CS 2601 Linear and Convex Optimization

9. Lagrange condition

Bo Jiang

John Hopcroft Center for Computer Science
Shanghai Jiao Tong University

Fall 2022

Outline

- Convex problems with equality constraints

- General equality constrained problems

Equality constrained convex problems

Consider the equality constrained convex optimization problem

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & f(\boldsymbol{x}) \\ \text{s.t.} \quad & \boldsymbol{a}_i^T \boldsymbol{x} = b_i, \quad i = 1, 2, \ldots, k \end{aligned}$$

where f is convex with $\operatorname{dom} f = \mathbb{R}^n$.

In a more compact form,

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & f(\boldsymbol{x}) \\ \text{s.t.} \quad & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} \end{aligned} \qquad \text{(EC)}$$

where $\boldsymbol{A}^T = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_k) \in \mathbb{R}^{n \times k}$, $\boldsymbol{b} = (b_1, \ldots, b_k)^T \in \mathbb{R}^k$.

We assume f is differentiable and the problem is feasible.

Feasible set

The feasible set is

$$X = \{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}\}$$

Given any $\boldsymbol{x}_0 \in X$,

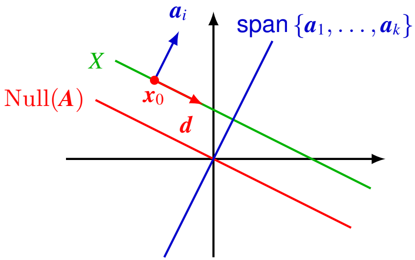
$$X = \boldsymbol{x}_0 + \mathrm{Null}(\boldsymbol{A})$$

where $\mathrm{Null}(\boldsymbol{A}) = \{\boldsymbol{x} : \boldsymbol{A}\boldsymbol{x} = \boldsymbol{0}\} = \{\boldsymbol{x} : \boldsymbol{a}_i^T\boldsymbol{x} = 0, i = 1, \ldots, k\}$ is the null space of $\boldsymbol{A}$.

$\mathrm{Null}(\boldsymbol{A})$ is precisely the set of feasible directions (at any $\boldsymbol{x}_0 \in X$)

$$\boldsymbol{x}_0 + \boldsymbol{d} \in X \iff \boldsymbol{a}_i^T\boldsymbol{d} = 0, \forall i$$

- $\boldsymbol{a}_i$ is a normal vector to X
- $\boldsymbol{d} \in \mathrm{Null}(\boldsymbol{A})$ is a tangent vector to X, the velocity $\boldsymbol{x}'(0)$ of a path $\boldsymbol{x}(t) = \boldsymbol{x}_0 + t\boldsymbol{d} \subset X$

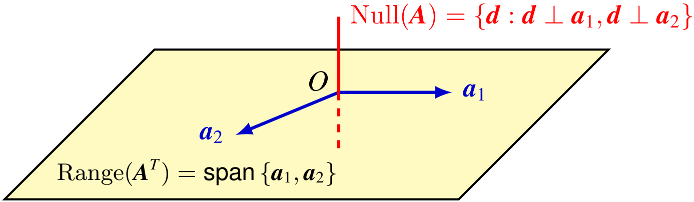


$\mathrm{Range}(\boldsymbol{A}^T) = \text{span}\,\{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_k\}$

Appendix

Lemma. $\mathrm{Null}(\boldsymbol{A})^\perp = \mathrm{Range}(\boldsymbol{A}^T)$, where $\mathrm{Range}(\boldsymbol{A}^T) = \{\boldsymbol{A}^T\boldsymbol{v} : \boldsymbol{v} \in \mathbb{R}^k\}$ and $\mathrm{Null}(\boldsymbol{A})^\perp$ is the orthogonal complement of $\mathrm{Null}(\boldsymbol{A})$, i.e.

$$\boldsymbol{x} \in \mathrm{Null}(\boldsymbol{A})^\perp \iff \boldsymbol{x} \perp \boldsymbol{d}, \quad \forall \boldsymbol{d} \in \mathrm{Null}(\boldsymbol{A})$$



Proof. Show $\mathrm{Range}(\boldsymbol{A}^T) \subset \mathrm{Null}(\boldsymbol{A})^\perp$ is a subspace with the same dimension, so $\mathrm{Range}(\boldsymbol{A}^T) = \mathrm{Null}(\boldsymbol{A})^\perp$.

- $\boldsymbol{x} \in \mathrm{Range}(\boldsymbol{A}^T) \implies \boldsymbol{x} = \boldsymbol{A}^T\boldsymbol{z}$ for some $\boldsymbol{z}$
- $\forall \boldsymbol{d} \in \mathrm{Null}(\boldsymbol{A})$, $\boldsymbol{x}^T\boldsymbol{d} = \boldsymbol{z}^T\boldsymbol{A}\boldsymbol{d} = \boldsymbol{z}^T\boldsymbol{0} = 0$, i.e. $\boldsymbol{x} \perp \boldsymbol{d}$, so $\boldsymbol{x} \in \mathrm{Null}(\boldsymbol{A})^\perp$.
- $\dim \mathrm{Range}(\boldsymbol{A}^T) = \mathrm{rank}\, \boldsymbol{A} = n - \dim \mathrm{Null}(\boldsymbol{A}) = \dim \mathrm{Null}(\boldsymbol{A})^\perp$

Optimality condition

Lemma. $\boldsymbol{x}^* \in X$ is optimal iff

$$\nabla f(\boldsymbol{x}^*) \perp \mathrm{Null}(\boldsymbol{A})$$

Note. Geometrically, $\nabla f(\boldsymbol{x}^*) \perp \mathrm{Null}(\boldsymbol{A})$ means $\nabla f(\boldsymbol{x}^*)$ is perpendicular to all feasible directions, which are also tangent vectors at $\boldsymbol{x}^*$.

