CS257 Linear and Convex Optimization

Lecture 13

Bo Jiang

John Hopcroft Center for Computer Science
Shanghai Jiao Tong University

November 30, 2020

Recap: Equality Constrained Convex Problem

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

where $\boldsymbol{A} \in \mathbb{R}^{k\times n}$, $\boldsymbol{b} \in \mathbb{R}^k$, f is differentiable and convex.

Lagrangian.

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}) = f(\boldsymbol{x}) + \boldsymbol{\lambda}^T(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b})$$

Lagrange condition. $\boldsymbol{x}^*$ is optimal iff $\exists$ Lagrange multiplier $\boldsymbol{\lambda}^* \in \mathbb{R}^k$ s.t.

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

Convex QP. $f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T\boldsymbol{Q}\boldsymbol{x} + \boldsymbol{g}^T\boldsymbol{x} + c$, where $\boldsymbol{Q} \succeq \boldsymbol{O}$. KKT system

$$\begin{cases} \boldsymbol{Q}\boldsymbol{x} + \boldsymbol{g} + \boldsymbol{A}^T\boldsymbol{\lambda} = \boldsymbol{0} \\ \boldsymbol{A}\boldsymbol{x} - \boldsymbol{b} = \boldsymbol{0} \end{cases} \quad \text{or} \quad \begin{bmatrix} \boldsymbol{Q} & \boldsymbol{A}^T \\ \boldsymbol{A} & \boldsymbol{O} \end{bmatrix} \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{\lambda} \end{bmatrix} = \begin{bmatrix} -\boldsymbol{g} \\ \boldsymbol{b} \end{bmatrix}$$

Recap: Newton's Method

Solve an approximate quadratic problem in each iteration.

$$\begin{aligned} \min_{\boldsymbol{d}} \quad & f(\boldsymbol{x}+\boldsymbol{d}) = \frac{1}{2}\boldsymbol{d}^T \nabla^2 f(\boldsymbol{x})\boldsymbol{d} + \nabla f(\boldsymbol{x})^T \boldsymbol{d} + f(\boldsymbol{x}) \\ \text{s.t.} \quad & \boldsymbol{A}\boldsymbol{d} = \boldsymbol{0} \end{aligned}$$

initialization $\boldsymbol{x} \leftarrow \boldsymbol{x}_0$ s.t. $\boldsymbol{A}\boldsymbol{x}_0 = \boldsymbol{b}$
repeat
    Compute Newton's direction $\boldsymbol{d}$ by solving

$$\begin{bmatrix} \nabla^2 f(\boldsymbol{x}) & \boldsymbol{A}^T \\ \boldsymbol{A} & \boldsymbol{O} \end{bmatrix} \begin{bmatrix} \boldsymbol{d} \\ \boldsymbol{\lambda} \end{bmatrix} = \begin{bmatrix} -\nabla f(\boldsymbol{x}) \\ \boldsymbol{0} \end{bmatrix}$$

    find stepsize t by backtracking line search
    $\boldsymbol{x} \leftarrow \boldsymbol{x} + t\boldsymbol{d}$
until $\|\boldsymbol{d}\| \leq \delta$
return $\boldsymbol{x}$

Contents

Optimization on 2D Circle

Let $\boldsymbol{x} = (x, y)^T \in \mathbb{R}^2$. Consider the following nonconvex (why?) problem

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

Parameterize the feasible set by $\boldsymbol{x}(t) = (\cos t, \sin t)^T$ and reduce the above constrained problem to the following unconstrained problem

$$\min_t \; g(t) \triangleq f(\boldsymbol{x}(t)) = f(\cos t, \sin t)$$

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

$$g'(t^*) = \frac{\partial f(\boldsymbol{x}^*)}{\partial x} x'(t^*) + \frac{\partial f(\boldsymbol{x}^*)}{\partial y} y'(t^*) = 0$$

On the other hand, $h(\boldsymbol{x}(t)) = 0$. Differentiating w.r.t. t at t^*,

$$\frac{\partial h(\boldsymbol{x}^*)}{\partial x} x'(t^*) + \frac{\partial h(\boldsymbol{x}^*)}{\partial y} y'(t^*) = 0$$

Optimization on 2D Circle (cont'd)

Combining the previous two equations,

$$\begin{bmatrix} \frac{\partial f(\boldsymbol{x}^*)}{\partial x} & \frac{\partial f(\boldsymbol{x}^*)}{\partial y} \\ \frac{\partial h(\boldsymbol{x}^*)}{\partial x} & \frac{\partial h(\boldsymbol{x}^*)}{\partial y} \end{bmatrix} \begin{bmatrix} x'(t^*) \\ y'(t^*) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \text{or} \quad \begin{bmatrix} \nabla f(\boldsymbol{x}^*)^T \\ \nabla h(\boldsymbol{x}^*)^T \end{bmatrix} \boldsymbol{x}'(t^*) = \boldsymbol{0}$$

- The linear system has a solution $\boldsymbol{x}'(t^*) = (-\sin t^*, \cos t^*)^T \neq \boldsymbol{0}$, so $\nabla f(\boldsymbol{x}^*)$ and $\nabla h(\boldsymbol{x}^*)$ must be linearly dependent.
- Note $\nabla h(\boldsymbol{x}^*) = \boldsymbol{x}^* \neq \boldsymbol{0}$ (why?), so there exists λ^* s.t.

