CS257 Linear and Convex Optimization

Lecture 11

Bo Jiang

John Hopcroft Center for Computer Science
Shanghai Jiao Tong University

November 16, 2020

Recap: Line Search

Exact line search.

$$t_k = \arg\min_s f(\boldsymbol{x}_k - s\nabla f(\boldsymbol{x}_k))$$

Backtracking line search (Armijo's rule).

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}_k - t_k\nabla f(\boldsymbol{x}_k)) \geq \alpha t_k \|\nabla f(\boldsymbol{x}_k)\|_2^2$$

initialization $\boldsymbol{x} \leftarrow \boldsymbol{x}_0 \in \mathbb{R}^n$
while $\|\nabla f(\boldsymbol{x})\| > \delta$ **do**
    choose direction $\boldsymbol{d}$ ▷ $\boldsymbol{d} = -\nabla f(\boldsymbol{x})$ for gradient descent
    $t \leftarrow t_0$
    while $f(\boldsymbol{x} + t\boldsymbol{d}) > f(\boldsymbol{x}) + \alpha t \nabla f(\boldsymbol{x})^T \boldsymbol{d}$ **do**
        $t \leftarrow \beta t$
    end while
    $\boldsymbol{x} \leftarrow \boldsymbol{x} + t\boldsymbol{d}$
end while
return $\boldsymbol{x}$

Recap: Convergence of Gradient Descent

For m-strongly convex and L-smooth f with minimum $\boldsymbol{x}^*$

- gradient descent with constant step size $t \in (0, \frac{1}{L}]$ satisfies

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*) \leq \frac{L(1 - mt)^k}{m}[f(\boldsymbol{x}_0) - f(\boldsymbol{x}^*)]$$

- gradient descent with exact line search satisfies

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*) \leq \left(1 - \frac{m}{L}\right)^k [f(\boldsymbol{x}_0) - f(\boldsymbol{x}^*)]$$

- gradient descent with backtracking line search satisfies

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*) \leq c^k[f(\boldsymbol{x}_0) - f(\boldsymbol{x}^*)]$$

  where

$$c = 1 - \min\left\{2m\alpha t_0, \frac{4m\beta\alpha(1-\alpha)}{L}\right\}$$

Recap: Newton's Method

Newton's method for solving optimization problem $\min_{\boldsymbol{x}} f(\boldsymbol{x})$

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - [\nabla^2 f(\boldsymbol{x}_k)]^{-1} \nabla f(\boldsymbol{x}_k)$$

Newton's method for solving $\boldsymbol{g}(\boldsymbol{x}) = \boldsymbol{0}$

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - [D\boldsymbol{g}(\boldsymbol{x}_k)]^{-1} \boldsymbol{g}(\boldsymbol{x}_k)$$

Connection. First-order optimality condition

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

Today

- Analysis of Newton's method
- Damped Newton's method
- Equality Constrained Optimization

Contents

Convergence of Newton's Method

Example. Consider the minimization of $f(x) = \sqrt{1+x^2}$.

$$f'(x) = \frac{x}{\sqrt{1+x^2}}, \quad f''(x) = \frac{1}{(1+x^2)^{3/2}}$$

The Newton direction is

$$d_k = -f'(x_k)/f''(x_k) = -x_k - x_k^3$$

The Newton step is

$$x_{k+1} = x_k + d_k = -x_k^3$$

Note $x_k \to x^* = 0$ iff $|x_0| < 1$. When $|x_0| > 1$, x_k diverges, and

$$f(x_{k+1}) > f(x_k)$$

In general, Newton's method does not guarantee global convergence. When it does converge, the convergence is usually very fast.

Convergence Analysis: 1D Case

Theorem. If f is m-strongly convex, f'' is M-Lipschitz continuous, and x^* is a minimum of f, then the sequence $\{x_k\}$ produced by Newton's method satisfies

$$|x_{k+1} - x^*| \leq \frac{M}{2m}|x_k - x^*|^2$$

Notes. Let $\xi_k = \frac{M}{2m}|x_k - x^*|$. The above inequality becomes $\xi_{k+1} \leq \xi_k^2$.

