we find that

$$\forall\, \boldsymbol{x} \in \Omega, \qquad \boldsymbol{A}(\boldsymbol{x} - \boldsymbol{x}_0) = \boldsymbol{A}\boldsymbol{x} - \boldsymbol{A}\boldsymbol{x}_0 = \boldsymbol{0}\,.$$

Conversely, if $\boldsymbol{A}\boldsymbol{v} = \boldsymbol{0}$, then $\boldsymbol{x} = \boldsymbol{x}_0 + \boldsymbol{v}$ satisfies $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ and thus is in Ω. So $\Omega - \boldsymbol{x}_0$ is the *kernel* (or the *null space*) of $\boldsymbol{A}$. Or equivalently,

$$\Omega = \boldsymbol{x}_0 + \ker(\boldsymbol{A})\,.$$

14.2 Lagrange multiplier method for convex optimization

Question

How to verify that a solution is optimal?

We may rewrite the optimization problem

$$\begin{aligned} \min_{\boldsymbol{x}\in\mathbb{R}^n} \quad & f(\boldsymbol{x}) \\ \text{subject to} \quad & g(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x} - \boldsymbol{b} = \mathbf{0} \end{aligned}$$

as

$$\min_{\boldsymbol{x}\in\Omega} \quad f(\boldsymbol{x})$$

where $\Omega = \{\boldsymbol{x} \in \mathbb{R}^n \mid \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}\}$ is the feasible set.

Recall that, in Lecture 2, we've showed that if x^* is a local minimum point of $f : D \subseteq \mathbb{R}^n \to \mathbb{R}$, then for any feasible direction v (i.e. $\exists\, \varepsilon > 0$ such that $x^* + \delta v \in D$ for any $0 < \delta < \varepsilon$),

$$\nabla_v f(x^*) = \langle \nabla f(x^*),\, v\rangle \geq 0\,.$$

If f is a convex function, D is a convex set. Thus for any $x \in D$, $x - x^*$ is a feasible direction. Then we have $\langle \nabla f(x^*), x - x^*\rangle \geq 0$ for any $x \in D$. Conversely, if $\langle \nabla f(x^*), x - x^*\rangle \geq 0$, then by convexity we have

$$f(x) \geq f(x^*) + \langle \nabla f(x^*), x - x^*\rangle \geq f(x^*)\,,$$

which implies that x^* is a global minimum point. Therefore, we have the following optimality theorem.

Theorem

Suppose $f : D \subseteq \mathbb{R}^n \to \mathbb{R}$ is a convex function. Then x^* is a *global* minimum point of f iff

$$\forall\, x \in D, \quad \langle \nabla f(x^*),\, x - x^*\rangle \geq 0\,.$$

In particular, if $D = \mathbb{R}^n$ or D is open, then x^* is a global minimum point iff $\nabla f(x^*) = \mathbf{0}$.

For unconstrained optimization problems, we often assume the domain $D = \mathbb{R}^n$, or D is an open set. So we can apply the criterion $\nabla f(\boldsymbol{x}^*) = \mathbf{0}$. Now, since we have constraints, the domain should be the feasible set Ω, which is probably not open.

However, the feasible set Ω is an affine set, which means that if $\boldsymbol{x}^* + \boldsymbol{v} \in \Omega$, then $\boldsymbol{x}^* - \boldsymbol{v} \in \Omega$. Thus, $\langle \nabla f(\boldsymbol{x}^*),\, \boldsymbol{x} - \boldsymbol{x}^*\rangle \geq 0$ for all $\boldsymbol{x} \in \Omega$ is equivalent to $\langle \nabla f(\boldsymbol{x}^*),\, \boldsymbol{x} - \boldsymbol{x}^*\rangle = 0$.

Note that for all i, $\boldsymbol{a}_i^\top \boldsymbol{x} = 0$ (or in a compact form, $\boldsymbol{A}(\boldsymbol{x} - \boldsymbol{x}^*) = \boldsymbol{0}$, where $\boldsymbol{A} = [\boldsymbol{a}_1, \dots, \boldsymbol{a}_m]^\top$). So if there exists $\lambda_1, \dots, \lambda_m \in \mathbb{R}$ such that $\nabla f(\boldsymbol{x}^*) + \lambda_1 \boldsymbol{a}_1 + \cdots + \lambda_m \boldsymbol{a}_m = \boldsymbol{0}$ (namely, $\nabla f(\boldsymbol{x}^*) \in \operatorname{span}\{\boldsymbol{a}_1, \dots, \boldsymbol{a}_m\}$), clearly we have $\langle \nabla f(\boldsymbol{x}^*), \boldsymbol{x} - \boldsymbol{x}^* \rangle = 0$ for all $\boldsymbol{x} \in \Omega$.

Conversely, if $\langle \nabla f(\boldsymbol{x}^*), \boldsymbol{v} \rangle = 0$ for all $\boldsymbol{v} \in \ker(\boldsymbol{A})$, it holds that $\nabla f(\boldsymbol{x}^*)$ is a linear combination of $\boldsymbol{a}_1, \dots, \boldsymbol{a}_m$ (or equivalently, there exists $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_m)^\top \in \mathbb{R}^m$ such that $\nabla f(\boldsymbol{x}^*) + \boldsymbol{A}^\top \boldsymbol{\lambda} = \boldsymbol{0}$). There are multiple ways to prove this proposition. Suppose for the sake of contradiction $\nabla f(\boldsymbol{x}^*)$ is not a linear combination of $\boldsymbol{a}_1, \dots, \boldsymbol{a}_m$. Then consider the *orthogonal decomposition*

$$\nabla f(\boldsymbol{x}^*) = \boldsymbol{u} + \boldsymbol{v}$$

where $\boldsymbol{u} \in \operatorname{span}\{\boldsymbol{a}_1, \dots, \boldsymbol{a}_m\}$ and $\boldsymbol{0} \neq \boldsymbol{v} \perp \operatorname{span}\{\boldsymbol{a}_1, \dots, \boldsymbol{a}_m\}$. By definition, $\boldsymbol{v} \in \ker(\boldsymbol{A})$. But

$$\langle \nabla f(\boldsymbol{x}^*), \boldsymbol{v} \rangle = \langle \boldsymbol{v}, \boldsymbol{v} \rangle > 0\,,$$

which contradicts to our assumption $\langle \nabla f(\boldsymbol{x}^*), \boldsymbol{v} \rangle = 0$ for all $\boldsymbol{v} \in \ker(\boldsymbol{A})$.