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

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.

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

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

Example

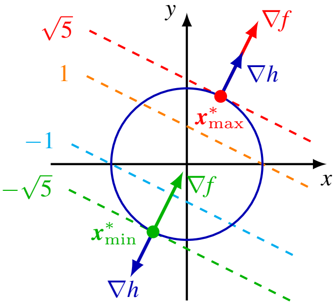
$$\begin{aligned} \min_{\boldsymbol{x}} \quad & f(\boldsymbol{x}) = x + 2y \\ \text{s.t.} \quad & h(\boldsymbol{x}) = \|\boldsymbol{x}\|^2 - 1 = 0 \end{aligned}$$

- 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

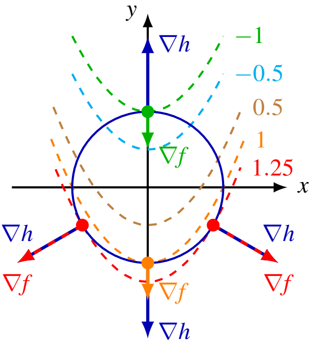
$$(1) \begin{cases} x = -\frac{\sqrt{5}}{5} \\ y = -\frac{2\sqrt{5}}{5} \\ \lambda = \frac{\sqrt{5}}{2} \end{cases} \qquad (2) \begin{cases} x = \frac{\sqrt{5}}{5} \\ y = \frac{2\sqrt{5}}{5} \\ \lambda = -\frac{\sqrt{5}}{2} \end{cases}$$



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

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

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

- 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, $\nabla f \parallel \nabla h \perp X$

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

Implicit Function Theorem in 2D

The derivation on slides 4-5 works for general h of two variables, as long as we can parameterize the feasible set in a neighborhood of $\boldsymbol{x}^*$ by $\boldsymbol{x}(t)$, i.e. $h(\boldsymbol{x}(t)) = 0$, s.t. $\boldsymbol{x}'(t^*) \neq \mathbf{0}$ and $\nabla h(\boldsymbol{x}^*) \neq \mathbf{0}$. The Implicit Function Theorem guarantees this is possible if $\nabla h(\boldsymbol{x}^*) \neq \mathbf{0}$.

Implicit Function Theorem. If $F(x, y)$ is continuously differentiable in a neighborhood of (x_0, y_0), and satisfies

$$F(x_0, y_0) = 0, \quad \frac{\partial F(x_0, y_0)}{\partial y} \neq 0$$

then there exists a continuously differentiable function $y = \phi(x)$ defined in a neighborhood of x_0 s.t.

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

Implicit Function Theorem and Parameterization

If $\nabla h(x_0, y_0) \neq \mathbf{0}$, then either $\frac{\partial h(x_0,y_0)}{\partial x} \neq 0$ or $\frac{\partial h(x_0,y_0)}{\partial y} \neq 0$.

- If $\frac{\partial h(x_0,y_0)}{\partial y} \neq 0$, we can parameterize the feasible set by $t = x$,

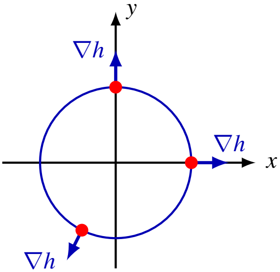
$$\boldsymbol{x}(t) = (t, \phi(t))^T \quad \text{with} \quad \boldsymbol{x}'(t) = (1, \phi'(t))^T \neq \mathbf{0}$$

- If $\frac{\partial h(x_0,y_0)}{\partial y} \neq 0$, we can parameterize the feasible set by $t = y$,

$$\boldsymbol{x}(t) = (\psi(t), t)^T \quad \text{with} \quad \boldsymbol{x}'(t) = (\psi'(t), 1)^T \neq \mathbf{0}$$

Example. For $h(\boldsymbol{x}) = \|\boldsymbol{x}\|^2 - 1$.