- If $\xi_k = 10^{-p}$, then $\xi_{k+1} \leq 10^{-2p}$, the number of significant digits doubles in each iteration!
- If $\xi_0 < 1$ i.e. $|x_0 - x^*| < \frac{2m}{M}$, then $\xi_k \leq \xi_0^{2^k}$ converges to 0 extremely fast. The number of iterations to ensure $\xi_k \leq \epsilon$ is $k \geq \log_2 \log_{\frac{1}{\xi_0}} \frac{1}{\epsilon}$. For $\epsilon = 10^{-p}$, $k \geq \log_2 p + \log_2 \log_{\frac{1}{\xi_0}} 10$, only logarithmic in the number of digits. Very few iterations are required!
- This theorem is a local convergence result. Fast convergence if x_0 is close enough to x^*, i.e. $|x_0 - x^*| < \frac{2m}{M}$. No guarantee if $|x_0 - x^*|$ is large.

Proof: 1D Case

$$
\begin{aligned}
& |x_{k+1} - x^*| \\
= & |x_k - x^* - [f''(x_k)]^{-1} f'(x_k)| && \text{Newton step} \\
= & |f''(x_k)|^{-1} \cdot |f'(x^*) - f'(x_k) - f''(x_k)(x^* - x_k)| && f'(x^*) = 0 \\
= & \frac{|x_k - x^*|}{|f''(x_k)|} \cdot \left| \int_0^1 [f''(x_k + t(x^* - x_k)) - f''(x_k)] dt \right| && \text{Newton-Leibniz} \\
\le & \frac{|x_k - x^*|}{|f''(x_k)|} \cdot \int_0^1 |f''(x_k + t(x^* - x_k)) - f''(x_k)| dt && \left| \int f \right| \le \int |f| \\
\le & \frac{|x_k - x^*|}{|f''(x_k)|} \cdot \int_0^1 Mt |x_k - x^*| dt && M\text{-Lipschitz of } f'' \\
= & \frac{M}{2|f''(x_k)|} |x_k - x^*|^2 \\
\le & \frac{M}{2m} |x_k - x^*|^2 && m\text{-strong convexity}
\end{aligned}
$$

Matrix Norm

The set of $m \times n$ matrices $\mathbb{R}^{m \times n}$ is a mn-dimensional vector space

A matrix norm on $\mathbb{R}^{m \times n}$ is a function $\| \cdot \| : \mathbb{R}^{m \times n} \to \mathbb{R}$ s.t.

1. $\|\boldsymbol{A}\| \geq 0$, $\forall \boldsymbol{A} \in \mathbb{R}^{m \times n}$
2. $\|\boldsymbol{A}\| = 0$ iff $\boldsymbol{A} = \boldsymbol{O}$
3. $\|c\boldsymbol{A}\| = |c| \cdot \|\boldsymbol{A}\|$, $\forall c \in \mathbb{R}, \boldsymbol{A} \in \mathbb{R}^{m \times n}$ (positive homogeneity)
4. $\|\boldsymbol{A} + \boldsymbol{B}\| \leq \|\boldsymbol{A}\| + \|\boldsymbol{B}\|$, $\forall \boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{m \times n}$ (triangle inequality)

Example. The Frobenius norm on $\mathbb{R}^{m \times n}$ is the 2-norm on $\mathbb{R}^{mn}$.

$$\|\boldsymbol{A}\|_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij}^2} \quad \text{for } \boldsymbol{A} = (a_{ij}) \in \mathbb{R}^{m \times n}$$

Operator Norm

A matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ defines a linear transformation from $\mathbb{R}^n$ to $\mathbb{R}^m$

$$\begin{aligned} \boldsymbol{A} : \mathbb{R}^n &\to \mathbb{R}^m \\ \boldsymbol{x} &\mapsto \boldsymbol{A}\boldsymbol{x} \end{aligned}$$