An alternate proof is to use the *rank-nullity theorem*: $\dim \ker(\boldsymbol{A}) + \operatorname{rank}(\boldsymbol{A}) = n$.

Lemma

Given a matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, $\operatorname{im}(\boldsymbol{A}^\top) = \operatorname{range}\{\boldsymbol{A}^\top\} \triangleq \{\boldsymbol{A}^\top \boldsymbol{v} \mid \boldsymbol{v} \in \mathbb{R}^m\}$ is the *orthogonal complement* of $\ker(\boldsymbol{A})$, denoted by $\operatorname{im}(\boldsymbol{A}^\top) = \ker(\boldsymbol{A})^\perp$.

Proof

We argue that $\operatorname{im}(\boldsymbol{A}^\top) \subseteq \ker(\boldsymbol{A})^\perp$ is a subspace with the same dimension, thus $\operatorname{im}(\boldsymbol{A}^\top) = \ker(\boldsymbol{A})^\perp$. Note that for all $\boldsymbol{x} \in \operatorname{im}(\boldsymbol{A}^\top)$, there exists $\boldsymbol{v} \in \mathbb{R}^m$ such that $\boldsymbol{x} = \boldsymbol{A}^\top \boldsymbol{v}$. For all $\boldsymbol{y} \in \ker(\boldsymbol{A})$, $\boldsymbol{A}\boldsymbol{y} = \boldsymbol{0}$. So

$$\boldsymbol{x}^\top \boldsymbol{y} = (\boldsymbol{A}^\top \boldsymbol{v})^\top \boldsymbol{y} = \boldsymbol{v}^\top \boldsymbol{A} \boldsymbol{y} = 0\,,$$

which gives that $\boldsymbol{x} \in \ker(\boldsymbol{A})^\perp$. Therefore $\operatorname{im}(\boldsymbol{A}^\top) \subseteq \ker(\boldsymbol{A})^\perp$. Moreover, we have $\dim \operatorname{im}(\boldsymbol{A}^\top) = \operatorname{rank}(\boldsymbol{A}) = n - \dim \ker(\boldsymbol{A}) = \dim \ker(\boldsymbol{A})^\perp$ by the rank-nullity theorem, which concludes that $\operatorname{im}(\boldsymbol{A}^\top) = \ker(\boldsymbol{A})^\perp$.

By this lemma we know that $\nabla f(\boldsymbol{x}^*) \in \operatorname{im}(\boldsymbol{A}^\top)$ since $\nabla f(\boldsymbol{x}^*) \perp \ker(\boldsymbol{A})$.

The third way is to apply the *Farkas' lemma*.

Theorem (*Farkas' lemma*)

Let $\boldsymbol{A} \in \mathbb{R}^{n\times m}, \boldsymbol{b} \in \mathbb{R}^n$. Then exactly one of the following sets is empty:

1. $\{\boldsymbol{x} \in \mathbb{R}^m \mid \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}, \boldsymbol{x} \geq \boldsymbol{0}\}$;
2. $\{\boldsymbol{y} \in \mathbb{R}^n \mid \boldsymbol{A}^\top \boldsymbol{y} \leq \boldsymbol{0}, \boldsymbol{b}^\top \boldsymbol{y} > 0\}$.

Recall that our goal is to show that $\nabla f(\boldsymbol{x}^*) \in \operatorname{span}(\boldsymbol{a}_1, \dots, \boldsymbol{a}_m)$ where $\langle \nabla f(\boldsymbol{x}^*), \boldsymbol{v}\rangle = 0$ for all $\boldsymbol{v} \in \ker(\boldsymbol{A})$. Let $\boldsymbol{A}' = (\boldsymbol{A}^\top \quad -\boldsymbol{A}^\top)$. Then $\boldsymbol{v} \in \ker(\boldsymbol{A})$ is equivalent to

$$(\boldsymbol{A}')^\top \boldsymbol{v} = \begin{pmatrix} \boldsymbol{A} \\ -\boldsymbol{A} \end{pmatrix} \boldsymbol{v} \leq \boldsymbol{0}\,.$$

Applying Farkas' lemma on $\boldsymbol{A}'$ and $\nabla f(\boldsymbol{x}^*)$, we know that item 2 is not satisfiable, since for all $\boldsymbol{v} \in \mathbb{R}^n$, if $(\boldsymbol{A}')^\top \boldsymbol{v} \leq \boldsymbol{0}$, then $\nabla f(\boldsymbol{x}^*)^\top \boldsymbol{v} = 0$. Thus, there exists $\boldsymbol{u} \in \mathbb{R}^{2m}$ such that $\boldsymbol{A}'\boldsymbol{u} = \nabla f(\boldsymbol{x}^*)$ and $\boldsymbol{u} \geq \boldsymbol{0}$. Let $\boldsymbol{\lambda} \in \mathbb{R}^m$ where $\lambda_i = u_i - u_{i+m}$ we have $\boldsymbol{A}^\top \boldsymbol{\lambda} = \nabla f(\boldsymbol{x}^*)$.

Overall, we obtain the following *Lagrange condition*. The method to find the global minimum point by this condition is called the *Lagrange multiplier method*.

