CS257 Linear and Convex Optimization

Lecture 12

Bo Jiang

John Hopcroft Center for Computer Science
Shanghai Jiao Tong University

November 23, 2020

Recap: Damped Newton's Method

initialization $\boldsymbol{x} \leftarrow \boldsymbol{x}_0 \in \mathbb{R}^n$
while $\|\nabla f(\boldsymbol{x})\| > \delta$ **do**
  $\boldsymbol{d} \leftarrow -[\nabla^2 f(\boldsymbol{x})]^{-1} \nabla f(\boldsymbol{x})$
  find t with backtracking line search
  $\boldsymbol{x} \leftarrow \boldsymbol{x} + t\boldsymbol{d}$
end while
return $\boldsymbol{x}$

Pure vs damped Newton's method (under appropriate conditions).

- Pure Newton's method has fast local convergence with no global guarantee

$$\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\| \leq \frac{M}{2m} \|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2$$

- Damped Newton's method guarantees global convergence with a slow damped phase and a fast pure phase.

Recap: Equality Constrained Optimization Problem

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

where $\boldsymbol{A}^T = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_k) \in \mathbb{R}^{n \times k}$, $\boldsymbol{b} \in \mathbb{R}^k$, f is differentiable and convex.

Assume the feasible set $X = \{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}\} \neq \emptyset$, and the constraints are independent, i.e. $\operatorname{rank} \boldsymbol{A} = k$.

First-order optimality condition. $\boldsymbol{x}^* \in X$ is optimal iff

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

or equivalently,

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

i.e.

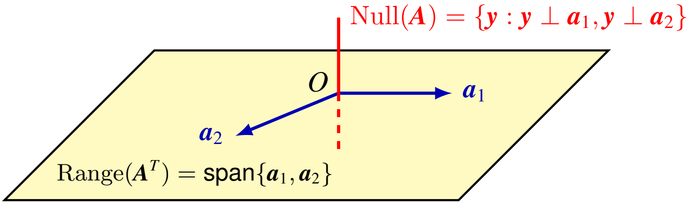
$$\nabla f(\boldsymbol{x}^*) + \boldsymbol{A}^T \boldsymbol{\lambda}^* = \nabla f(\boldsymbol{x}^*) + \sum_{i=1}^{k} \lambda_i^* \boldsymbol{a}_i = \boldsymbol{0}, \quad \text{for some } \boldsymbol{\lambda}^* \in \mathbb{R}^k$$

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

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.

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



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

Lagrange Condition

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 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 gradient[2] w.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 such.

[2] We use a similar notation $\nabla_{\boldsymbol{d}} f(\boldsymbol{x}) = \nabla f(\boldsymbol{x})^T \boldsymbol{d}$ to denote the directional derivative of f along the direction $\boldsymbol{d}$. The context should make it clear which is which.

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

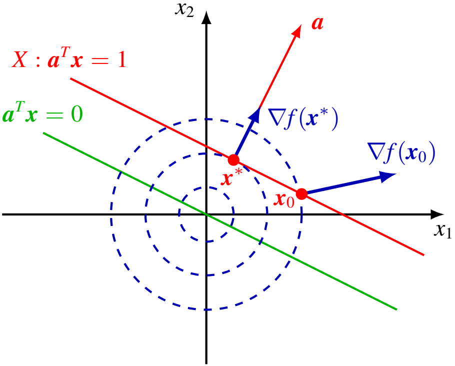
$$\boldsymbol{a} = (1,2)^T$$

- gradient

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

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

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



Example

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

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)

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

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

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

- Solve for $\boldsymbol{x}, \boldsymbol{\lambda}$ e.g. by 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

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

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

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

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

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

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

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

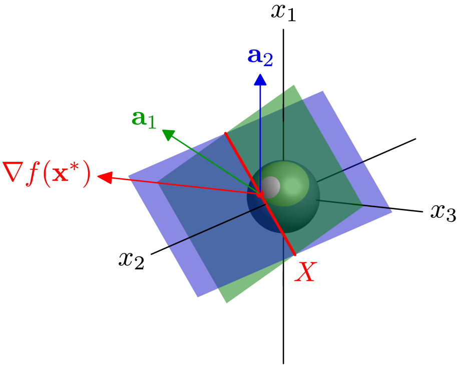
Example (cont'd)