Given two vector norms $\| \cdot \|_a$ and $\| \cdot \|_b$ on $\mathbb{R}^n$ and $\mathbb{R}^m$, respectively, the operator norm or induced norm of $\boldsymbol{A}$ is defined by

$$\|\boldsymbol{A}\|_{a,b} = \max_{\boldsymbol{x}:\boldsymbol{x} \neq \boldsymbol{0}} \frac{\|\boldsymbol{A}\boldsymbol{x}\|_b}{\|\boldsymbol{x}\|_a} = \max_{\boldsymbol{x}:\|\boldsymbol{x}\|_a = 1} \|\boldsymbol{A}\boldsymbol{x}\|_b = \max_{\boldsymbol{x}:\|\boldsymbol{x}\|_a \leq 1} \|\boldsymbol{A}\boldsymbol{x}\|_b$$

Exercise. Show the three definitions are equivalent.

The induced norm has the following important property.

Proposition (compatibility of norms).

$$\|\boldsymbol{A}\boldsymbol{x}\|_b \leq \|\boldsymbol{A}\|_{a,b} \|\boldsymbol{x}\|_a$$

Spectral Norm

When the norms on $\mathbb{R}^n$ and $\mathbb{R}^m$ are both 2-norms, the induced norm on $\mathbb{R}^{n\times m}$ is simply called the 2-norm or spectral norm, denoted by $\|\cdot\|_2$.

Proposition.

$$\|\boldsymbol{A}\|_2 = \sqrt{\lambda_{\max}(\boldsymbol{A}^T\boldsymbol{A})},$$

where $\lambda_{\max}(\boldsymbol{A}^T\boldsymbol{A})$ is the maximum eigenvalue of $\boldsymbol{A}^T\boldsymbol{A}$.

Proof. Let $\|\boldsymbol{x}\|_2 = 1$. By slide 15 of Lecture 8,

$$\|\boldsymbol{Ax}\|_2^2 = \boldsymbol{x}^T\boldsymbol{A}^T\boldsymbol{Ax} \le \lambda_{\max}(\boldsymbol{A}^T\boldsymbol{A})\|\boldsymbol{x}\|_2^2 = \lambda_{\max}(\boldsymbol{A}^T\boldsymbol{A}), \quad \forall \boldsymbol{x} \in \mathbb{R}^n$$

with equality iff $\boldsymbol{x}$ is an eigenvector of $\boldsymbol{A}^T\boldsymbol{A}$ associated with $\lambda_{\max}(\boldsymbol{A}^T\boldsymbol{A})$.

Corollary. If $\boldsymbol{A}$ is symmetric,

$$\|\boldsymbol{A}\|_2 = \max\{|\lambda_{\max}(\boldsymbol{A})|, |\lambda_{\min}(\boldsymbol{A})|\}$$

If $\boldsymbol{A} \succeq \boldsymbol{O}$, then $\|\boldsymbol{A}\|_2 = \lambda_{\max}(\boldsymbol{A})$.

Examples

Example.

$$\boldsymbol{A} = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$$

To find the 2-norm,

$$\boldsymbol{A}^T\boldsymbol{A} = \begin{pmatrix} 1 & 3 \\ 2 & 4 \end{pmatrix} \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} = \begin{pmatrix} 10 & 14 \\ 14 & 20 \end{pmatrix}$$

$$\|\boldsymbol{A}\|_2 = \sqrt{\lambda_{\max}(\boldsymbol{A}^T\boldsymbol{A})} = \sqrt{15 + \sqrt{221}} \approx 5.465$$

Example.

$$\boldsymbol{A} = \begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix} \succeq \boldsymbol{O}$$

$$\|\boldsymbol{A}\|_2 = \sqrt{\lambda_{\max}(\boldsymbol{A}^T\boldsymbol{A})} = \sqrt{\lambda_{\max}(\boldsymbol{A}^2)} = \sqrt{\lambda_{\max}^2(\boldsymbol{A})} = \lambda_{\max}(\boldsymbol{A}) = 5$$