Proof. Recall (slide 7 of §5 part 1) $\boldsymbol{x}^* \in X$ is optimal iff

$$\nabla f(\boldsymbol{x}^*)^T(\boldsymbol{x} - \boldsymbol{x}^*) \geq 0, \quad \forall \boldsymbol{x} \in X$$

Note $\boldsymbol{x} \in X$ iff $\boldsymbol{d} = \boldsymbol{x} - \boldsymbol{x}^* \in \mathrm{Null}(\boldsymbol{A})$. The above condition becomes

$$\nabla f(\boldsymbol{x}^*)^T \boldsymbol{d} \geq 0, \quad \forall \boldsymbol{d} \in \mathrm{Null}(\boldsymbol{A})$$

Since $\boldsymbol{d} \in \mathrm{Null}(\boldsymbol{A}) \iff -\boldsymbol{d} \in \mathrm{Null}(\boldsymbol{A})$, the condition then reduces to

$$\nabla f(\boldsymbol{x}^*)^T \boldsymbol{d} = 0, \quad \forall \boldsymbol{d} \in \mathrm{Null}(\boldsymbol{A})$$

Note. If f is nonconvex and $\boldsymbol{x}^*$ a local minimum, then $\nabla f(\boldsymbol{x}^*) \perp \mathrm{Null}(\boldsymbol{A})$ is a necessary condition. For a proof, note $t = 0$ is a local minimum of $g(t) = f(\boldsymbol{x}^* + t\boldsymbol{d})$, so $g'(0) = \nabla f(\boldsymbol{x}^*)^T \boldsymbol{d} = 0$.

Lagrange condition

Theorem. $\boldsymbol{x}^* \in X$ is optimal iff there exists $\boldsymbol{\lambda}^* = (\lambda_1^*, \ldots, \lambda_k^*)^T \in \mathbb{R}^k$ s.t.

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

or written out,

$$\nabla f(\boldsymbol{x}^*) + \sum_{i=1}^{k} \lambda_i^* \boldsymbol{a}_i = \boldsymbol{0}.$$

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

Proof. By the previous lemma, $\boldsymbol{x}^* \in X$ is optimal iff $\nabla f(\boldsymbol{x}^*) \perp \mathrm{Null}(\boldsymbol{A})$. Since

$$\mathrm{Null}(\boldsymbol{A})^{\perp} = \mathrm{Range}(\boldsymbol{A}^T)$$

$\boldsymbol{x}^*$ is optimal iff

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

i.e. there exists $\boldsymbol{v}^*$ s.t. $\nabla f(\boldsymbol{x}^*) = \boldsymbol{A}^T \boldsymbol{v}^* = -\boldsymbol{A}^T \boldsymbol{\lambda}^*$ with $\boldsymbol{\lambda}^* = -\boldsymbol{v}^*$.

Lagrange condition (cont'd)

Define Lagrangian (or Lagrange function) by

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}) = f(\boldsymbol{x}) + \boldsymbol{\lambda}^T(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}) = f(\boldsymbol{x}) + \sum_{i=1}^{k} \lambda_i(\boldsymbol{a}_i^T \boldsymbol{x} - b_i)$$

The optimality condition becomes the following Lagrange condition, aka KKT equations[1]

$$\begin{cases} \nabla_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}^*, \boldsymbol{\lambda}^*) = \nabla f(\boldsymbol{x}^*) + \boldsymbol{A}^T \boldsymbol{\lambda}^* = \boldsymbol{0} \\ \nabla_{\boldsymbol{\lambda}} \mathcal{L}(\boldsymbol{x}^*, \boldsymbol{\lambda}^*) = \boldsymbol{A}\boldsymbol{x}^* - \boldsymbol{b} = \boldsymbol{0} \end{cases}$$

where $\nabla_{\boldsymbol{x}}$ and $\nabla_{\boldsymbol{\lambda}}$ are partial gradientw.r.t. $\boldsymbol{x}$ and $\boldsymbol{\lambda}$, or

$$\nabla \mathcal{L}(\boldsymbol{x}^*, \boldsymbol{\lambda}^*) = \boldsymbol{0}$$

i.e. $(\boldsymbol{x}^*, \boldsymbol{\lambda}^*)$ is a stationary point of $\mathcal{L}$.

[1] KKT stands for Karush-Kuhn-Tucker. We'll see later why it is called as such.

Example

Consider

$$\min_{x_1,x_2} \quad f(x_1,x_2) = \frac{1}{2}x_1^2 + \frac{1}{2}x_2^2$$
$$\text{s.t.} \quad x_1 + 2x_2 = 1$$

Method 1. Reduction to an equivalent unconstrained problem.

$$g(x_2) \triangleq f(1-2x_2, x_2) = \frac{1}{2}(1-2x_2)^2 + \frac{1}{2}x_2^2$$

$$\min_{x_2} g(x_2) \implies g'(x_2^*) = 0 \implies x_2^* = \frac{2}{5} \implies x_1^* = 1 - 2x_2^* = \frac{1}{5}$$

Method 2. Lagrangian multipliers method. The Lagrangian is

$$\mathcal{L}(x_1,x_2,\lambda) = \frac{1}{2}x_1^2 + \frac{1}{2}x_2^2 + \lambda(x_1 + 2x_2 - 1)$$

By the Lagrange condition,

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial x_1} = x_1 + \lambda = 0 \\ \frac{\partial \mathcal{L}}{\partial x_2} = x_2 + 2\lambda = 0 \\ \frac{\partial \mathcal{L}}{\partial \lambda} = x_1 + 2x_2 - 1 = 0 \end{cases} \implies \begin{cases} x_1^* = \frac{1}{5} \\ x_2^* = \frac{2}{5} \\ \lambda^* = -\frac{1}{5} \end{cases}$$

Example (cont'd)

$$\min_{x_1, x_2} \quad f(x_1, x_2) = \frac{1}{2}x_1^2 + \frac{1}{2}x_2^2$$
$$\text{s.t.} \quad x_1 + 2x_2 = 1$$