Theorem (*Lagrange condition for convex optimization problems*)

Consider a convex optimization problem

$$\begin{aligned} \min_{\boldsymbol{x} \in \mathbb{R}^n} \quad & f(\boldsymbol{x}) \\ \text{subject to} \quad & g(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x} - \boldsymbol{b} = \boldsymbol{0}\,. \end{aligned}$$

Then $\boldsymbol{x}^*$ is a *global minimum point* iff $\boldsymbol{x}^* \in \Omega$ and there exists $\boldsymbol{\lambda}^* \in \mathbb{R}^m$ such that

$$\nabla f(\boldsymbol{x}^*) + \boldsymbol{A}^\top \boldsymbol{\lambda}^* = \boldsymbol{0}\,,$$

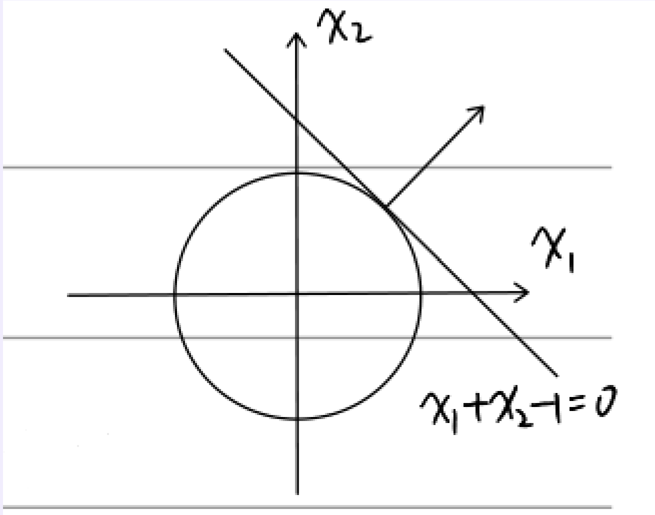
or equivalently,

$$\nabla f(\boldsymbol{x}^*) + \lambda_1^* \boldsymbol{a}_1 + \cdots + \lambda_m^* \boldsymbol{a}_m = \boldsymbol{0}\,,$$

where $\boldsymbol{A} = [\boldsymbol{a}_1, \dots, \boldsymbol{a}_m]^\top$ and $\boldsymbol{\lambda}^* = (\lambda_1^*, \dots, \lambda_m^*)^\top$.
The constants $\lambda_1^*, \dots, \lambda_m^*$ are called *Lagrange multipliers*.

Example

$$\begin{aligned} \min \quad & f(x_1, x_2) = x_1^2 + x_2^2 \\ \text{subject to} \quad & g(x_1, x_2) = x_1 + x_2 - 1 = 0 \end{aligned}$$

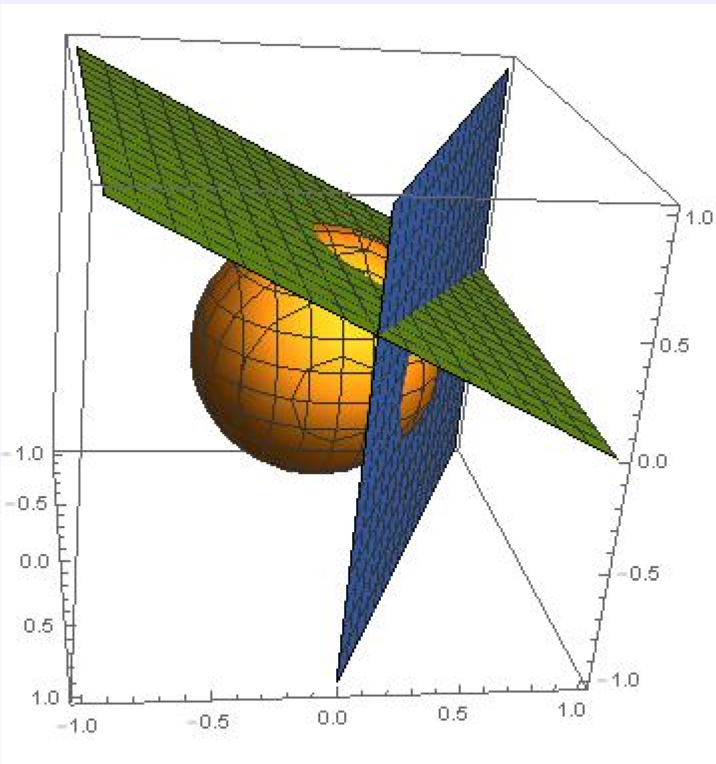


The optimal solution is $x_1 = \frac{1}{2}, x_2 = \frac{1}{2}$, where $\nabla f(x_1, x_2) = (1, 1)^\mathsf{T}$. The Lagrange condition has a geometric interpretation: the *level set* of f should be *tangent* to the feasible set at the minimum point.

Here is an example in high dimensions.

Example

$$\begin{aligned} \min \quad & f(x_1, x_2, x_3) = x_1^2 + x_2^2 + x_3^2 \\ \text{subject to} \quad & g_1(x_1, x_2, x_3) = x_1 + 2x_2 - 1 = 0 \\ & g_2(x_1, x_2, x_3) = x_2 + 2x_3 - 1 = 0 \end{aligned}$$



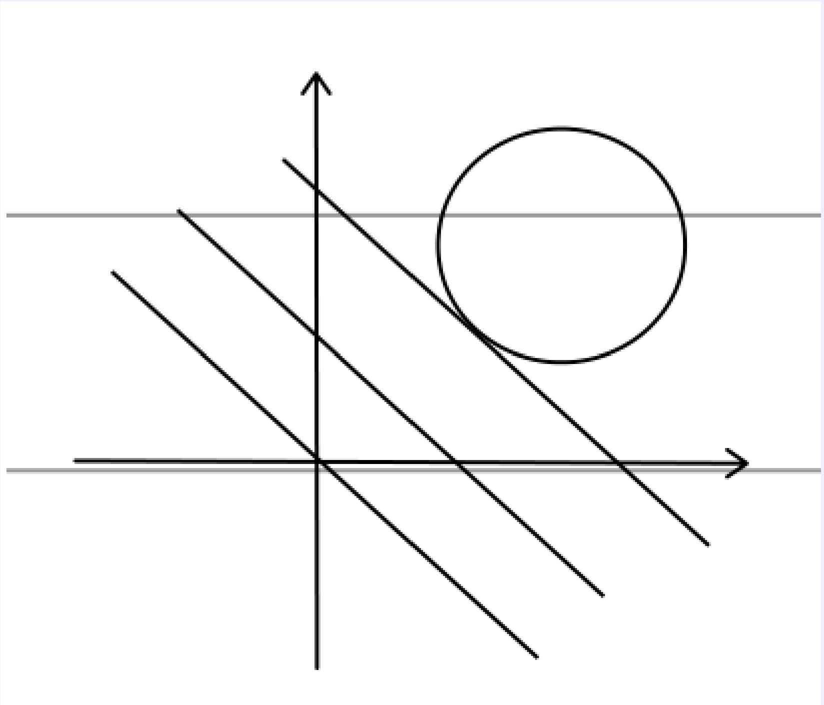
The optimal solution is $x_1 = \frac{1}{7}$, $x_2 = \frac{3}{7}$, $x_3 = \frac{2}{7}$, where the level set $\{\boldsymbol{x} \in \mathbb{R}^3 \mid f(\boldsymbol{x}) = \frac{2}{7}\}$ is tangent to the feasible set.