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

- 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 $\boldsymbol{x}^*$,

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

Equality Constrained Convex QP

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T\boldsymbol{Q}\boldsymbol{x} + \boldsymbol{g}^T\boldsymbol{x} + c \\ \text{s.t.} \quad & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} \end{aligned} \tag{$\star$}$$

where $\boldsymbol{Q} \in \mathbb{R}^n$, $\boldsymbol{Q} \succeq \boldsymbol{O}$, $\boldsymbol{A} \in \mathbb{R}^{k\times n}$, $\operatorname{rank} \boldsymbol{A} = k$.

Note. This is the basis for an extension of Newton's method to equality constrained problems.

- The Lagrangian is

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

- The Lagrange condition is

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

  This is the KKT system of the problem ($\star$). The coefficient matrix $\boldsymbol{K} = \begin{bmatrix} \boldsymbol{Q} & \boldsymbol{A}^T \\ \boldsymbol{A} & \boldsymbol{O} \end{bmatrix}$ is called the KKT matrix.

Solving KKT System When $\boldsymbol{Q} \succ \boldsymbol{O}$

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

1. Solving for $\boldsymbol{x}$ in term of $\boldsymbol{\lambda}$ from $\boldsymbol{Q}\boldsymbol{x} + \boldsymbol{g} + \boldsymbol{A}^T\boldsymbol{\lambda} = \boldsymbol{0}$,

$$\boldsymbol{x} = -\boldsymbol{Q}^{-1}\boldsymbol{g} - \boldsymbol{Q}^{-1}\boldsymbol{A}^T\boldsymbol{\lambda}$$

2. Substituting into $\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b} = \boldsymbol{0}$,

$$-\boldsymbol{A}\boldsymbol{Q}^{-1}\boldsymbol{g} - \boldsymbol{A}\boldsymbol{Q}^{-1}\boldsymbol{A}^T\boldsymbol{\lambda} = \boldsymbol{b}$$

3. Since $\boldsymbol{A}\boldsymbol{Q}^{-1}\boldsymbol{A}^T \succ \boldsymbol{O}$ (why?), solving for $\boldsymbol{\lambda}$,

$$\boldsymbol{\lambda} = -[\boldsymbol{A}\boldsymbol{Q}^{-1}\boldsymbol{A}^T]^{-1}[\boldsymbol{A}\boldsymbol{Q}^{-1}\boldsymbol{g} + \boldsymbol{b}]$$

4. Plugging into step 1,

$$\boldsymbol{x} = -\boldsymbol{Q}^{-1}\boldsymbol{g} + \boldsymbol{Q}^{-1}\boldsymbol{A}^T[\boldsymbol{A}\boldsymbol{Q}^{-1}\boldsymbol{A}^T]^{-1}[\boldsymbol{A}\boldsymbol{Q}^{-1}\boldsymbol{g} + \boldsymbol{b}]$$

Note. We can also use block Gaussian elimination (cf. slide 9).

Unsolvable KKT System

Example.

$$\min_{x_1,x_2} \quad f(x_1,x_2) = \frac{1}{2}x_2^2 + x_1$$
$$\text{s.t.} \quad x_2 = 0$$

This is a convex QP with

$$\boldsymbol{Q} = \text{diag}\{0,1\}, \quad \boldsymbol{g} = (1,0)^T, \quad \boldsymbol{A} = (0,1), \quad b = 0$$

The KKT system is

$$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \lambda \end{bmatrix} = \begin{bmatrix} -1 \\ 0 \\ 0 \end{bmatrix}$$

which has no solution, since $0 \cdot x_1 + 0 \cdot x_2 + 0 \cdot \lambda \neq -1$.