- normal vector to the feasible set X

$$\boldsymbol{a} = (1, 2)^T$$

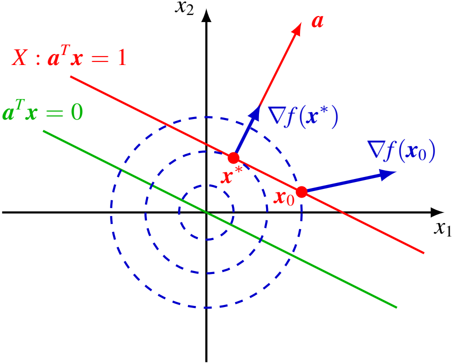
- gradient

$$\nabla f(\boldsymbol{x}) = \boldsymbol{x}$$

- at $\boldsymbol{x}^*$,

$$\nabla f(\boldsymbol{x}^*) = -\lambda^* \boldsymbol{a} \perp X$$

Note X is parallel to $\mathrm{Null}(\boldsymbol{a}^T)$.



Example

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & f(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{x}\|^2, \\ \text{s.t.} \quad & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} \end{aligned} \quad \text{where } \boldsymbol{A} = \begin{bmatrix} 1 & 2 & 0 \\ 2 & 2 & 1 \end{bmatrix}, \boldsymbol{b} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

Method 1. Reduction to an equivalent unconstrained problem.

- $\operatorname{rank} \boldsymbol{A} = 2$. Find two independent columns of $\boldsymbol{A}$, e.g. the first and third columns, and solve for the corresponding x_i's in terms of the others. Let $\boldsymbol{A}_1 = \begin{bmatrix} 1 & 0 \\ 2 & 1 \end{bmatrix}$, $\boldsymbol{A}_2 = \begin{bmatrix} 2 \\ 2 \end{bmatrix}$. The constraints become

$$\boldsymbol{A}_1 \begin{bmatrix} x_1 \\ x_3 \end{bmatrix} + \boldsymbol{A}_2 x_2 = \boldsymbol{b} \implies \begin{bmatrix} x_1 \\ x_3 \end{bmatrix} = \boldsymbol{A}_1^{-1}\boldsymbol{b} - \boldsymbol{A}_1^{-1}\boldsymbol{A}_2 x_2 = \begin{bmatrix} 1 - 2x_2 \\ 2x_2 - 1 \end{bmatrix}$$

- Substitution into f yields

$$g(x_2) = f(1 - 2x_2, x_2, 2x_2 - 1) = (2x_2 - 1)^2 + \frac{1}{2}x_2^2 \implies x_2^* = \frac{4}{9}$$

- $x_1^* = 1 - 2x_2^* = \frac{1}{9}$, $x_3^* = 2x_2^* - 1 = -\frac{1}{9}$

Example (cont'd)

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & f(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{x}\|^2, \\ \text{s.t.} \quad & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} \end{aligned} \qquad \text{where } \boldsymbol{A} = \begin{bmatrix} 1 & 2 & 0 \\ 2 & 2 & 1 \end{bmatrix}, \boldsymbol{b} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

Method 2. Lagrange multipliers method.

- The Lagrangian is

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}) = \frac{1}{2}\|\boldsymbol{x}\|^2 + \boldsymbol{\lambda}^T(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b})$$

- Lagrange condition

$$\begin{cases} \nabla_{\boldsymbol{x}}\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}) = \boldsymbol{x} + \boldsymbol{A}^T\boldsymbol{\lambda} = \boldsymbol{0} \\ \nabla_{\boldsymbol{\lambda}}\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}) = \boldsymbol{A}\boldsymbol{x} - \boldsymbol{b} = \boldsymbol{0} \end{cases} \quad \text{or} \quad \begin{bmatrix} \boldsymbol{I} & \boldsymbol{A}^T \\ \boldsymbol{A} & \boldsymbol{O} \end{bmatrix} \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{\lambda} \end{bmatrix} = \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{b} \end{bmatrix}$$

- Solve for $\boldsymbol{x}, \boldsymbol{\lambda}$ e.g. by substitution or block Gaussian elimination,

$$\begin{cases} \boldsymbol{x}^* = -\boldsymbol{A}^T\boldsymbol{\lambda}^* = \boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{A}^T)^{-1}\boldsymbol{b} \\ \boldsymbol{\lambda}^* = -(\boldsymbol{A}\boldsymbol{A}^T)^{-1}\boldsymbol{b} \end{cases} \implies \begin{cases} \boldsymbol{x}^* = (\frac{1}{9}, \frac{4}{9}, -\frac{1}{9})^T \\ \boldsymbol{\lambda}^* = (-\frac{1}{3}, \frac{1}{9})^T \end{cases}$$

Example (cont'd)

Block Gaussian elimination.

- The augmented matrix is

$$\begin{bmatrix} \boldsymbol{I} & \boldsymbol{A}^T & \boldsymbol{0} \\ \boldsymbol{A} & \boldsymbol{O} & \boldsymbol{b} \end{bmatrix}$$

- Left multiply the first "row" by $-\boldsymbol{A}$ and add to the second "row",

$$\begin{bmatrix} \boldsymbol{I} & \boldsymbol{A}^T & \boldsymbol{0} \\ \boldsymbol{O} & -\boldsymbol{A}\boldsymbol{A}^T & \boldsymbol{b} \end{bmatrix}$$

- Left multiply the second "row" by $-(\boldsymbol{A}\boldsymbol{A}^T)^{-1}$ (why invertible?),

$$\begin{bmatrix} \boldsymbol{I} & \boldsymbol{A}^T & \boldsymbol{0} \\ \boldsymbol{O} & \boldsymbol{I} & -(\boldsymbol{A}\boldsymbol{A}^T)^{-1}\boldsymbol{b} \end{bmatrix}$$

- Left multiply the second "row" by $-\boldsymbol{A}^T$ and add to the first "row",

$$\begin{bmatrix} \boldsymbol{I} & \boldsymbol{O} & \boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{A}^T)^{-1}\boldsymbol{b} \\ \boldsymbol{O} & \boldsymbol{I} & -(\boldsymbol{A}\boldsymbol{A}^T)^{-1}\boldsymbol{b} \end{bmatrix}$$

Example (cont'd)

$$\min_{\boldsymbol{x}} \quad f(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{x}\|^2, \qquad \text{where } \boldsymbol{A} = \begin{bmatrix} 1 & 2 & 0 \\ 2 & 2 & 1 \end{bmatrix}, \boldsymbol{b} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$
$$\text{s.t.} \quad \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$$

- vectors normal to the feasible set X

$$\text{span}\,\{\boldsymbol{a}_1, \boldsymbol{a}_2\}$$

with $\boldsymbol{a}_1 = (1,2,0)^T$, $\boldsymbol{a}_2 = (2,2,1)^T$.