- at $\boldsymbol{x}_0 = (1, 0)^T$, use $\boldsymbol{x}(t) = (\sqrt{1 - t^2}, t)^T$
- at $\boldsymbol{x}_0 = (0, 1)^T$, use $\boldsymbol{x}(t) = (t, \sqrt{1 - t^2})^T$
- at $\boldsymbol{x}_0$ in the 3rd quadrant, we can use $\boldsymbol{x}(t) = (t, -\sqrt{1 - t^2})^T$ or $\boldsymbol{x}(t) = (-\sqrt{1 - t^2}, t)^T$



First-order Necessary Condition in 2D

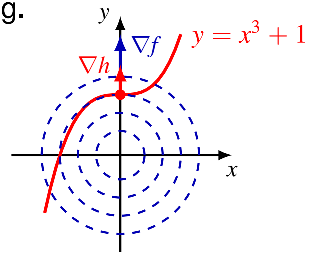
A point $\boldsymbol{x}$ is called a regular point of a function h if $\nabla h(\boldsymbol{x}) \neq \mathbf{0}$; otherwise it is called a critical point.

Theorem. If $\boldsymbol{x}^*$ is a local extremum (maximum or minimum) of f s.t. $h(\boldsymbol{x}) = 0$, and $\boldsymbol{x}^*$ is a regular point of h, then there exists λ^* s.t.

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

Note. $\boldsymbol{x}^*$ satisfying the above 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 $\boldsymbol{x}^* = (0, 1)^T$,

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

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

Critical Points

The Lagrange condition may fail at critical points.

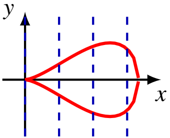
Example.

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

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.

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



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.

First-order Necessary Condition

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

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

A point $\boldsymbol{x}$ is a regular point of $\boldsymbol{h} = (h_1, \ldots, h_k)^T$ if $\nabla h_1(\boldsymbol{x}), \ldots, \nabla h_k(\boldsymbol{x})$ are linearly independent; otherwise it is a critical point of $\boldsymbol{h}$.

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 $\lambda_1^*, \ldots, \lambda_k^* \in \mathbb{R}$ s.t.

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

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

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 $\boldsymbol{y} = \boldsymbol{\phi}(\boldsymbol{x})$ defined in a neighborhood of $\boldsymbol{x}_0$ s.t.

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

Appendix: Proof of Lagrange Condition

Let $\boldsymbol{h} = (h_1, \ldots, h_k)^T$. Note that the Jacobian matrix of $\boldsymbol{h}$ is

$$\frac{\partial \boldsymbol{h}(\boldsymbol{x}^*)}{\partial \boldsymbol{x}} = \begin{bmatrix} \frac{\partial h_1(\boldsymbol{x}^*)}{\partial x_1} & \cdots & \frac{\partial h_1(\boldsymbol{x}^*)}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial h_k(\boldsymbol{x}^*)}{\partial x_1} & \cdots & \frac{\partial h_k(\boldsymbol{x}^*)}{\partial x_n} \end{bmatrix} = \begin{bmatrix} \nabla h_1(\boldsymbol{x}^*)^T \\ \vdots \\ \nabla h_k(\boldsymbol{x}^*)^T \end{bmatrix}$$

Since $\boldsymbol{x}^*$ is regular, $\operatorname{rank} \frac{\partial \boldsymbol{h}(\boldsymbol{x}^*)}{\partial \boldsymbol{x}} = k$. By re-indexing $x_1, \ldots x_n$ if necessary, we assume the last m columns are linearly independent. Let $\boldsymbol{y} = (x_1, \ldots, x_{n-k})^T$, $\boldsymbol{z} = (x_{n-k+1}, \ldots, x_n)^T$.

By the Implicit Function Theorem, there is a continuously differentiable function $\boldsymbol{z} = \phi(\boldsymbol{y})$ s.t. $\boldsymbol{h}(\boldsymbol{y}, \phi(\boldsymbol{y})) = \boldsymbol{0}$, i.e. we can parameterize the feasible set X by[1]

$$\boldsymbol{x}(\boldsymbol{y}) = (\boldsymbol{y}, \phi(\boldsymbol{y}))$$

(ECP) reduces to

$$\min_{\boldsymbol{y}} \; g(\boldsymbol{y}) = f(\boldsymbol{y}, \phi(\boldsymbol{y}))$$

[1] we are sloppy about the shape here, but it should not cause any confusion.

Proof (cont'd)

Since $\boldsymbol{x}^* = \boldsymbol{x}(\boldsymbol{y}^*) = (\boldsymbol{y}^*, \phi(\boldsymbol{y}^*))$ is a local extremum of (ECP), $\boldsymbol{y}^*$ is a local extreme of g. Recalling $y_\ell = x_\ell$ for $\ell = 1, \ldots, n-k$,

$$\frac{\partial g(\boldsymbol{y}^*)}{\partial y_\ell} = \frac{\partial f(\boldsymbol{x}^*)}{\partial x_\ell} + \sum_{j=n-k+1}^{n} \frac{\partial f(\boldsymbol{x}^*)}{\partial x_j} \frac{\partial \phi_{j-(n-k)}(\boldsymbol{y}^*)}{\partial y_\ell} = 0, \quad \ell = 1, \ldots, n-k$$