14.3 Lagrange multiplier method for general cases

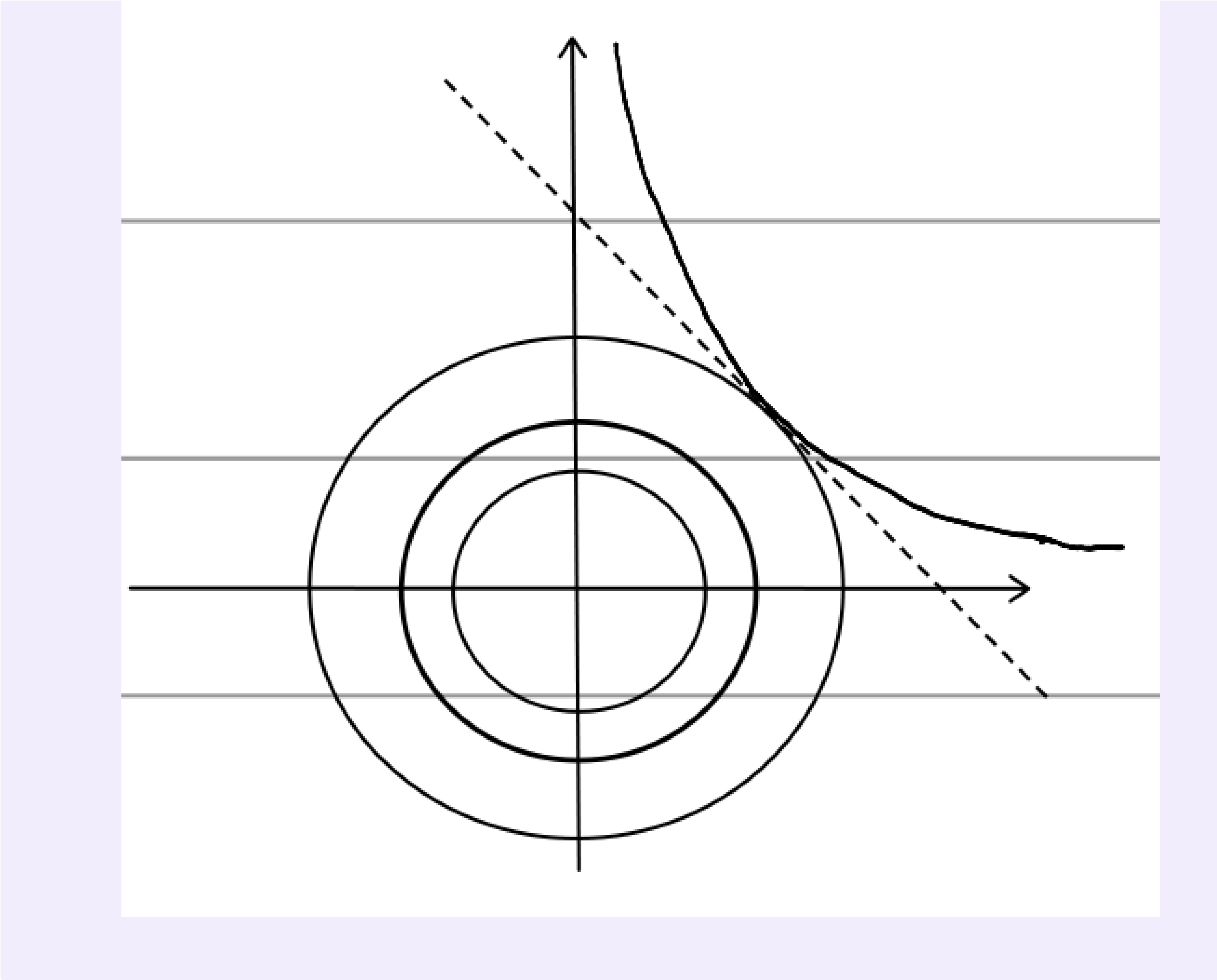
It seems that the geometric observation (the level set and the feasible set should be tangent) still holds in general cases. We now introduce some examples.

Example

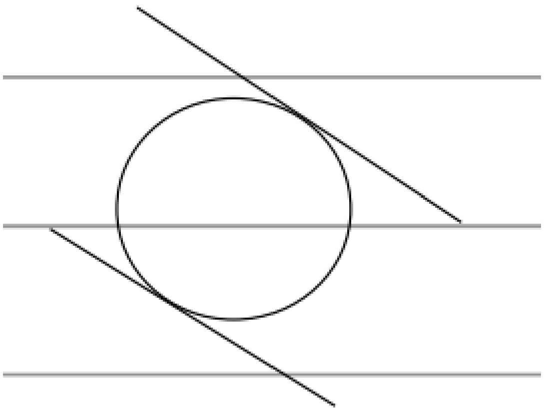
1. $$\begin{aligned} \min \quad & f(x_1, x_2) = x_1 + x_2 \\ \text{subject to} \quad & g(x_1, x_2) = (x_1 - 2)^2 + (x_2 - 1)^2 - 1 = 0 \end{aligned}$$



2. $$\begin{aligned} \min \quad & f(x_1, x_2) = x_1^2 + x_2^2 \\ \text{subject to} \quad & g(x_1, x_2) = x_1 x_2 - 1 = 0 \end{aligned}$$



Both examples look good! However, note that in Example 1, there exists two parallel lines $\{(x_1, x_2)^\top \mid x_1 + x_2 = c\}$ tangent to the level set $\{(x_1, x_2)^\top \mid g(x_1, x_2) = 0\}$.