Note $f^* = -\infty$ for this problem.

Unsolvable KKT System (cont'd)

If the KKT system has no solution, then the problem ($\star$) is either infeasible or unbounded below.

- KKT system has no solution iff

$$\begin{bmatrix} -\boldsymbol{g} \\ \boldsymbol{b} \end{bmatrix} \notin \mathrm{Range}(\boldsymbol{K}) = \mathrm{Range}(\boldsymbol{K}^T) = \mathrm{Null}(\boldsymbol{K})^{\perp}$$

- There exists $\begin{bmatrix} \boldsymbol{v} \\ \boldsymbol{w} \end{bmatrix} \in \mathrm{Null}(\boldsymbol{K})$ s.t. $\begin{bmatrix} -\boldsymbol{g} \\ \boldsymbol{b} \end{bmatrix}^T \begin{bmatrix} \boldsymbol{v} \\ \boldsymbol{w} \end{bmatrix} \neq 0$, i.e.

$$\boldsymbol{Q}\boldsymbol{v} + \boldsymbol{A}^T\boldsymbol{w} = \boldsymbol{0}, \quad \boldsymbol{A}\boldsymbol{v} = \boldsymbol{0}, \quad -\boldsymbol{g}^T\boldsymbol{v} + \boldsymbol{b}^T\boldsymbol{w} \neq 0$$

- If $\boldsymbol{x}_0$ is feasible, then $\boldsymbol{x}_0 + t\boldsymbol{v}$ is feasible for any $t \in \mathbb{R}$,

$$\begin{aligned} f(\boldsymbol{x}_0 + t\boldsymbol{v}) &= f(\boldsymbol{x}_0) + t(\boldsymbol{x}_0^T\boldsymbol{Q}\boldsymbol{v} + \boldsymbol{g}^T\boldsymbol{v}) + \frac{1}{2}t^2\boldsymbol{v}^T\boldsymbol{Q}\boldsymbol{v} \\ &= f(\boldsymbol{x}_0) + t(-\boldsymbol{x}_0^T\boldsymbol{A}^T\boldsymbol{w} + \boldsymbol{g}^T\boldsymbol{v}) - \frac{1}{2}t^2\boldsymbol{w}^T\boldsymbol{A}\boldsymbol{v} \quad (\text{use } \boldsymbol{Q}\boldsymbol{v} = -\boldsymbol{A}^T\boldsymbol{w}) \\ &= f(\boldsymbol{x}_0) - t(\boldsymbol{b}^T\boldsymbol{w} - \boldsymbol{g}^T\boldsymbol{v}) \quad (\text{use } \boldsymbol{A}\boldsymbol{v} = \boldsymbol{0} \text{ and } \boldsymbol{A}\boldsymbol{x}_0 = \boldsymbol{b}) \end{aligned}$$

which goes to $-\infty$, as $t \to \mathrm{sign}(\boldsymbol{b}^T\boldsymbol{w} - \boldsymbol{g}^T\boldsymbol{v}) \cdot \infty$.

Nonsingularity of KKT Matrix

If the KKT matrix $\boldsymbol{K}$ is nonsingular, then the KKT system has a unique solution, which is optimal.

Recall $\boldsymbol{Q} \succeq \boldsymbol{O}$ and $\operatorname{rank} \boldsymbol{A} = k$. The following conditions are equivalent

1. $\boldsymbol{K}$ is nonsingular
2. $\mathrm{Null}(\boldsymbol{Q}) \cap \mathrm{Null}(\boldsymbol{A}) = \{\boldsymbol{0}\}$, i.e. $\boldsymbol{Q}$ and $\boldsymbol{A}$ have no nontrivial common nullspace, i.e. $\boldsymbol{Ax} = \boldsymbol{0}$, $\boldsymbol{Qx} = \boldsymbol{0}$ only have the trivial solution $\boldsymbol{x} = \boldsymbol{0}$.
3. $\boldsymbol{Ax} = \boldsymbol{0}, \boldsymbol{x} \neq \boldsymbol{0} \implies \boldsymbol{x}^T\boldsymbol{Qx} > 0$, i.e. $\boldsymbol{Q}$ is positive definite on the nullspace of $\boldsymbol{A}$.
4. $\boldsymbol{F}^T\boldsymbol{QF} \succ \boldsymbol{O}$ for any $\boldsymbol{F} \in \mathbb{R}^{n \times (n-k)}$ s.t. $\mathrm{Range}(\boldsymbol{F}) = \mathrm{Null}(\boldsymbol{A})$, i.e. the columns of $\boldsymbol{F}$ are linearly independent solutions of $\boldsymbol{Ax} = \boldsymbol{0}$.