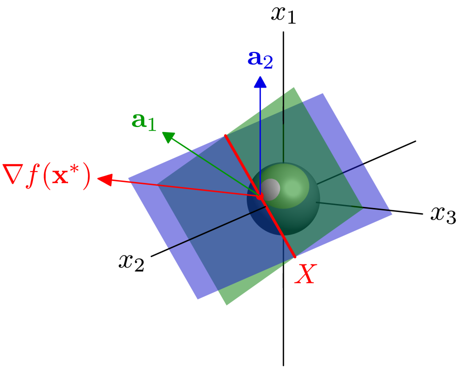
- gradient

$$\nabla f(\boldsymbol{x}) = \boldsymbol{x}$$

- at $\boldsymbol{x}^*$,

$$\nabla f(\boldsymbol{x}^*) = -\lambda_1^* \boldsymbol{a}_1 - \lambda_2^* \boldsymbol{a}_2 \perp X$$

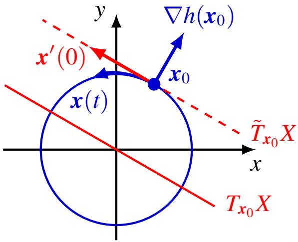
Note X is parallel to $\text{Null}(\boldsymbol{A}^T)$.



Outline

- Convex problems with equality constraints
- General equality constrained problems

Optimization on 2D circle



Consider the constraint in $\mathbb{R}^2$,

$$h(\boldsymbol{x}) = \|\boldsymbol{x}\|^2 - 1 = 0$$

Feasible set $X = \{\boldsymbol{x} : \|\boldsymbol{x}\| = 1\}$. At $\boldsymbol{x}_0 \in X$,

- A tangent vector is the initial velocity $\boldsymbol{x}'(0)$ of a feasible local path $\boldsymbol{x}(t)$ starting at $\boldsymbol{x}_0$, i.e. $\boldsymbol{x}(0) = \boldsymbol{x}_0$, $h(\boldsymbol{x}(t)) = 0$ for small t. Note

$$h'(\boldsymbol{x}_0)\boldsymbol{x}'(0) = \nabla h(\boldsymbol{x}_0)^T \boldsymbol{x}'(0) = 0 \quad \text{i.e.} \quad \boldsymbol{x}'(0) \in \mathrm{Null}(h'(\boldsymbol{x}_0))$$

- A tangent vector $\boldsymbol{d}$ is a feasible direction in the sense that there is a feasible path $\boldsymbol{x}(t)$ in that direction, i.e. $\boldsymbol{d} = \boldsymbol{x}'(0)$.
- The tangent space $T_{\boldsymbol{x}_0}X$ is the set of tangent vectors. It turns out

$$T_{\boldsymbol{x}_0}X = \mathrm{Null}(h'(\boldsymbol{x}_0)) = \{\boldsymbol{d} : \nabla h(\boldsymbol{x}_0)^T \boldsymbol{d} = 0\}$$

- Think of the tangent line $\tilde{T}_{\boldsymbol{x}_0}X$ as $T_{\boldsymbol{x}_0}X$ attached at $\boldsymbol{x}_0$

Optimization on 2D circle

Consider the smooth nonconvex (why?) problem

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & f(\boldsymbol{x}) \\ \text{s.t.} \quad & h(\boldsymbol{x}) = \|\boldsymbol{x}\|^2 - 1 = 0 \end{aligned}$$

Let $\boldsymbol{x}^* \in X$ be a local minimum. Given $\boldsymbol{d} \in \mathrm{Null}(h'(\boldsymbol{x}^*))$, let $\boldsymbol{x}(t)$ be a feasible local path[2] with $\boldsymbol{x}(0) = \boldsymbol{x}^*$, $\boldsymbol{x}'(0) = \boldsymbol{d}$ and $h(\boldsymbol{x}(t)) = 0$ for small t.

Since $\boldsymbol{x}^* = \boldsymbol{x}(0)$ is a local minimum of the constrained problem, $t = 0$ is a local minimum of $g(t) = f(\boldsymbol{x}(t))$, so

$$0 = g'(0) = \nabla f(\boldsymbol{x}^*)^T \boldsymbol{x}'(0) = \nabla f(\boldsymbol{x}^*)^T \boldsymbol{d}$$

Since $\boldsymbol{d} \in \mathrm{Null}(h'(\boldsymbol{x}^*))$ is arbitrary,

$$\nabla f(\boldsymbol{x}^*) \perp \mathrm{Null}(h'(\boldsymbol{x}^*))$$

[2]For example, if $\boldsymbol{x}^* = (\cos\phi_0, \sin\phi_0)$, then $\boldsymbol{d} = (-a\sin\phi_0, a\sin\phi_0)$ for some $a \in \mathbb{R}$. Then $\boldsymbol{x}(t) = (\cos(at+\phi_0), \sin(at+\phi_0))$ satisfies the requirement.