Convergence Analysis

$\nabla^2 f$ is M-Lipschitz continuous if

$$\|\nabla^2 f(\boldsymbol{x}) - \nabla^2 f(\boldsymbol{y})\|_2 \leq M\|\boldsymbol{x} - \boldsymbol{y}\|_2, \quad \forall \boldsymbol{x}, \boldsymbol{y}$$

Theorem. If f is m-strongly convex, $\nabla^2 f$ is M-Lipschitz continuous, and $\boldsymbol{x}^*$ is a minimum of f, then the sequence $\{\boldsymbol{x}_k\}$ produced by Newton's method satisfies

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

Note. The same remarks on slide 7 apply here with $|x_k - x^*|$ replaced by $\|\boldsymbol{x}_k - \boldsymbol{x}^*\|$. In particular, if $\|\boldsymbol{x}_0 - \boldsymbol{x}^*\| < \frac{2m}{M}$, then

$$\|\boldsymbol{x}_k - \boldsymbol{x}^*\| \leq \frac{2m}{M}\left(\frac{M}{2m}\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|\right)^{2^k}$$

The proof is also very similar with only minor modifications.

Proof

$$
\begin{aligned}
& \|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\| \\
= \; & \|\boldsymbol{x}_k - \boldsymbol{x}^* - [\nabla^2 f(\boldsymbol{x}_k)]^{-1}[\nabla f(\boldsymbol{x}_k) - \nabla f(\boldsymbol{x}^*)]\| && (1) \\
\leq \; & \|[\nabla^2 f(\boldsymbol{x}_k)]^{-1}\| \cdot \|\nabla f(\boldsymbol{x}^*) - \nabla f(\boldsymbol{x}_k) - \nabla^2 f(\boldsymbol{x}_k)(\boldsymbol{x}^* - \boldsymbol{x}_k)\| && (2) \\
= \; & \|[\nabla^2 f(\boldsymbol{x}_k)]^{-1}\| \cdot \left\| \int_0^1 [\nabla^2 f(\boldsymbol{x}_k + t(\boldsymbol{x}^* - \boldsymbol{x}_k)) - \nabla^2 f(\boldsymbol{x}_k)](\boldsymbol{x}^* - \boldsymbol{x}_k) dt \right\| && (3) \\
\leq \; & \|[\nabla^2 f(\boldsymbol{x}_k)]^{-1}\| \int_0^1 \left\| [\nabla^2 f(\boldsymbol{x}_k + t(\boldsymbol{x}^* - \boldsymbol{x}_k)) - \nabla^2 f(\boldsymbol{x}_k)](\boldsymbol{x}^* - \boldsymbol{x}_k) \right\| dt && (4) \\
\leq \; & \|[\nabla^2 f(\boldsymbol{x}_k)]^{-1}\| \int_0^1 \left\| \nabla^2 f(\boldsymbol{x}_k + t(\boldsymbol{x}^* - \boldsymbol{x}_k)) - \nabla^2 f(\boldsymbol{x}_k) \right\| \cdot \|\boldsymbol{x}^* - \boldsymbol{x}_k\| dt && (5) \\
\leq \; & \|[\nabla^2 f(\boldsymbol{x}_k)]^{-1}\| \int_0^1 Mt \|\boldsymbol{x}^* - \boldsymbol{x}_k\|^2 dt && (6) \\
= \; & \|[\nabla^2 f(\boldsymbol{x}_k)]^{-1}\| \cdot \frac{M}{2} \|\boldsymbol{x}^* - \boldsymbol{x}_k\|^2 && (7) \\
\leq \; & \frac{M}{2m} \|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2 && (8)
\end{aligned}
$$

Proof (cont'd)