In particular, if $\boldsymbol{Q} \succ \boldsymbol{O}$, then $\boldsymbol{K}$ is nonsingular (by 3).

Proof

We show $1 \Rightarrow 2 \Rightarrow 3 \Rightarrow 4 \Rightarrow 1$.

- $(1 \Rightarrow 2)$. If $\mathbf{0} \neq \boldsymbol{x} \in \mathrm{Null}(\boldsymbol{Q}) \cap \mathrm{Null}(\boldsymbol{A})$, then

$$\begin{bmatrix} \boldsymbol{Q} & \boldsymbol{A}^T \\ \boldsymbol{A} & \boldsymbol{O} \end{bmatrix} \begin{bmatrix} \boldsymbol{x} \\ \mathbf{0} \end{bmatrix} = \mathbf{0}$$

  contradicting the nonsingularity of $\boldsymbol{K}$.
- $(2 \Rightarrow 3.)$ Assume $\boldsymbol{A}\boldsymbol{x} = \mathbf{0}$ and $\boldsymbol{x}^T\boldsymbol{Q}\boldsymbol{x} = 0$. We show $\boldsymbol{x} = \mathbf{0}$. Since $\boldsymbol{Q} \succeq \boldsymbol{O}$, $\boldsymbol{x}^T\boldsymbol{Q}\boldsymbol{x} = 0$ iff $\boldsymbol{Q}\boldsymbol{x} = \mathbf{0}$ [3]. By 2, $\boldsymbol{A}\boldsymbol{x} = \mathbf{0}$ and $\boldsymbol{Q}\boldsymbol{x} = \mathbf{0}$ implies $\boldsymbol{x} = \mathbf{0}$.

[3] Proof of necessity: Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ be an orthonormal eigenbasis of $\boldsymbol{Q}$ and $\boldsymbol{Q}\boldsymbol{x}_i = \nu_i\boldsymbol{x}_i$. Then $\boldsymbol{Q} = \sum_{i=1}^{n} \nu_i\boldsymbol{x}_i\boldsymbol{x}_i^T$. Note $\nu_i \geq 0$, since $\boldsymbol{Q} \succeq \boldsymbol{O}$. Then

$$0 = \boldsymbol{x}^T\boldsymbol{Q}\boldsymbol{x} = \boldsymbol{x}^T \left(\sum_{i=1}^{n} \nu_i\boldsymbol{x}_i\boldsymbol{x}_i^T \right) \boldsymbol{x} = \sum_{i=1}^{n} \nu_i \|\boldsymbol{x}_i^T\boldsymbol{x}\|^2 \implies \boldsymbol{x}_i^T\boldsymbol{x} = 0 \text{ if } \nu_i > 0$$

i.e. either $\nu_i = 0$ or $\boldsymbol{x}_i^T\boldsymbol{x} = 0$. Thus $\boldsymbol{Q}\boldsymbol{x} = \sum_{i=1}^{n} \nu_i\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{x} = \mathbf{0}$.