Optimization on 2D circle (cont'd)

By $\mathrm{Null}(\boldsymbol{A})^\perp = \mathrm{Range}(\boldsymbol{A}^T)$,

$$\nabla f(\boldsymbol{x}^*) \in \mathrm{Range}(h'(\boldsymbol{x}^*)^T) = \text{span}\left\{\nabla h(\boldsymbol{x}^*)\right\}$$

so there exists a λ^* s.t.

$$\nabla f(\boldsymbol{x}^*) + \lambda^* \nabla h(\boldsymbol{x}^*) = \mathbf{0}$$

Define the Lagrangian by

$$\mathcal{L}(\boldsymbol{x}, \lambda) = f(\boldsymbol{x}) + \lambda h(\boldsymbol{x})$$

Lagrange condition. $\boldsymbol{x}^*$ is a local optimum only if there exists λ^* s.t.

$$\nabla \mathcal{L}(\boldsymbol{x}^*, \lambda^*) = \mathbf{0}, \quad \text{i.e.} \quad \begin{cases} \nabla_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}^*, \lambda^*) = \nabla f(\boldsymbol{x}^*) + \lambda^* \nabla h(\boldsymbol{x}^*) \\ \nabla_\lambda \mathcal{L}(\boldsymbol{x}^*, \lambda^*) = h(\boldsymbol{x}^*) = 0 \end{cases}$$

Note. This is only a necessary condition for nonconvex problems.

Example

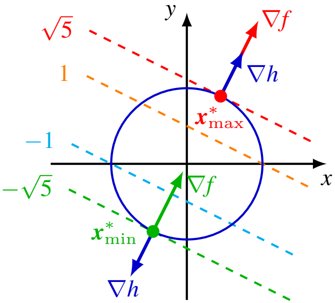
$$\min_{\boldsymbol{x}} \quad f(\boldsymbol{x}) = x + 2y$$
$$\text{s.t.} \quad h(\boldsymbol{x}) = \|\boldsymbol{x}\|^2 - 1 = 0$$

- Lagrange condition

$$\begin{cases} \frac{\partial f(\boldsymbol{x})}{\partial x} + \lambda \frac{\partial h(\boldsymbol{x})}{\partial x} = 1 + 2\lambda x = 0 \implies x = -\frac{1}{2\lambda} \\ \frac{\partial f(\boldsymbol{x})}{\partial y} + \lambda \frac{\partial h(\boldsymbol{x})}{\partial y} = 2 + 2\lambda y = 0 \implies y = -\frac{1}{\lambda} \\ h(\boldsymbol{x}^*) = x^2 + y^2 - 1 = 0 \end{cases}$$

- solutions to the above equations

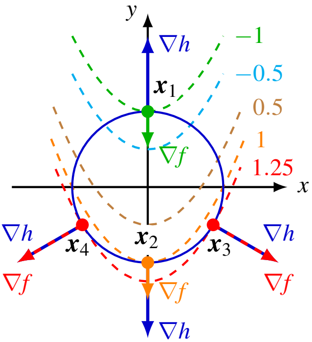
$$(1) \begin{cases} x = -\frac{\sqrt{5}}{5} \\ y = -\frac{2\sqrt{5}}{5} \\ \lambda = \frac{\sqrt{5}}{2} \end{cases} \qquad (2) \begin{cases} x = \frac{\sqrt{5}}{5} \\ y = \frac{2\sqrt{5}}{5} \\ \lambda = -\frac{\sqrt{5}}{2} \end{cases}$$



- (1) global minimum, (2) global maximum
- at all extrema, $\nabla f \parallel \nabla h$ and $\nabla f \perp X$

Example

$$\min_{\boldsymbol{x}} \quad f(\boldsymbol{x}) = x^2 - y$$
$$\text{s.t.} \quad h(\boldsymbol{x}) = \|\boldsymbol{x}\|^2 - 1 = 0$$



- Lagrange condition

$$\begin{cases} \dfrac{\partial f(\boldsymbol{x})}{\partial x} + \lambda \dfrac{\partial h(\boldsymbol{x})}{\partial x} = 2x + 2\lambda x = 0 \\ \dfrac{\partial f(\boldsymbol{x})}{\partial y} + \lambda \dfrac{\partial h(\boldsymbol{x})}{\partial y} = -1 + 2\lambda y = 0 \\ h(\boldsymbol{x}^*) = x^2 + y^2 - 1 = 0 \end{cases}$$

- solutions to above equations

$$(1) \begin{cases} x = 0 \\ y = 1 \\ \lambda = \frac{1}{2} \end{cases} \quad (2) \begin{cases} x = 0 \\ y = -1 \\ \lambda = -\frac{1}{2} \end{cases} \quad (3) \begin{cases} x = \frac{\sqrt{3}}{2} \\ y = -\frac{1}{2} \\ \lambda = -1 \end{cases} \quad (4) \begin{cases} x = -\frac{\sqrt{3}}{2} \\ y = -\frac{1}{2} \\ \lambda = -1 \end{cases}$$

- (1) global minimum, (2) local minimum, (3)(4) global maxima
- at all extrema (and certain other points), $\nabla f \parallel \nabla h$ and $\nabla f \perp X$

Exercise. Solve equivalent problem $g(y) = 1 - y^2 - y$ s.t. $|y| \leq 1$.

General equality constraints

Consider a general equality constraint function $\boldsymbol{h}$, where $\boldsymbol{h} : \mathbb{R}^n \to \mathbb{R}^k$ has smooth components $h_1, \ldots, h_k$. The feasible set is

$$X = \{\boldsymbol{x} : \boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{0}\}$$

A point $\boldsymbol{x}_0$ is a regular point of $\boldsymbol{h}$ if

$$\boldsymbol{h}'(\boldsymbol{x}_0) = \begin{bmatrix} \nabla h_1(\boldsymbol{x}_0)^T \\ \vdots \\ \nabla h_k(\boldsymbol{x}_0)^T \end{bmatrix}$$

has full (row) rank k, or equivalently, $\nabla h_1(\boldsymbol{x}_0), \ldots, \nabla h_k(\boldsymbol{x}_0)$ are linearly independent; otherwise it is a critical point of $\boldsymbol{h}$.