1. Step (1) uses the Newton updating rule

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - [\nabla^2 f(\boldsymbol{x}_k)]^{-1} \nabla f(\boldsymbol{x}_k)$$

   and the optimality condition $\nabla f(\boldsymbol{x}^*) = \boldsymbol{0}$.
2. Step (2) applies the compatibility of norms on slide 10 to

$$[\nabla^2 f(\boldsymbol{x}_k)]^{-1} [\nabla f(\boldsymbol{x}^*) - \nabla f(\boldsymbol{x}_k) - \nabla^2 f(\boldsymbol{x}_k)(\boldsymbol{x}^* - \boldsymbol{x}_k)]$$

3. Step (3) applies the Newton-Leibniz formula to the function $\boldsymbol{h}(t) = \nabla f(\boldsymbol{x}_k + t(\boldsymbol{x}^* - \boldsymbol{x}_k))$,

$$\nabla f(\boldsymbol{x}^*) - \nabla f(\boldsymbol{x}_k) = \boldsymbol{h}(1) - \boldsymbol{h}(0) = \int_0^1 \boldsymbol{h}'(t) dt$$

   where $\boldsymbol{h}'(t)$ is given by the chain rule,

$$\boldsymbol{h}'(t) = \nabla^2 f(\boldsymbol{x}_k + t(\boldsymbol{x}^* - \boldsymbol{x}_k))(\boldsymbol{x}^* - \boldsymbol{x}_k)$$

Proof (cont'd)

4. Step (4) uses the following inequality

$$\left\| \int \boldsymbol{f}(t)dt \right\| \leq \int \|\boldsymbol{f}(t)\|dt$$

Proof. Let $\boldsymbol{z} = \int \boldsymbol{f}(t)dt$.

$$\|\boldsymbol{z}\|^2 = \boldsymbol{z}^T \int \boldsymbol{f}(t)dt \overset{(a)}{=} \int \boldsymbol{z}^T \boldsymbol{f}(t)dt \overset{(b)}{\leq} \int \|\boldsymbol{z}\| \cdot \|\boldsymbol{f}(t)\|dt = \|\boldsymbol{z}\| \int \|\boldsymbol{f}(t)\|dt,$$

where (a) uses linearity of integration and (b) Cauchy-Schwarz.

5. Step (5) again applies the compatibility of norms on slide 10
6. Step (6) uses the Lipschitz continuity of $\nabla^2 f$
7. Step (7) performs the integration over t
8. Step (8) uses the m-strong convexity of f

$$\|[\nabla^2 f(\boldsymbol{x}_k)]^{-1}\| = \lambda_{\max}([\nabla^2 f(\boldsymbol{x}_k)]^{-1}) = \frac{1}{\lambda_{\min}(\nabla^2 f(\boldsymbol{x}_k))} \leq \frac{1}{m}$$

Contents

Damped Newton's Method

The Newton direction $-[\nabla^2 f(\boldsymbol{x})]^{-1}\nabla f(\boldsymbol{x})$ is a descent direction, but with step size 1, Newton's method does not guarantee $f(\boldsymbol{x}_{k+1}) < f(\boldsymbol{x}_k)$.

To ensure $f(\boldsymbol{x}_{k+1}) < f(\boldsymbol{x}_k)$, damped Newton's method does backtracking line search along the Newton direction.

Damped Newton's method

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

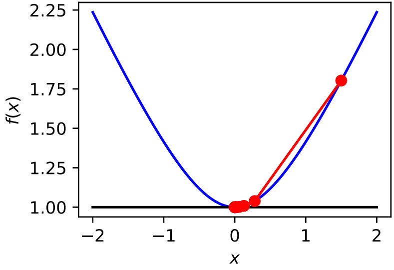
where $\alpha, \beta \in (0, 1)$

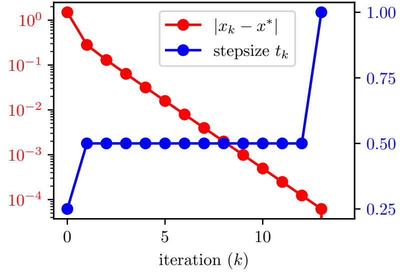
Example

$$f(x) = \sqrt{1+x^2}$$

Recall pure Newton's method converges iff $|x_0| < 1$.