Proof (cont'd)

- ($3 \Rightarrow 4.$) $\operatorname{rank} \boldsymbol{F} = \dim \operatorname{Null}(\boldsymbol{A}) = n - \operatorname{rank} \boldsymbol{A} = n - k$, so $\boldsymbol{F}$ has full column rank. If $\boldsymbol{z} \neq \boldsymbol{0}$, then $\boldsymbol{x} = \boldsymbol{F}\boldsymbol{z} \neq \boldsymbol{0}$ and $\boldsymbol{x} \in \operatorname{Range}(\boldsymbol{F}) = \operatorname{Null}(\boldsymbol{A})$. By 3, $\boldsymbol{z}^T(\boldsymbol{F}^T\boldsymbol{Q}\boldsymbol{F})\boldsymbol{z} = \boldsymbol{x}^T\boldsymbol{Q}\boldsymbol{x} > 0$.
- ($4 \Rightarrow 1.$) To show $\boldsymbol{K}$ is nonsingular, we assume

$$\begin{bmatrix} \boldsymbol{Q} & \boldsymbol{A}^T \\ \boldsymbol{A} & \boldsymbol{O} \end{bmatrix} \begin{bmatrix} \boldsymbol{v} \\ \boldsymbol{w} \end{bmatrix} = \boldsymbol{0}$$

and show $\boldsymbol{v} = \boldsymbol{0}$ and $\boldsymbol{w} = \boldsymbol{0}$. Note $\boldsymbol{v} \in \operatorname{Null}(\boldsymbol{A})$ and

$$\boldsymbol{Q}\boldsymbol{v} = -\boldsymbol{A}^T\boldsymbol{w} \implies \boldsymbol{v}^T\boldsymbol{Q}\boldsymbol{v} = -\boldsymbol{v}^T\boldsymbol{A}^T\boldsymbol{w} = -(\boldsymbol{A}\boldsymbol{v})^T\boldsymbol{w} = 0$$

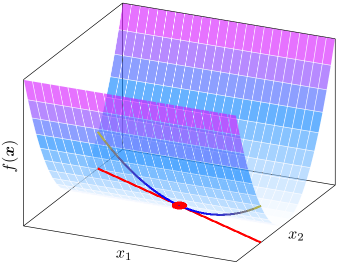
Let $\boldsymbol{F} \in \mathbb{R}^{n\times(n-k)}$ be a matrix whose columns consist of a basis of $\operatorname{Null}(\boldsymbol{A})$. Then $\operatorname{Range}(\boldsymbol{F}) = \operatorname{Null}(\boldsymbol{A})$ and $\boldsymbol{v} = \boldsymbol{F}\boldsymbol{z}$ for some $\boldsymbol{z}$. Now

$$0 = \boldsymbol{v}^T\boldsymbol{Q}\boldsymbol{v} = \boldsymbol{z}^T\boldsymbol{F}^T\boldsymbol{Q}\boldsymbol{F}\boldsymbol{z}$$

By 4, $\boldsymbol{z} = \boldsymbol{0}$, so $\boldsymbol{v} = \boldsymbol{F}\boldsymbol{z} = \boldsymbol{0}$. Then $\boldsymbol{A}^T\boldsymbol{w} = -\boldsymbol{Q}\boldsymbol{v} = \boldsymbol{0}$. Since $\boldsymbol{A}^T$ has full column rank, $\boldsymbol{w} = \boldsymbol{0}$.

Example

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



Trivial with solution $x_1^* = b, x_2^* = 0$.

But let's check the condition on slide 15.

$$\boldsymbol{Q} = \text{diag}\{0,1\}, \quad \boldsymbol{A} = (1,2)$$

Let $\boldsymbol{F} = (2,-1)^T$. Then $\mathrm{Range}(\boldsymbol{F}) = \mathrm{Null}(\boldsymbol{A})$, and

$$\boldsymbol{F}^T\boldsymbol{Q}\boldsymbol{F} = [1] \succ \boldsymbol{O}$$

By 4 of slide 15, the KKT matrix in nonsingular, so $\exists$ a unique solution.

Note. The unconstrained problem $\min_{\boldsymbol{x}} f(\boldsymbol{x})$ has infinitely many solutions. But this does not prevent the constrained problem from having a unique solution, as $\boldsymbol{Q} \succ \boldsymbol{O}$ on $\mathrm{Null}(\boldsymbol{A})$ (see 3 on slide 15).