At a regular point $\boldsymbol{x}_0$, the local geometry of X can be well characterized by the first order information $\boldsymbol{h}'(\boldsymbol{x}_0)$, or $\nabla h_1(\boldsymbol{x}_0), \ldots, \nabla h_k(\boldsymbol{x}_0)$, and the derivation on slides 16-17 carries over.

Tangent space and normal space

A tangent vector of X at $\boldsymbol{x}_0 \in X$ is the initial velocity $\boldsymbol{x}'(0)$ of a feasible local path $\boldsymbol{x}(t)$ starting at $\boldsymbol{x}_0$, i.e. $\boldsymbol{x}(0) = \boldsymbol{x}_0$, $h(\boldsymbol{x}(t)) = 0$ for small t. Note

$$\frac{d}{dt}\boldsymbol{h}(\boldsymbol{x}(t))\Big|_{t=0} = \boldsymbol{h}'(\boldsymbol{x}_0)\boldsymbol{x}'(0) = \boldsymbol{0} \quad \text{i.e.} \quad \boldsymbol{x}'(0) \in \mathrm{Null}(\boldsymbol{h}'(\boldsymbol{x}_0))$$

The tangent space $T_{\boldsymbol{x}_0}X$ of X at $\boldsymbol{x}_0$ is the set of all tangent vectors at $\boldsymbol{x}_0$.

The normal space $N_{\boldsymbol{x}_0}X$ of X at $\boldsymbol{x}_0$ is the orthogonal complement of $T_{\boldsymbol{x}_0}X$,

$$N_{\boldsymbol{x}_0}X = [T_{\boldsymbol{x}_0}X]^{\perp}$$

Theorem. At a regular point $\boldsymbol{x}_0 \in X$,

$$T_{\boldsymbol{x}_0}X = \mathrm{Null}(\boldsymbol{h}'(\boldsymbol{x}_0)) = \{\boldsymbol{d} : \nabla h_i(\boldsymbol{x}_0)^T\boldsymbol{d} = 0, \quad i = 1, 2, \ldots, k\}$$

and

$$N_{\boldsymbol{x}_0}X = \text{span}\,\{\nabla h_1(\boldsymbol{x}_0), \ldots, \nabla h_k(\boldsymbol{x}_0)\}$$

Tangent space and normal space

Proof

We already know

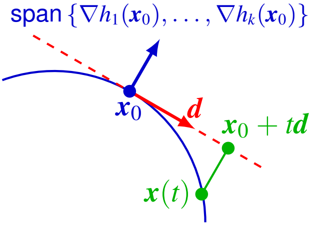
$$T_{\boldsymbol{x}_0}X \subset \mathrm{Null}(\boldsymbol{h}'(\boldsymbol{x}_0))$$

For $\mathrm{Null}(\boldsymbol{h}'(\boldsymbol{x}_0)) \subset T_{\boldsymbol{x}_0}X$, we have the following

Lemma. If $\boldsymbol{x}_0$ is a regular point, then for any $\boldsymbol{d}$ s.t. $\boldsymbol{h}'(\boldsymbol{x}_0)\boldsymbol{d} = \boldsymbol{0}$, there exists a local path $\boldsymbol{x}(t)$ s.t. $\boldsymbol{h}(\boldsymbol{x}(t)) = \boldsymbol{0}$, $\boldsymbol{x}(0) = \boldsymbol{x}_0$ and $\boldsymbol{x}'(0) = \boldsymbol{d}$.

Proof. Let

$$\begin{aligned}\tilde{\boldsymbol{x}}(t, \boldsymbol{\alpha}) &= \boldsymbol{x}_0 + t\boldsymbol{d} + \boldsymbol{h}'(\boldsymbol{x}_0)^T \boldsymbol{\alpha}, \\ &= \boldsymbol{x}_0 + t\boldsymbol{d} + \sum_{i=1}^{k} \alpha_i \nabla h_i(\boldsymbol{x}_0)\end{aligned}$$



and

$$\boldsymbol{F}(t, \boldsymbol{\alpha}) = \boldsymbol{h}(\tilde{\boldsymbol{x}}(t, \boldsymbol{\alpha}))$$

Proof of lemma (cont'd)

Note

$$\boldsymbol{F}(0,\mathbf{0}) = \boldsymbol{h}(\boldsymbol{x}_0) = \mathbf{0}, \quad \frac{\partial \boldsymbol{F}(0,\mathbf{0})}{\partial \boldsymbol{\alpha}} = \boldsymbol{h}'(\boldsymbol{x}_0)\boldsymbol{h}'(\boldsymbol{x}_0)^T \succ \boldsymbol{O}$$

since $\boldsymbol{h}'(\boldsymbol{x}_0)^T$ has full rank k by regularity at $\boldsymbol{x}_0$.

By the Implicit Function Theorem, there exists $\boldsymbol{\alpha} = \boldsymbol{\phi}(t)$ for small t s.t. $\boldsymbol{\phi}(0) = \mathbf{0}$, $\boldsymbol{F}(t, \boldsymbol{\phi}(t)) = \mathbf{0}$ and

$$\boldsymbol{\phi}'(0) = -\left[\frac{\partial \boldsymbol{F}(0,\mathbf{0})}{\partial \boldsymbol{\alpha}}\right]^{-1} \frac{\partial \boldsymbol{F}(0,\mathbf{0})}{\partial t} = -\left[\frac{\partial \boldsymbol{F}(0,\mathbf{0})}{\partial \boldsymbol{\alpha}}\right]^{-1} \boldsymbol{h}'(\boldsymbol{x}_0)\boldsymbol{d} = 0$$

Then

$$\boldsymbol{x}(t) = \tilde{\boldsymbol{x}}(t, \boldsymbol{\phi}(t)) = \boldsymbol{x}_0 + t\boldsymbol{d} + \boldsymbol{h}'(\boldsymbol{x}_0)^T \boldsymbol{\phi}(t) = \boldsymbol{x}_0 + t\boldsymbol{d} + \sum_{i=1}^{k} \phi_i(t) \nabla h_i(\boldsymbol{x}_0)$$

satisfies the requirement.