Differentiating $h_i(\boldsymbol{x}(\boldsymbol{y})) = h_i(\boldsymbol{y}, \phi(\boldsymbol{y})) = 0$ at $\boldsymbol{y}^*$,

$$\frac{\partial h_i(\boldsymbol{x}^*)}{\partial y_\ell} = \frac{\partial h_i(\boldsymbol{x}^*)}{\partial y_\ell} + \sum_{j=n-k+1}^{n} \frac{\partial h_j(\boldsymbol{x}^*)}{\partial x_j} \frac{\partial \phi_{j-(n-k)}(\boldsymbol{y}^*)}{\partial y_\ell} = 0, \quad \ell = 1, \ldots, n-k$$

In matrix form,

$$\begin{bmatrix} 1 & \ldots & 0 & \frac{\partial \phi_1(\boldsymbol{x}^*)}{\partial y_1} & \ldots & \frac{\partial \phi_k(\boldsymbol{x}^*)}{\partial y_1} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \ldots & 1 & \frac{\partial \phi_1(\boldsymbol{x}^*)}{\partial y_{n-k}} & \ldots & \frac{\partial \phi_k(\boldsymbol{x}^*)}{\partial y_{n-k}} \end{bmatrix} \begin{bmatrix} \frac{\partial f(\boldsymbol{x}^*)}{\partial x_1} & \frac{\partial h_1(\boldsymbol{x}^*)}{\partial x_1} & \ldots & \frac{\partial h_k(\boldsymbol{x}^*)}{\partial x_1} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f(\boldsymbol{x}^*)}{\partial x_n} & \frac{\partial h_1(\boldsymbol{x}^*)}{\partial x_n} & \ldots & \frac{\partial h_k(\boldsymbol{x}^*)}{\partial x_n} \end{bmatrix} = \boldsymbol{O}$$

Proof (cont'd)

The matrix equation takes the form

$$\begin{bmatrix} \boldsymbol{I}_{n-k} & \frac{\partial \boldsymbol{\phi}(\boldsymbol{x}^*)}{\partial \boldsymbol{y}}^T \end{bmatrix} \begin{bmatrix} \nabla f(\boldsymbol{x}^*) & \nabla h_1(\boldsymbol{x}^*) & \dots & \nabla h_k(\boldsymbol{x}^*) \end{bmatrix} = \boldsymbol{O}_{(n-k)\times(k+1)}$$

meaning $\nabla f(\boldsymbol{x}^*), \nabla h_1(\boldsymbol{x}^*), \dots, \nabla h_k(\boldsymbol{x}^*)$ are all in $\mathrm{Null}(\boldsymbol{A})$, where

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{I}_{n-k} & \frac{\partial \boldsymbol{\phi}(\boldsymbol{x}^*)}{\partial \boldsymbol{y}}^T \end{bmatrix} \in \mathbb{R}^{(n-k)\times n}.$$

Note $\dim \mathrm{Null}(\boldsymbol{A}) = k$, so $\nabla f(\boldsymbol{x}^*), \nabla h_1(\boldsymbol{x}^*), \dots, \nabla h_k(\boldsymbol{x}^*)$ are linearly dependent. But $\nabla h_1(\boldsymbol{x}^*), \dots, \nabla h_k(\boldsymbol{x}^*)$ are linearly independent, since $\boldsymbol{x}^*$ is a regular point. Thus there exist $\lambda_1^*, \dots, \lambda_k^* \in \mathbb{R}$ s.t.

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

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 (8) 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

Contents

Active and Inactive Constraints

Let $\boldsymbol{x} \in \mathbb{R}^n$ and $n > k$. Consider

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & f(\boldsymbol{x}) \\ \text{s.t.} \quad & h_i(\boldsymbol{x}) = 0,\ i = 1, 2, \ldots, k \\ & g_j(\boldsymbol{x}) \leq 0,\ j = 1, 2, \ldots, m \end{aligned} \tag{ICP}$$

We do not assume it is a convex problem. Assume the domain is $\mathbb{R}^n$. The feasible set is

$$X = \{\boldsymbol{x} : h_i(\boldsymbol{x}) = 0,\ 1 \leq i \leq k;\ g_j(\boldsymbol{x}) \leq 0,\ 1 \leq j \leq m\}$$

Let $\boldsymbol{x}_0 \in X$. The j-th inequality constraint $g_j(\boldsymbol{x}) \leq 0$ is called active at $\boldsymbol{x}_0$ if $g_j(\boldsymbol{x}_0) = 0$, and inactive at $\boldsymbol{x}_0$ if $g_j(\boldsymbol{x}_0) < 0$. Denote by $J(\boldsymbol{x}_0)$ the set of indices of the active inequality constraints at $\boldsymbol{x}_0$,

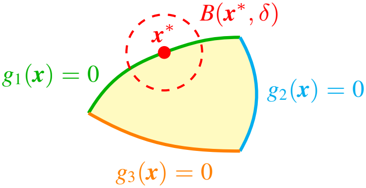
$$J(\boldsymbol{x}_0) = \{j : g_j(\boldsymbol{x}_0) = 0\}$$

By convention, equality constraints are considered active at all $\boldsymbol{x} \in X$.