Newton Direction for Equality Constrained Problem

Consider the second-order Taylor approximation for f at a feasible $\boldsymbol{x}_k$,

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

Using $\boldsymbol{A}\boldsymbol{x}_k = \boldsymbol{b}$,

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

KKT system for this quadratic problem is

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

The Newton direction $\boldsymbol{d}_k$ is given by the solution to the KKT system.
We will assume the KKT matrix is nonsingular (cf. slide 15)

Newton's Method for Equality Constrained Problem

1: initialization $\boldsymbol{x} \leftarrow \boldsymbol{x}_0 \in X$ $\triangleright$ $\boldsymbol{x}_0$ is feasible, i.e. $\boldsymbol{A}\boldsymbol{x}_0 = \boldsymbol{b}$
2: **repeat**
3: Compute Newton's direction $\boldsymbol{d}$ by solving

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

4: $t \leftarrow 1$ $\triangleright$ backtracking line search on lines 4-7
5: **while** $f(\boldsymbol{x} + t\boldsymbol{d}) > f(\boldsymbol{x}) + \alpha t \nabla f(\boldsymbol{x})^T \boldsymbol{d}$ **do**
6: $t \leftarrow \beta t$
7: **end while**
8: $\boldsymbol{x} \leftarrow \boldsymbol{x} + t\boldsymbol{d}$
9: **until** $\|\boldsymbol{d}\| \leq \delta$
10: **return** $\boldsymbol{x}$

Note. We cannot use $\|\nabla f(\boldsymbol{x})\| \leq \delta$ as stopping criterion now, as $\nabla f(\boldsymbol{x}^*) = \boldsymbol{0}$ no longer holds in general. [BV] uses $\sqrt{\boldsymbol{d}^T \nabla^2 f(\boldsymbol{x}) \boldsymbol{d}} \leq \delta$.

Note. This is called a feasible descent method, since all $\boldsymbol{x}_k$ are feasible and $f(\boldsymbol{x}_{k+1}) < f(\boldsymbol{x}_k)$ unless $\boldsymbol{x}_k$ is optimal.

Newton's Method and Constraint Elimination

Let $\boldsymbol{F} \in \mathbb{R}^{n\times(n-k)}$ be a matrix whose columns are linearly independent solutions to $\boldsymbol{A}\boldsymbol{x} = \mathbf{0}$. For a fixed feasible $\tilde{\boldsymbol{x}} \in \boldsymbol{X}$,

$$\boldsymbol{X} = \{\boldsymbol{x} : \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}\} = \{\tilde{\boldsymbol{x}} + \boldsymbol{F}\boldsymbol{z} : \boldsymbol{z} \in \mathbb{R}^{n-k}\}$$

Constrained problem reduces to unconstrained problem by $\boldsymbol{x} = \boldsymbol{F}\boldsymbol{z} + \tilde{\boldsymbol{x}}$,

$$\begin{cases} \min\limits_{\boldsymbol{x}} & f(\boldsymbol{x}) \\ \text{s.t.} & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} \end{cases} \iff \min_{\boldsymbol{z}} \; g(\boldsymbol{z}) = f(\boldsymbol{F}\boldsymbol{z} + \tilde{\boldsymbol{x}})$$

Note (slides 8 and 17 of Lecture 2),

$$\nabla g(\boldsymbol{z}) = \boldsymbol{F}^T \nabla f(\boldsymbol{F}\boldsymbol{z} + \tilde{\boldsymbol{x}}), \quad \nabla^2 g(\boldsymbol{z}) = \boldsymbol{F}^T \nabla^2 f(\boldsymbol{F}\boldsymbol{z} + \tilde{\boldsymbol{x}})\boldsymbol{F} \tag{$\dagger$}$$

Apply Newton's method to both problems with initial points $\boldsymbol{x}_0$ and $\boldsymbol{z}_0$. If $\boldsymbol{x}_0 = \boldsymbol{F}\boldsymbol{z}_0 + \tilde{\boldsymbol{x}}$, we show by induction that $\boldsymbol{x}_k = \boldsymbol{F}\boldsymbol{z}_k + \tilde{\boldsymbol{x}}$, so Newton's method converges for the constrained problem if it does for the unconstrained problem.