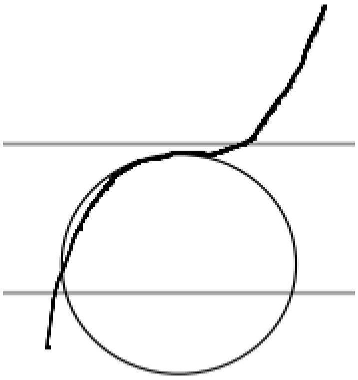


So we know that the geometric observation (the level set and the feasible set should be tangent) is no longer a sufficient condition for minimum points. In fact, it is even not a sufficient condition for optimality. Consider the following example

$$\begin{array}{rl} \min & f(x_1, x_2) = x_1^2 + x_2^2 \\ \text{subject to} & g(x_1, x_2) = x_1^3 - x_2 + 1 = 0 \end{array}$$

Two curves $\{\boldsymbol{x} \in \mathbb{R}^2 \mid f(\boldsymbol{x}) = 1\}$ and $\{\boldsymbol{x} \in \mathbb{R}^2 \mid g(\boldsymbol{x}) = 0\}$ are tangent at $(0, 1)$.

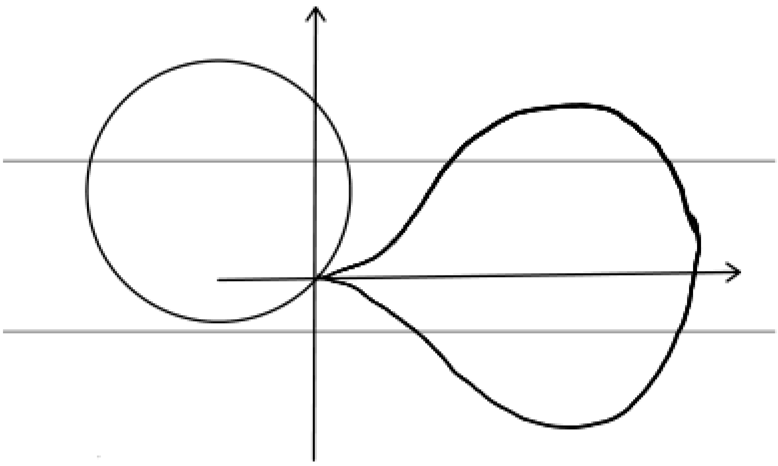
Unfortunately, this is neither a minimum point nor a maximum point.



We may guess that the condition is necessary for optimality, although it is not sufficient. However, consider the following example

$$\begin{aligned} \min \quad & (x_1+1)^2 + (x_2-1)^2 \\ \text{subject to} \quad & g(x_1,x_2) = (x_1^2+x_2^2)^2 - 2x_1(x_1^2+x_2^2) + 3x_2^2 = 0\,. \end{aligned}$$

The optimal solution is $x^* = (0,0)$, where two curves are not tangent.



So what can we say about the minimum points? We first need to clarify the concept of "tangency", and then we try to give a necessary condition for optimality and rule out the final example. We now introduce the following *Lagrange condition for general cases*.

Definition (*Regular point and critical point*)

Suppose $\boldsymbol{g}(\boldsymbol{x}) = (g_1(\boldsymbol{x}),\ldots,g_m(\boldsymbol{x}))^\mathsf{T}$ is a function $\mathbb{R}^n \to \mathbb{R}^m$ where $m \le n$. A point $\boldsymbol{x}_0$ is a *regular point* of $\boldsymbol{g}$ if

$$\boldsymbol{g}'(\boldsymbol{x}_0) = \begin{pmatrix} \nabla g_1(\boldsymbol{x}_0)^\mathsf{T} \\ \vdots \\ g_m(\boldsymbol{x}_0)^\mathsf{T} \end{pmatrix}$$

has full (row) rank m. Or equivalently, $\nabla g_1(\boldsymbol{x}_0),\ldots,\nabla g_m(\boldsymbol{x}_0)$ are linearly

independent.
Otherwise, $\boldsymbol{x}_0$ is called a *critical point*.

Theorem (*Lagrange condition for general optimization problems*)

Consider a general optimization problem

$$\begin{aligned} \min_{\boldsymbol{x}\in\mathbb{R}^n} \quad & f(\boldsymbol{x}) \\ \text{subject to} \quad & g_i(\boldsymbol{x}) = 0\,, \quad \forall\, i = 1, 2, \ldots, m\,. \end{aligned}$$

If $\boldsymbol{x}^*$ is a *local minimum* in the feasible set, and $\boldsymbol{x}^*$ is a *regular point* of $\boldsymbol{g} = (g_1, \ldots, g_m)^\top$, then there exists $\boldsymbol{\lambda}^* = (\lambda_1^*, \ldots, \lambda_m^*)^\top \in \mathbb{R}^m$ such that

$$\nabla f(\boldsymbol{x}^*) + \lambda_1^* \nabla g_1(\boldsymbol{x}^*) + \cdots + \lambda_m^* \nabla g_m(\boldsymbol{x}^*) = \boldsymbol{0}\,.$$

The constants $\lambda_1^*, \ldots, \lambda_m^*$ are called *Lagrange multipliers*.

Example

We can give an alternate proof of the *Rayleigh quotient* by Lagrange condition. Given a symmetric $\boldsymbol{Q} \in \mathbb{R}^{n\times n}$, consider the following optimizations

$$\begin{aligned} \max / \min \quad & \boldsymbol{v}^\top \boldsymbol{Q} \boldsymbol{v} \\ \text{subject to} \quad & \boldsymbol{v}^\top \boldsymbol{v} = 1\,. \end{aligned}$$

This problem is to find extreme values of a continuous function on a compact set. So both maximum and minimum exist.
For a maximum (or minimum) solution $\boldsymbol{v}^*$, clearly $\boldsymbol{v}^*$ is regular. By Lagrange multiplier condition, there exists $\lambda^* \in \mathbb{R}$ such that
$(\nabla \boldsymbol{v}^\top \boldsymbol{Q} \boldsymbol{v})\big|_{\boldsymbol{v}=\boldsymbol{v}^*} + \lambda^* (\nabla \boldsymbol{v}^\top \boldsymbol{v})\big|_{\boldsymbol{v}=\boldsymbol{v}^*} = \boldsymbol{0}$. Namely,

$$\boldsymbol{Q}\boldsymbol{v}^* + \lambda \boldsymbol{v}^* = \boldsymbol{0}\,.$$