Reduction to Equality Constrained Problem

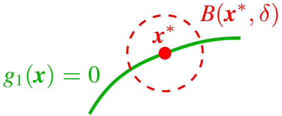
Suppose $\boldsymbol{x}^*$ is a local minimum of (ICP). It is the solution to

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & f(\boldsymbol{x}) \\ \text{s.t.} \quad & h_i(\boldsymbol{x}) = 0,\ i = 1, 2, \ldots, k \\ & g_j(\boldsymbol{x}) \leq 0,\ j = 1, 2, \ldots, m \\ & \boldsymbol{x} \in B(\boldsymbol{x}^*, \delta) \end{aligned}$$



for some small enough δ. It is equivalent to

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & f(\boldsymbol{x}) \\ \text{s.t.} \quad & h_i(\boldsymbol{x}) = 0,\ i = 1, 2, \ldots, k \\ & g_j(\boldsymbol{x}) = 0,\ j \in J(\boldsymbol{x}^*) \\ & \boldsymbol{x} \in B(\boldsymbol{x}^*, \delta) \end{aligned}$$



If it is known a priori which constraints are active at $\boldsymbol{x}^*$, we can find $\boldsymbol{x}^*$ by solving the above equality constrained problem.

Reduction to Equality Constrained Problem (cont'd)

A local minimum $\boldsymbol{x}^*$ of (ICP) is also a local minimum of the following

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & f(\boldsymbol{x}) \\ \text{s.t.} \quad & h_i(\boldsymbol{x}) = 0,\ i = 1, 2, \ldots, k \\ & g_j(\boldsymbol{x}) = 0,\ j \in J(\boldsymbol{x}^*) \end{aligned}$$

$\boldsymbol{x}^* \in X$ is a regular point if $\nabla h_i(\boldsymbol{x}^*), 1 \le i \le k$ and $\nabla g_j(\boldsymbol{x}^*), j \in J(\boldsymbol{x}^*)$ are linearly independent.

At a regular local minimum, Lagrange condition yields

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

Setting $\mu_j^* = 0$ for inactive constraints, i.e. $j \notin J(\boldsymbol{x}^*)$,

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

Karush-Kuhn-Tucker (KKT) Conditions

Theorem. If $\boldsymbol{x}^*$ is a local minimum of (ICP) and also a regular point, then there exist Lagrange multipliers[2] $\lambda_1^*, \ldots, \lambda_k^*, \mu_1^*, \ldots, \mu_m^* \in \mathbb{R}$ s.t. the following KKT conditions hold,

1. $\mu_j^* \geq 0, j = 1, 2, \ldots, m$
2. $\nabla f(\boldsymbol{x}^*) + \sum_{i=1}^{k} \lambda_i^* \nabla h_i(\boldsymbol{x}^*) + \sum_{j=1}^{m} \mu_j^* \nabla g_j(\boldsymbol{x}^*) = \mathbf{0}$
3. $\mu_j^* g_j(\boldsymbol{x}^*) = 0, j = 1, 2, \ldots, m$

Note. Condition 2 says $\nabla_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*) = \mathbf{0}$ for the Lagrangian

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

Note. Condition 3 is called complementary slackness condition, as it, together with 1 and $g_j(\boldsymbol{x}) \leq 0$, implies either $\mu_j^* = 0$ or $g_j(\boldsymbol{x}^*) = 0$.

[2] Sometimes also called KKT multipliers. Sometimes λ_i are called Lagrange multipliers while μ_j are called KKT multipliers.

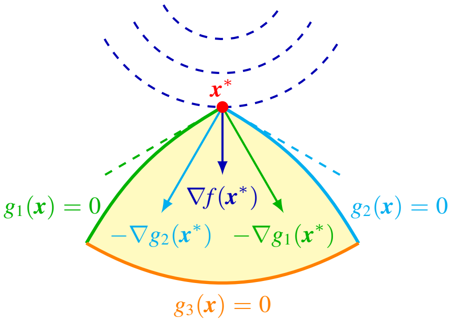
Geometric Interpretation

Let $\boldsymbol{x} \in \mathbb{R}^2$. Consider

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & f(\boldsymbol{x}) \\ \text{s.t.} \quad & g_j(\boldsymbol{x}) \le 0,\ j = 1, 2, 3 \end{aligned}$$

Suppose $\boldsymbol{x}^*$ is a local minimum and only g_1 and g_2 are active at $\boldsymbol{x}^*$. The KKT condition says $\mu_1^* \ge 0$, $\mu_2^* \ge 0$, $\mu_3^* = 0$ and

$$\nabla f(\boldsymbol{x}^*) = -\mu_1^* \nabla g_1(\boldsymbol{x}^*) - \mu_2^* \nabla g_2(\boldsymbol{x}^*)$$



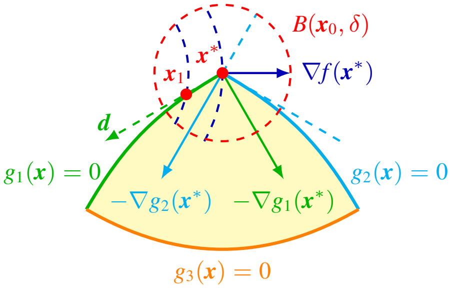
Geometric Interpretation (cont'd)

Why $\mu_j^* \geq 0$? Assume $\mu_2^* < 0$ and we show a contradiction.