Appendix: Implicit function theorem

Write $\boldsymbol{F} : \mathbb{R}^{n+k} \to \mathbb{R}^k$ as $\boldsymbol{F}(\boldsymbol{x}, \boldsymbol{y})$ with $\boldsymbol{x} \in \mathbb{R}^n$ and $\boldsymbol{y} \in \mathbb{R}^k$. Let $\boldsymbol{F} = (F_1, \ldots, F_k)^T$, and

$$\frac{\partial \boldsymbol{F}}{\partial \boldsymbol{x}} = \begin{bmatrix} \frac{\partial F_1}{\partial x_1} & \cdots & \frac{\partial F_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial F_k}{\partial x_1} & \cdots & \frac{\partial F_k}{\partial x_n} \end{bmatrix}, \quad \frac{\partial \boldsymbol{F}}{\partial \boldsymbol{y}} = \begin{bmatrix} \frac{\partial F_1}{\partial y_1} & \cdots & \frac{\partial F_1}{\partial y_k} \\ \vdots & \ddots & \vdots \\ \frac{\partial F_k}{\partial y_1} & \cdots & \frac{\partial F_k}{\partial y_k} \end{bmatrix}$$

Implicit Function Theorem. If $\boldsymbol{F} : \mathbb{R}^{n+k} \to \mathbb{R}^k$ is continuously differentiable in a neighborhood $(\boldsymbol{x}_0, \boldsymbol{y}_0)$, and satisfies

$$\boldsymbol{F}(\boldsymbol{x}_0, \boldsymbol{y}_0) = \boldsymbol{0}, \quad \det \frac{\partial \boldsymbol{F}(\boldsymbol{x}_0, \boldsymbol{y}_0)}{\partial \boldsymbol{y}} \neq 0,$$

then there exists continuously differentiable function $\boldsymbol{y} = \phi(\boldsymbol{x})$ defined in a neighborhood of $\boldsymbol{x}_0$ s.t.

$$\boldsymbol{F}(\boldsymbol{x}, \phi(\boldsymbol{x})) = 0, \quad \frac{\partial \phi(\boldsymbol{x})}{\partial \boldsymbol{x}} = -\left[\frac{\partial \boldsymbol{F}(\boldsymbol{x}, \phi(\boldsymbol{x}))}{\partial \boldsymbol{y}}\right]^{-1} \frac{\partial \boldsymbol{F}(\boldsymbol{x}, \phi(\boldsymbol{x}))}{\partial \boldsymbol{x}}$$

First-order necessary condition

Let $\boldsymbol{x} \in \mathbb{R}^n$ and $n \geq k$. Consider the equality constrained problem

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & f(\boldsymbol{x}) \\ \text{s.t.} \quad & h_i(\boldsymbol{x}) = 0, \ i = 1, 2, \dots, k \end{aligned} \tag{ECP}$$

Theorem. If $\boldsymbol{x}^*$ is a local extremum of f s.t. $\boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{0}$, and $\boldsymbol{x}^*$ is a regular point of $\boldsymbol{h}$, then there exist Lagrange multipliers $\lambda_1^*, \dots, \lambda_k^* \in \mathbb{R}$ s.t.

$$\nabla f(\boldsymbol{x}^*) + \sum_{i=1}^{k} \lambda_i^* \nabla h_i(\boldsymbol{x}^*) = \boldsymbol{0}$$

Note. This simply says $\nabla f(\boldsymbol{x}^*) \in N_{\boldsymbol{x}_0} X =$ span $\{\nabla h_1(\boldsymbol{x}_0), \dots, \nabla h_k(\boldsymbol{x}_0)\}$.

Define the Lagrangian of (ECP) by

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}) = f(\boldsymbol{x}) + \boldsymbol{\lambda}^T \boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{f}(\boldsymbol{x}) + \sum_{i=1}^{k} \lambda_i h_i(\boldsymbol{x})$$

Then the Lagrange condition is $\nabla \mathcal{L}(\boldsymbol{x}^*, \boldsymbol{\lambda}^*) = \boldsymbol{0}$.

Proof

Let $\boldsymbol{d} \in T_{\boldsymbol{x}_0}X$ and $\boldsymbol{x}(t)$ a feasible local path at $\boldsymbol{x}^*$ with $\boldsymbol{x}'(0) = \boldsymbol{d}$. Then $t = 0$ is a local minimum of $g(t) = f(\boldsymbol{x}(t))$, so

$$0 = g'(0) = \nabla f(\boldsymbol{x}^*)^T \boldsymbol{d}$$

Since $\boldsymbol{d}$ is arbitrary,

$$\nabla f(\boldsymbol{x}^*) \perp T_{\boldsymbol{x}_0}X$$

and hence

$$\nabla f(\boldsymbol{x}^*) \in [T_{\boldsymbol{x}_0}X]^\perp = N_{\boldsymbol{x}_0}X$$

Since

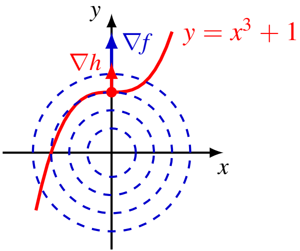
$$N_{\boldsymbol{x}_0}X = \text{span}\,\{\nabla h_1(\boldsymbol{x}_0), \ldots, \nabla h_k(\boldsymbol{x}_0)\}$$

at a regular point $\boldsymbol{x}_0$, there exist $\lambda_1^*, \ldots, \lambda_k^* \in \mathbb{R}$ s.t.

$$\nabla f(\boldsymbol{x}^*) + \sum_{i=1}^{k} \lambda_i^* \nabla h_i(\boldsymbol{x}^*) = \boldsymbol{0}$$