Proof

The Newton direction $\Delta \boldsymbol{x}_k$ for the constrained problem satisfies

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

1. By the induction hypothesis $\boldsymbol{x}_k = \boldsymbol{F}\boldsymbol{z}_k + \tilde{\boldsymbol{x}}$ and (†) on the previous slide
$$\boldsymbol{F}^T \nabla^2 f(\boldsymbol{x}_k) \boldsymbol{F} = \nabla^2 g(\boldsymbol{z}_k), \quad \boldsymbol{F}^T \nabla f(\boldsymbol{x}_k) = \nabla g(\boldsymbol{z}_k)$$
2. By 1 above and 3 on slide 15, the KKT matrix in 1 in nonsingular iff $\nabla^2 g(\boldsymbol{z}_k) \succ \boldsymbol{O}$, so $\Delta \boldsymbol{x}_k$ is well-defined iff the Newton direction $\Delta \boldsymbol{z}_k = -[\nabla^2 g(\boldsymbol{z}_k)]^{-1} \nabla g(\boldsymbol{z}_k)$ is well-defined.
3. Since $\mathrm{Null}(\boldsymbol{A}) = \mathrm{Range}(\boldsymbol{F})$, $\boldsymbol{A}\Delta \boldsymbol{x}_k = \boldsymbol{0} \iff \Delta \boldsymbol{x}_k = \boldsymbol{F}\boldsymbol{u}$ for some $\boldsymbol{u}$.
4. Plugging $\Delta \boldsymbol{x}_k = \boldsymbol{F}\boldsymbol{u}$ into the first KKT equation,
$$\nabla^2 f(\boldsymbol{x}_k) \boldsymbol{F}\boldsymbol{u} + \boldsymbol{A}^T \boldsymbol{\lambda}_k = -\nabla f(\boldsymbol{x}_k)$$
5. Pre-multiplying by $\boldsymbol{F}^T$,
$$\boldsymbol{F}^T \nabla^2 (\boldsymbol{x}_k) \boldsymbol{F}\boldsymbol{u} + (\boldsymbol{A}\boldsymbol{F})^T \boldsymbol{\lambda}_k = -\boldsymbol{F}^T \nabla f(\boldsymbol{x}_k)$$

Proof (cont'd)

6. Since the columns of $\boldsymbol{F}$ are solutions to $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{0}$, $\boldsymbol{A}\boldsymbol{F} = \boldsymbol{O}$
7. By 1, 5 and 6,

$$\nabla^2 g(z_k)\boldsymbol{u} = -\nabla g(z_k) \implies \boldsymbol{u} = -[\nabla^2 g(z_k)]^{-1}\nabla g(z_k) = \Delta z_k$$

so

$$\Delta \boldsymbol{x}_k = \boldsymbol{F}\boldsymbol{u} = \boldsymbol{F}\Delta z_k$$

8. By 7, backtracking line search gives the same step size t_k, since

$$f(\boldsymbol{x}_k + t\Delta \boldsymbol{x}_k) = f(\boldsymbol{F}(z_k + t\Delta z_k) + \tilde{\boldsymbol{x}}) = g(z_k + t\Delta z_k)$$

9. By 7, 8, and the induction hypothesis $\boldsymbol{x}_k = \boldsymbol{F}z_k + \tilde{\boldsymbol{x}}$,

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + t_k\Delta \boldsymbol{x}_k = \boldsymbol{F}z_k + \tilde{\boldsymbol{x}} + t_k\boldsymbol{F}\Delta z_k = \boldsymbol{F}(z_k + t_k\Delta z_k) + \tilde{\boldsymbol{x}} = \boldsymbol{F}z_{k+1} + \tilde{\boldsymbol{x}}$$

completing the induction.