- Let $\boldsymbol{d}$ be a tangent vector of $g_1(\boldsymbol{x}) = 0$ at $\boldsymbol{x}^*$, so $\boldsymbol{d} \perp \nabla g_1(\boldsymbol{x}^*)$.
- $\boldsymbol{d}^T \nabla g_2(\boldsymbol{x}^*) \neq 0$; otherwise, $\boldsymbol{d} \perp g_2(\boldsymbol{x}^*)$, so $\nabla g_1(\boldsymbol{x}^*) \parallel \nabla g_2(\boldsymbol{x}^*)$, contradicting the regularity of $\boldsymbol{x}^*$.
- Replacing $\boldsymbol{d}$ by $-\boldsymbol{d}$ if necessary, we can assume $\boldsymbol{d}^T \nabla g_2(\boldsymbol{x}^*) < 0$.
- Move along the curve $g_1(\boldsymbol{x}) = 0$ in the direction of $\boldsymbol{d}$ from $\boldsymbol{x}^*$ to $\boldsymbol{x}_1$.

$$\boldsymbol{d}^T \nabla f(\boldsymbol{x}^*) = \boldsymbol{d}^T [-\mu_1 \nabla g_1(\boldsymbol{x}^*) - \mu_2 \nabla g_2(\boldsymbol{x}^*)] = -\mu_2^* \boldsymbol{d}^T \nabla g_2(\boldsymbol{x}^*) < 0.$$

For a small move, $f(\boldsymbol{x}_1) < f(\boldsymbol{x}^*)$, contradicting minimality of $f(\boldsymbol{x}^*)$.



Appendix: Proof for $\boldsymbol{\mu} \geq \mathbf{0}$

Suppose $\mu^*_{j_0} < 0$ for some j_0. Let $J'(\boldsymbol{x}^*) = J(\boldsymbol{x}^*) \setminus \{j_0\}$, and S the set determined by all active constraints other than g_{j_0},

$$S = \{\boldsymbol{x} : h_i(\boldsymbol{x}) = 0, i = 1, 2, \ldots, k;\; g_j(\boldsymbol{x}) = 0,\ j \in J'(\boldsymbol{x}^*)\}$$

We will show we can move away from $\boldsymbol{x}^*$ on S so that feasibility is maintained but f decreases, contradicting the minimality of $\boldsymbol{x}^*$.

1. There exists a direction $\boldsymbol{d}_0$ tangent to S s.t. $\nabla g_{j_0}(\boldsymbol{x}^*)^T \boldsymbol{d}_0 < 0$
2. KKT then implies $\nabla f(\boldsymbol{x}^*)^T \boldsymbol{d}_0 < 0$
3. By Implicit Function Theorem, there exists a curve $\boldsymbol{x}(t) \subset S$ s.t. $\boldsymbol{x}(0) = \boldsymbol{x}^*$, $\boldsymbol{x}'(0) = \boldsymbol{d}_0$. Thus $g_j(\boldsymbol{x}(t)) = 0$ for $j \in J'(\boldsymbol{x}^*)$.
4. By continuity, $g_j(\boldsymbol{x}(t)) < 0$ for small t and $j \notin J(\boldsymbol{x}^*)$
5. By the chain rule,

$$\left.\frac{d}{dt} g_{j_0}(\boldsymbol{x}(t))\right|_{t=0} = \nabla g_{j_0}(\boldsymbol{x}^*)^T \boldsymbol{x}'(0) = \nabla g_{j_0}(\boldsymbol{x}^*)^T \boldsymbol{d}_0 < 0$$

   For small $t > 0$, $g_{j_0}(\boldsymbol{x}(t)) < g_{j_0}(\boldsymbol{x}^*) = 0$. Similarly, $f(\boldsymbol{x}(t)) < f(\boldsymbol{x}^*)$.

Proof for $\mu \geq \mathbf{0}$ (cont'd)

1. There exists a direction $\boldsymbol{d}_0$ tangent to S s.t. $\nabla g_{j_0}(\boldsymbol{x}^*)^T \boldsymbol{d}_0 < 0$

Proof.