It implies that $\boldsymbol{v}^*$ is an eigenvector of $\boldsymbol{Q}$. Thus,

$$\max_{\boldsymbol{v}^\top\boldsymbol{v}=1} \boldsymbol{v}^\top \boldsymbol{Q} \boldsymbol{v} = \lambda_{\max}(\boldsymbol{Q})\,, \qquad \min_{\boldsymbol{v}^\top\boldsymbol{v}=1} \boldsymbol{v}^\top \boldsymbol{Q} \boldsymbol{v} = \lambda_{\min}(\boldsymbol{Q})\,.$$

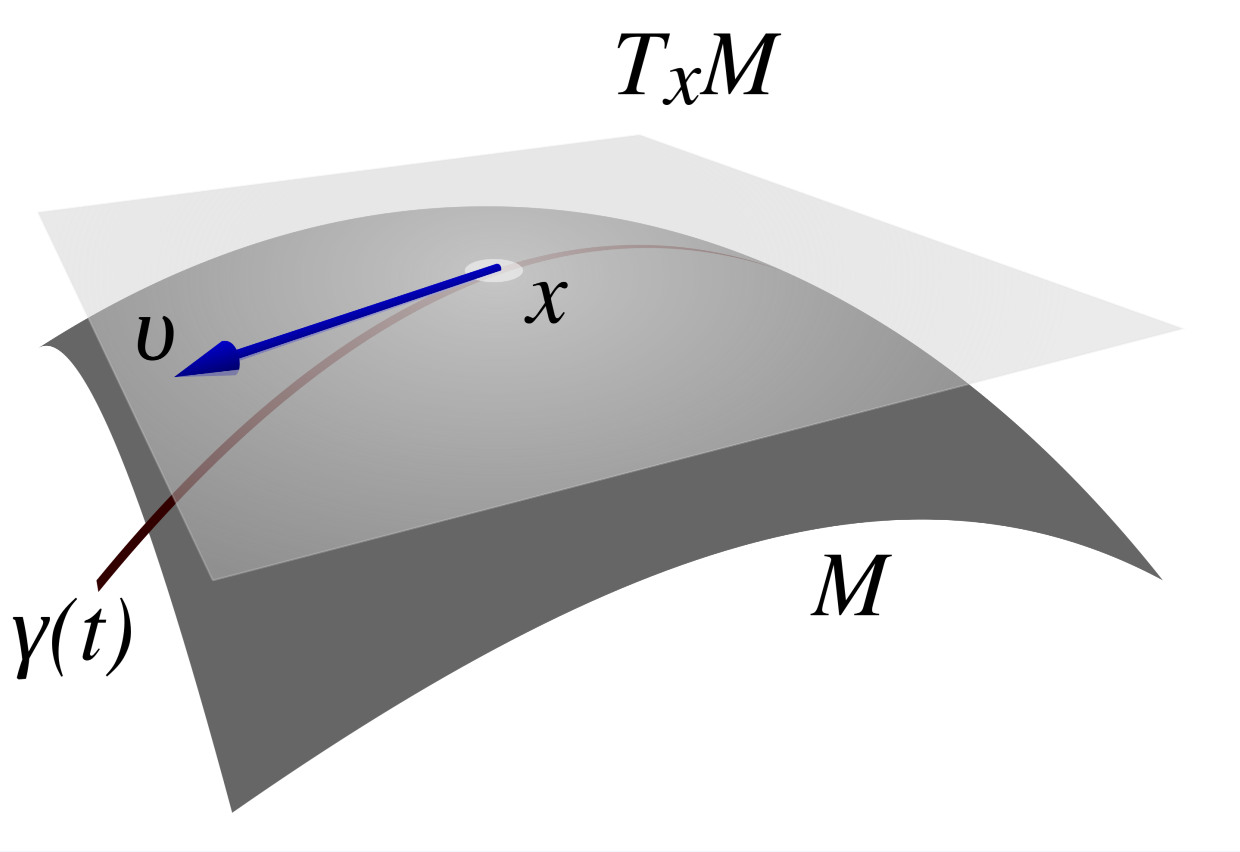
14.4 Tangent space and normal space

Recall the definition of "tangent lines". Given a differentiable curve $y = f(x)$ in $\mathbb{R}^2$, its tangent line at $(x_0, f(x_0))$ is $y = f'(x_0)(x - x_0) + f(x_0)$. In other words, a tangent line is a line passing through $(x_0, f(x_0))$ with slope $f'(x_0)$. We now generalize this concept to $\mathbb{R}^n$.

Definition (*Tangent vector and tangent space*)

Given a differentiable curve $\boldsymbol{\gamma}(t) : \mathbb{R} \to \mathbb{R}^n$, the *tangent vector* of $\boldsymbol{\gamma}$ at $\boldsymbol{x}_0 = \boldsymbol{\gamma}(t_0)$ is $\boldsymbol{\gamma}'(t_0)$.
Given a differentiable surface (*differentiable submanifold*) M, its *tangent space* at some $\boldsymbol{x}_0 \in M$, denoted by $T_{\boldsymbol{x}_0}M$, is the set of *tangent vectors* at $\boldsymbol{x}_0$ of all differentiable curves on M and passing through $\boldsymbol{x}_0$. Namely,

$$T_{\boldsymbol{x}_0}M \triangleq \{\boldsymbol{\gamma}'(0) \mid \boldsymbol{\gamma} : (-\varepsilon, \varepsilon) \to M, \boldsymbol{\gamma}(0) = \boldsymbol{x}_0\}.$$