Insufficiency of Lagrange condition

$\boldsymbol{x}^*$ satisfying the Lagrange condition may be neither a maximum nor a minimum. E.g.



$$f(\boldsymbol{x}) = \|\boldsymbol{x}\|^2$$
$$h(\boldsymbol{x}) = y - x^3 - 1$$

At $\boldsymbol{x}^* = (0, 1)^T$,

$$\nabla f(\boldsymbol{x}^*) = (0, 2)^T, \quad \nabla h(\boldsymbol{x}^*) = (0, 1)^T$$

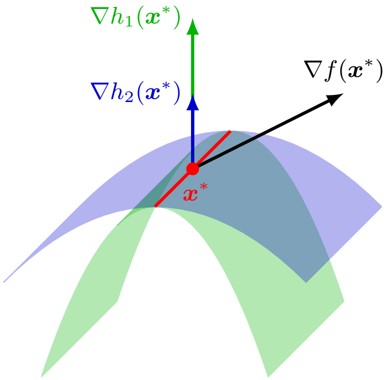
so

$$\nabla f(\boldsymbol{x}^*) - 2\nabla h(\boldsymbol{x}^*) = \mathbf{0}$$

but $\boldsymbol{x}^*$ is neither a maximum nor a minimum.

Second-order conditions can help distinguish different cases ([CZ, LY])

Critical points



Critical points

The Lagrange condition may fail at critical points.

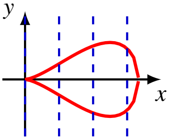
Example.

$$\min_{x,y} \quad f(x,y) = x + y$$
$$\text{s.t.} \quad h(x,y) = x^2 + y^2 = 0$$

The feasible set is $X = \{\mathbf{0}\}$, so $\boldsymbol{x}^* = \mathbf{0}$ is the global minimum. There is no $\lambda^* \in \mathbb{R}$ satisfying the Lagrange condition $\nabla f(\boldsymbol{x}^*) + \lambda^* \nabla h(\boldsymbol{x}^*) = \mathbf{0}$, as $\nabla f(\boldsymbol{x}^*) = (1,1)^T$ and $\nabla h(\boldsymbol{x}^*) = \mathbf{0}$.

Example.

$$\min_{x,y} \quad f(x,y) = x$$
$$\text{s.t.} \quad h(x,y) = y^2 + x^4 - x^3 = 0$$



Note $x^3 - x^4 = y^2 \geq 0$ implies $x \in [0,1]$, so $\boldsymbol{x}^* = \mathbf{0}$ is the global minimum. Lagrange condition fails as $\nabla f(\boldsymbol{x}^*) = (1,0)^T, \nabla h(\boldsymbol{x}^*) = \mathbf{0}$.

Note. To find the minimum, we need to check both regular points satisfying the Lagrange condition and feasible critical points.

Example

$$
\begin{aligned}
\min_{\boldsymbol{x}\in\mathbb{R}^3} \quad & f(\boldsymbol{x}) = x_1 + 2x_2 + x_3 \\
\text{s.t.} \quad & h_1(\boldsymbol{x}) = x_1 + x_2 + 2x_3 = 0 \\
& h_2(\boldsymbol{x}) = \|\boldsymbol{x}\|^2 - 1 = 0
\end{aligned}
$$

A critical point $\boldsymbol{x}$ satisfies $\nabla h_2(\boldsymbol{x}) \parallel \nabla h_1(\boldsymbol{x})$, so $\boldsymbol{x} \propto (1,1,2)^T$, infeasible.

The Lagrangian is

$$
\mathcal{L}(\boldsymbol{x},\boldsymbol{\lambda}) = x_1 + 2x_2 + x_3 + \lambda_1(x_1 + x_2 + 2x_3) + \lambda_2(x_1^2 + x_2^2 + x_3^2 - 1)
$$

The Lagrange condition is

$$
\begin{cases}
\partial_{x_1}\mathcal{L} = 1 + \lambda_1 + 2\lambda_2 x_1 = 0 & (1) \\
\partial_{x_2}\mathcal{L} = 2 + \lambda_1 + 2\lambda_2 x_2 = 0 & (2) \\
\partial_{x_3}\mathcal{L} = 1 + 2\lambda_1 + 2\lambda_2 x_3 = 0 & (3) \\
\partial_{\lambda_1}\mathcal{L} = x_1 + x_2 + 2x_3 = 0 & (4) \\
\partial_{\lambda_2}\mathcal{L} = x_1^2 + x_2^2 + x_3^2 - 1 = 0 & (5)
\end{cases}
$$

Example (cont'd)

- (1)+(2)+(3)$\times 2$,

$$5 + 6\lambda_1 + 2\lambda_2(x_1 + x_2 + 2x_3) = 0 \tag{6}$$

- Plugging (4) into (6) yields $\lambda_1 = -\frac{5}{6}$.
- Plugging λ_1 into (1)(2)(3), and noting that $\lambda_2 \neq 0$,

$$x_1 = -\frac{1}{12\lambda_2}, \quad x_2 = -\frac{7}{12\lambda_2}, \quad x_3 = \frac{1}{3\lambda_2} \tag{7}$$

- Plugging (7) into (5) yields $\lambda_2 = \pm\sqrt{\frac{33}{72}}$, so

$$(1)\begin{cases} x_1 = -\frac{1}{\sqrt{66}} \\ x_2 = -\frac{7}{\sqrt{66}} \\ x_3 = \frac{4}{\sqrt{66}} \\ \lambda_1 = -\frac{5}{6} \\ \lambda_2 = \sqrt{\frac{33}{72}} \end{cases} \quad \text{or} \quad (2)\begin{cases} x_1 = \frac{1}{\sqrt{66}} \\ x_2 = \frac{7}{\sqrt{66}} \\ x_3 = -\frac{4}{\sqrt{66}} \\ \lambda_1 = -\frac{5}{6} \\ \lambda_2 = -\sqrt{\frac{33}{72}} \end{cases}$$

- (1) global minimum, (2) global maximum

Example (cont'd)