- Let $\boldsymbol{A}$ be a matrix whose columns are $\nabla h_i(\boldsymbol{x}^*)$ and $\nabla g_j(\boldsymbol{x}^*)$, i.e.

$$\boldsymbol{A} = [\nabla h_i(\boldsymbol{x}^*), i = 1, \ldots, k;\ \nabla g_j(\boldsymbol{x}^*),\ j \in J'(\boldsymbol{x}^*)]$$

- The tangent "plane" (or more precisely, tangent space) of S at $\boldsymbol{x}^*$ is

$$T(\boldsymbol{x}^*) = \mathrm{Null}(\boldsymbol{A}^T) = \{\boldsymbol{d} : \nabla h_i(\boldsymbol{x}^*)^T \boldsymbol{d} = 0, \forall i;\ \nabla g_j(\boldsymbol{x}^*)^T \boldsymbol{d} = 0,\ j \in J'(\boldsymbol{x}^*)\}$$

- By regularity of $\boldsymbol{x}^*$,

$$\nabla g_{j_0}(\boldsymbol{x}^*) \notin \text{span}\{\nabla h_i(\boldsymbol{x}^*), \forall i; \nabla g_j(\boldsymbol{x}^*), j \in J'(\boldsymbol{x}^*)\} = \mathrm{Range}(\boldsymbol{A}) = \mathrm{Null}(\boldsymbol{A}^T)^{\perp}$$

  so there exists $\boldsymbol{d}_0 \in T(\boldsymbol{x}^*)$ s.t. $\nabla g_{j_0}(\boldsymbol{x}^*)^T \boldsymbol{d}_0 \neq 0$.

- Replacing $\boldsymbol{d}_0$ by $-\boldsymbol{d}_0 \in T(\boldsymbol{x}^*)$ if necessary, we have

$$\nabla g_{j_0}(\boldsymbol{x}^*)^T \boldsymbol{d}_0 < 0$$

Proof for $\boldsymbol{\mu} \geq \mathbf{0}$ (cont'd)

2. For $\boldsymbol{d}_0$ given by step 1, $\nabla f(\boldsymbol{x}^*)^T \boldsymbol{d}_0 < 0$

Proof.

- By KKT Conditions 2 and 3

$$\nabla f(\boldsymbol{x}^*) = -\sum_{i=1}^{k} \lambda_i^* \nabla h_i(\boldsymbol{x}^*) - \sum_{j \in J(\boldsymbol{x}^*)} \mu_j^* \nabla g_j(\boldsymbol{x}^*) - \sum_{j \notin J(\boldsymbol{x}^*)} \underbrace{\mu_j^* \nabla g_j(\boldsymbol{x}^*)}_{=\mathbf{0}}$$

- Since $\boldsymbol{d}_0 \in \mathrm{Null}(\boldsymbol{A}^T)$, $\mu_{j_0}^* < 0$,

$$\nabla f(\boldsymbol{x}^*)^T \boldsymbol{d}_0 = -\sum_{i=1}^{k} \lambda_i^* \underbrace{\nabla h_i(\boldsymbol{x}^*)^T \boldsymbol{d}_0}_{=0} - \sum_{j \in J'(\boldsymbol{x}^*)} \mu_j^* \underbrace{\nabla g_j(\boldsymbol{x}^*)^T \boldsymbol{d}_0}_{=0} - \mu_{j_0}^* \underbrace{\nabla g_{j_0}(\boldsymbol{x}^*)^T \boldsymbol{d}_0}_{<0}$$

$$< 0$$

Proof for $\boldsymbol{\mu} \geq \mathbf{0}$ (cont'd)

3. There exists a curve $\boldsymbol{x}(t) \subset S$ s.t. $\boldsymbol{x}(0) = \boldsymbol{x}^*$, $\boldsymbol{x}'(0) = \boldsymbol{d}_0$

Proof. For notational simplicity, denote $g_j, j \in J'(\boldsymbol{x}^*)$ by $h_{k+1}, \ldots, h_K$.

- Define $\tilde{\boldsymbol{x}}(t, \boldsymbol{\alpha}) = \boldsymbol{x}^* + t\boldsymbol{d}_0 + \sum_{i=1}^{K} \alpha_i \nabla h_i(\boldsymbol{x}^*)$ and

$$\tilde{h}_p(t, \boldsymbol{\alpha}) = h_p(\tilde{\boldsymbol{x}}(t, \boldsymbol{\alpha})), \quad 1 \leq p \leq K$$

- Note $\tilde{h}_p(0, \mathbf{0}) = h_p(\boldsymbol{x}^*) = 0$, and

$$\frac{\partial \tilde{h}_p(0, \mathbf{0})}{\partial \alpha_q} = \sum_{\ell=1}^{n} \frac{\partial h_p(\boldsymbol{x}^*)}{\partial x_\ell} \frac{\partial \tilde{x}_\ell(0, \mathbf{0})}{\partial \alpha_q} = \nabla h_p(\boldsymbol{x}^*)^T \nabla h_q(\boldsymbol{x}^*)$$

- By regularity of $\boldsymbol{x}^*$, $\boldsymbol{A} = [\nabla h_1(\boldsymbol{x}^*), \ldots, \nabla \tilde{h}_K(\boldsymbol{x}^*)]$ has rank K, so

$$\frac{\partial \tilde{\boldsymbol{h}}(0, \mathbf{0})}{\partial \boldsymbol{\alpha}} = \boldsymbol{A}^T \boldsymbol{A} \succ \boldsymbol{O}$$

and hence nonsingular.