An important (but probably surprising) fact is that the tangent space is a linear space. As an example, we show that the tangent space of $\Omega = \{\boldsymbol{x} \mid \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}\}$ is $\ker(\boldsymbol{A})$. Fix a point $\boldsymbol{x}_0 \in \Omega$. For any differentiable $\boldsymbol{\gamma}(t) : (-\varepsilon, \varepsilon) \to \Omega$ where $\boldsymbol{\gamma}(0) = \boldsymbol{x}_0$, we have

$$\boldsymbol{A}\boldsymbol{\gamma}'(0) = \lim_{\delta \to 0} \frac{\boldsymbol{A}(\boldsymbol{\gamma}(\delta) - \boldsymbol{\gamma}(0))}{\delta} = \mathbf{0}.$$

(An alternate proof is by considering the derivative of $f(t) = \boldsymbol{A}\boldsymbol{\gamma}(t)$, and noting that $\mathbf{0} = f'(t) = \boldsymbol{A}\boldsymbol{\gamma}'(t)$.) Thus it holds that $\boldsymbol{\gamma}(0) \in \ker(\boldsymbol{A})$. Conversely, for any $\boldsymbol{v} \in \ker(\boldsymbol{A})$, the curve $\boldsymbol{\gamma}(t) = \boldsymbol{x}_0 + t\boldsymbol{v}$ has derivative $\boldsymbol{\gamma}'(0) = \boldsymbol{v}$. So we have $T_{\boldsymbol{x}_0}\Omega = \ker(\boldsymbol{A})$.

The *normal space* is the orthogonal complement of the tangent space.

Definition (*Normal vector and normal space*)

A vector $\boldsymbol{v}$ is a *normal vector* of M at $\boldsymbol{x}_0$, if $\boldsymbol{v}$ is orthogonal to the tangent space $T_{\boldsymbol{x}_0}M$, i.e., $\langle \boldsymbol{u}, \boldsymbol{v} \rangle = 0$ for all $\boldsymbol{u} \in T_{\boldsymbol{x}_0}M$.
The *normal space* of M at $\boldsymbol{x}_0$, denoted by $N_{\boldsymbol{x}_0}M$, is the set of all normal vectors at $\boldsymbol{x}_0$. Namely, $N_{\boldsymbol{x}_0}M = (T_{\boldsymbol{x}_0}M)^\perp$ is the orthogonal complement.

The key fact is that, for any optimization problem, $\nabla f(\boldsymbol{x}^*)$ is a normal vector of the feasible set Ω at $\boldsymbol{x}^*$, if $\boldsymbol{x}^*$ is a minimum point.

Lemma

For any optimization problem with objective function f and feasible set Ω, if $\boldsymbol{x}^* \in \Omega$ is a *local minimum* of f over Ω, then $\nabla f(\boldsymbol{x}^*) \in N_{\boldsymbol{x}^*}\Omega$ (or equivalently, $\nabla f(\boldsymbol{x}^*) \perp T_{\boldsymbol{x}^*}\Omega$).

Proof

By definition, for every $\boldsymbol{v} \in T_{\boldsymbol{x}^*}\Omega$, there exists $\boldsymbol{\gamma} : (-\varepsilon, \varepsilon) \to \Omega$ such that $\boldsymbol{\gamma}(0) = \boldsymbol{x}^*$ and $\boldsymbol{\gamma}'(0) = \boldsymbol{v}$. Note that $\boldsymbol{x}^* \in \Omega$ is a local minimum of f over Ω, so 0 is a local minimum of $f(\boldsymbol{\gamma}(t))$ over $(-\varepsilon, \varepsilon)$. Thus we have $f'(\boldsymbol{\gamma}(0)) = 0$. Since

$$f'(\boldsymbol{\gamma}(0)) = \langle \nabla f(\boldsymbol{\gamma}(0)), \boldsymbol{\gamma}'(0) \rangle = \langle \nabla f(\boldsymbol{x}^*), \boldsymbol{v} \rangle ,$$

we conclude that $\nabla f(\boldsymbol{x}^*) \perp \boldsymbol{v}$ and hence $\nabla f(\boldsymbol{x}^*) \perp T_{\boldsymbol{x}^*}\Omega$.

The remaining ingredient is how to represent $N_{\boldsymbol{x}^*}\Omega$. Suppose the optimization problem has m constraints $g_1(\boldsymbol{x}) = 0, \dots, g_m(\boldsymbol{x}) = 0$. Using a similar argument we have

$$\forall \boldsymbol{x} \in \Omega, \qquad \nabla g_i(\boldsymbol{x}) \perp T_{\boldsymbol{x}}\Omega ,$$

since $g_i(\boldsymbol{\gamma}(t)) \equiv 0$ is a constant function. Let $\boldsymbol{g}(\boldsymbol{x}) = (g_1(\boldsymbol{x}), \dots, g_m(\boldsymbol{x}))^\mathsf{T}$. Then

$$\boldsymbol{g}'(\boldsymbol{x}) = \begin{pmatrix} \nabla g_1(\boldsymbol{x})^\mathsf{T} \\ \vdots \\ \nabla g_m(\boldsymbol{x})^\mathsf{T} \end{pmatrix}.$$

For any $\boldsymbol{x} \in \Omega$, we have $T_{\boldsymbol{x}}\Omega \subseteq \ker(\boldsymbol{g}'(\boldsymbol{x}))$ (or equivalently, $\operatorname{span}\{\nabla g_1(\boldsymbol{x}), \dots, \nabla g_m(\boldsymbol{x})\} \subseteq N_{\boldsymbol{x}}\Omega$). Finally, we show that if $\boldsymbol{x}^*$ is *regular*, then $T_{\boldsymbol{x}}\Omega = \ker(\boldsymbol{g}'(\boldsymbol{x}))$. Combining with $\nabla f(\boldsymbol{x}^*) \perp T_{\boldsymbol{x}^*}\Omega$, it follows that $\nabla f(\boldsymbol{x}^*) \in \operatorname{span}\{\nabla g_1(\boldsymbol{x}^*), \dots, \nabla g_m(\boldsymbol{x}^*)\}$.

Lemma

If $\boldsymbol{x}_0 \in \Omega$ is a regular point, then for any $\boldsymbol{v} \in \ker(\boldsymbol{g}'(\boldsymbol{x}_0))$, there exists a differentiable curve $\boldsymbol{\gamma} : (-\varepsilon, \varepsilon) \to \Omega$ such that $\boldsymbol{\gamma}(0) = \boldsymbol{x}_0$ and $\boldsymbol{\gamma}(0)' = \boldsymbol{v}$.

The proof is an application of the *implicit function theorem*, and we omit the details here.

Example

Consider the following examples

$$\begin{aligned} \min_{\boldsymbol{x} \in \mathbb{R}^3} \quad & f(\boldsymbol{x}) = x_1^2 + (x_2 - 2)^2 + (x_3 - 1)^2 \\ \text{subject to} \quad & x_1^2 + x_2^2 + x_3^2 = 1 \\ & x_1^2 + x_2^2 = 1\,, \end{aligned}$$

and

$$\begin{aligned} \min_{\boldsymbol{x} \in \mathbb{R}^3} \quad & f(\boldsymbol{x}) = x_1^2 + (x_2 - 2)^2 + (x_3 - 1)^2 \\ \text{subject to} \quad & x_1^2 + x_2^2 + x_3^2 = 1 \\ & x_3 = 0\,. \end{aligned}$$

Note that they have the same feasible set $\{\boldsymbol{x} \in \mathbb{R}^3 \mid x_1^2 + x_2^2 = 1, x_3 = 0\}$. So the optimal solutions are the same ($x_1^* = 0, x_2^* = 1, x_3^* = 0$). The tangent space of the feasible set at the optimal solution is $\{\boldsymbol{v} \in \mathbb{R}^3 \mid v_2 = v_3 = 0\}$. We can verify that $\nabla f(\boldsymbol{x}^*) = \big(2x_1^*, 2(x_2^* - 2), 2(x_3^* - 1)\big)^\mathsf{T} = (0, -2, -2)^\mathsf{T}$ is orthogonal to the tangent space.
However, in the first example, the gradients of constraints at $\boldsymbol{x}^*$ are $(0, -2, 0)^\mathsf{T}$ and $(0, -2, 0)^\mathsf{T}$, which yields that $\boldsymbol{x}^*$ is not regular and Lagrange multipliers

do not exist. In contrast, in the second example, the gradients of constraints are $(0,-2,0)^\mathsf{T}$ and $(0,0,1)^\mathsf{T}$, whose linear span is exactly the normal space at $\boldsymbol{x}^*$. So the Lagrange condition applies.

14.5 Second order condition

It is not difficult to write down a second-order condition. Again, note that if $\boldsymbol{x}^* \in \Omega$ is a local minimum of f over Ω, then for any $\boldsymbol{\gamma} : (-\varepsilon, \varepsilon) \to \Omega$ with $\boldsymbol{\gamma}(0) = \boldsymbol{x}^*$, 0 is a local minimum of $f(\boldsymbol{\gamma}(t))$. Applying the second-order condition for local minima, we have $f''(\boldsymbol{\gamma}(0)) \geq 0$, where

$$f''(\boldsymbol{\gamma}(0)) = \boldsymbol{\gamma}'(0)^\mathsf{T}\nabla^2 f(\boldsymbol{x}^*)\boldsymbol{\gamma}'(0) + \langle \boldsymbol{\gamma}''(0),\, \nabla f(\boldsymbol{x}^*)\rangle\,.$$

Similarly, we also have

$$g_i''(\boldsymbol{\gamma}(0)) = \boldsymbol{\gamma}'(0)^\mathsf{T}\nabla^2 g_i(\boldsymbol{x}^*)\boldsymbol{\gamma}'(0) + \langle \boldsymbol{\gamma}''(0),\, \nabla g_i(\boldsymbol{x}^*)\rangle = 0$$

for every $i = 1, 2, \ldots, m$. By the Lagrange condition, there exists multipliers $\lambda_1^*, \ldots, \lambda_m^*$ such that

$$\nabla f(\boldsymbol{x}^*) + \lambda_1 \nabla g_1(\boldsymbol{x}^*) + \cdots + \lambda_m \nabla g_m(\boldsymbol{x}^*) = \mathbf{0}\,,$$

so it follows that

$$\boldsymbol{\gamma}'(0)^\mathsf{T}\nabla^2 f(\boldsymbol{x}^*)\boldsymbol{\gamma}'(0) + \lambda_1^*\boldsymbol{\gamma}'(0)^\mathsf{T}\nabla^2 g_1(\boldsymbol{x}^*)\boldsymbol{\gamma}'(0) + \cdots + \lambda_m^*\boldsymbol{\gamma}'(0)^\mathsf{T}\nabla^2 g_m(\boldsymbol{x}^*)\boldsymbol{\gamma}'(0) \geq 0\,.$$

Overall, we have the *second-order Lagrange condition*.

Theorem (*Lagrange condition for general optimization problems, second-order*)

Consider a general optimization problem

$$\begin{aligned} \min_{\boldsymbol{x}\in\mathbb{R}^n} \quad & f(\boldsymbol{x}) \\ \text{subject to} \quad & g_i(\boldsymbol{x}) = 0\,, \quad \forall\, i = 1, 2, \ldots, m\,. \end{aligned}$$

If $\boldsymbol{x}^*$ is a *local minimum* in the feasible set, and $\boldsymbol{x}^*$ is a *regular point* of $\boldsymbol{g} = (g_1, \ldots, g_m)^\mathsf{T}$, then there exists $\boldsymbol{\lambda}^* = (\lambda_1^*, \ldots, \lambda_m^*)^\mathsf{T} \in \mathbb{R}^m$ such that

$$\nabla f(\boldsymbol{x}^*) + \lambda_1^* \nabla g_1(\boldsymbol{x}^*) + \cdots + \lambda_m^* \nabla g_m(\boldsymbol{x}^*) = \mathbf{0}\,,$$

and for all $\boldsymbol{v} \in \ker(\boldsymbol{g}'(\boldsymbol{x}^*))$,

$$\boldsymbol{v}^{\mathsf{T}}\Big(\nabla^2 f(\boldsymbol{x}^*) + \lambda_1^* \nabla^2 g_1(\boldsymbol{x}^*) + \cdots + \lambda_m^* \nabla^2 g_m(\boldsymbol{x}^*)\Big)\boldsymbol{v} \geq 0\,.$$