Proof for $\boldsymbol{\mu} \geq \mathbf{0}$ (cont'd)

Proof (cont'd).

- By Implicit Function Theorem, there exists $\boldsymbol{\alpha}(t)$ for small $|t|$ s.t. $\boldsymbol{\alpha}(0) = \mathbf{0}$ and $\tilde{\boldsymbol{h}}(t, \boldsymbol{\alpha}(t)) = \boldsymbol{h}(\tilde{\boldsymbol{x}}(t, \alpha(t))) = \mathbf{0}$. Furthermore,

$$\boldsymbol{\alpha}'(0) = \left[\frac{\partial \tilde{\boldsymbol{h}}(0, \mathbf{0})}{\partial \boldsymbol{\alpha}}\right]^{-1} \frac{\partial \tilde{\boldsymbol{h}}(0, \mathbf{0})}{\partial t} = \left[\frac{\partial \tilde{\boldsymbol{h}}(0, \mathbf{0})}{\partial \boldsymbol{\alpha}}\right]^{-1} \begin{bmatrix} \nabla h_1(\boldsymbol{x}^*)^T \boldsymbol{d}_0 \\ \vdots \\ \nabla h_K(\boldsymbol{x}^*)^T \boldsymbol{d}_0 \end{bmatrix} = \mathbf{0}$$

  since $\boldsymbol{d}_0 \in T(\boldsymbol{x}^*) = \mathrm{Null}(\boldsymbol{A}^T)$.

- Let $\boldsymbol{x}(t) = \tilde{\boldsymbol{x}}(t, \boldsymbol{\alpha}(t))$. Then $\boldsymbol{h}(\boldsymbol{x}(t)) = \mathbf{0}$, so $\boldsymbol{x}(t) \subset S$.

$$\boldsymbol{x}'(0) = \boldsymbol{d}_0 + \sum_{i=1}^{K} \alpha_i'(0) \nabla h_i(\boldsymbol{x}^*) = \boldsymbol{d}_0$$

Sufficiency of KKT Conditions for Convex Problems

Theorem. For convex (ICP), i.e. f and g_j are convex, and h_i are affine, if there exist $\lambda_1^*, \ldots, \lambda_k^*$ and $\mu_1^*, \ldots, \mu_m^*$ s.t. the KKT conditions are satisfied at a feasible $\boldsymbol{x}^* \in X$, then $\boldsymbol{x}^*$ is a global minimum of (ICP).

Note. The previous necessary conditions assume $\boldsymbol{x}^*$ is regular point. The sufficient conditions here assume convexity but not regularity.

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

1. By the KKT conditions,

$$\nabla f(\boldsymbol{x}^*)^T(\boldsymbol{x}-\boldsymbol{x}^*) = -\sum_i \lambda^* \nabla h_i(\boldsymbol{x}^*)^T(\boldsymbol{x}-\boldsymbol{x}^*) - \sum_{j \in J(\boldsymbol{x}^*)} \mu_j^* \nabla g_j(\boldsymbol{x}^*)^T(\boldsymbol{x}-\boldsymbol{x}^*)$$

   It suffices to show $\nabla h_i(\boldsymbol{x}^*)^T(\boldsymbol{x} - \boldsymbol{x}^*) = 0$ and $\nabla g_j(\boldsymbol{x}^*)^T(\boldsymbol{x} - \boldsymbol{x}^*) \leq 0$.

2. Since $h_i(\boldsymbol{x}) = \boldsymbol{a}_i^T \boldsymbol{x} - b_i$ is affine, and $h_i(\boldsymbol{x}) = h(\boldsymbol{x}^*) = 0$ by feasibility,

$$\nabla h_i(\boldsymbol{x}^*)^T(\boldsymbol{x} - \boldsymbol{x}^*) = \boldsymbol{a}_i^T(\boldsymbol{x} - \boldsymbol{x}^*) = h_i(\boldsymbol{x}) - h(\boldsymbol{x}^*) = 0$$

3. For $j \in J(\boldsymbol{x}^*)$, $g_j(\boldsymbol{x}^*) = 0$ and $g_j(\boldsymbol{x}) \leq 0$. By the convexity of g_j,

$$\nabla g_j(\boldsymbol{x}^*)^T(\boldsymbol{x} - \boldsymbol{x}^*) \leq g_j(\boldsymbol{x}) - g_j(\boldsymbol{x}^*) \leq 0$$