Damped Newton's method converges globally, e.g. for $x_0 = 1.5$.





Convergence Analysis

Theorem. Assume f is m-strongly convex and L-smooth, $\nabla^2 f$ is M-Lipschitz, and $\boldsymbol{x}^*$ is a minimum of f. Damped Newton's method satisfies the following error bounds

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*) \leq \begin{cases} f(\boldsymbol{x}_0) - f(\boldsymbol{x}^*) - \gamma k, & \text{if } k \leq k_0 \\ \frac{2m^3}{M^2} \left(\frac{1}{2}\right)^{2^{k-k_0+1}}, & \text{if } k > k_0 \end{cases}$$

where $\gamma = 2\alpha\bar{\alpha}\beta\eta^2 m/L^2$, $\eta = \min\{1, 3(1-2\alpha)\}m^2/M$, and k_0 is the number of steps until $\|\nabla f(\boldsymbol{x}_{k_0+1})\| \leq \eta$.

Notes.

- Damped Newton's method guarantees global convergence.
- To get $f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*) \leq \epsilon$, we need at most

$$\frac{f(\boldsymbol{x}_0) - f(\boldsymbol{x}^*)}{\gamma} + \log_2 \log_2 \frac{\epsilon_0}{\epsilon}$$

where $\epsilon_0 = \frac{2m^3}{M^2}$. It can be slow if γ is small.

Convergence Analysis (cont'd)

Detailed analysis shows that the convergence follows two stages

- **Damped Newton phase.** When $\|\nabla f(\boldsymbol{x}_k)\| > \eta$, backtracking selects a step size $t_k \le 1$, and

$$f(\boldsymbol{x}_{k+1}) - f(\boldsymbol{x}_k) \le -\gamma$$

  Summing over k from 0 to $k_0 - 1$,

$$f(\boldsymbol{x}^*) - f(\boldsymbol{x}_0) \le f(\boldsymbol{x}_{k_0}) - f(\boldsymbol{x}_0) \le -k_0\gamma \implies k_0 \le \frac{f(\boldsymbol{x}_0) - f(\boldsymbol{x}^*)}{\gamma}$$

- **Pure Newton phase.** When $\|\nabla f(\boldsymbol{x}_k)\| \le \eta$, backtracking always selects step size $t_k = 1$, and

$$\|\nabla f(\boldsymbol{x}_{k+1})\| \le \frac{M}{2m^2}\|\nabla f(\boldsymbol{x}_k)\|^2 \le \frac{1}{2}\|\nabla f(\boldsymbol{x}_k)\|$$

  Once we are in the pure Newton phase, we will remain so.

Contents

Equality Constrained Optimization 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,

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

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

The feasible set is

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

We assume $X \neq \emptyset$. We also assume the constraints are independent, i.e. $\operatorname{rank} \boldsymbol{A} = k$ (What if $\operatorname{rank} \boldsymbol{A} < k$?)

Optimality Condition

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

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

where $\mathrm{Null}(\boldsymbol{A}) = \{\boldsymbol{x} : \boldsymbol{A}\boldsymbol{x} = \boldsymbol{0}\}$ is the null space of $\boldsymbol{A}$.

Proof. Recall (slide 20 of Lecture 6) $\boldsymbol{x}^* \in X$ is optimal iff

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

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

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

Note $\boldsymbol{y} \in \mathrm{Null}(\boldsymbol{A}) \iff -\boldsymbol{y} \in \mathrm{Null}(\boldsymbol{A})$. The condition then reduces to

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

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

Optimality Condition (cont'd)

Second Proof. Let $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_{n-k}$ be a basis of $\mathrm{Null}(\boldsymbol{A})$. Then $\boldsymbol{x} \in X$ iff

$$\boldsymbol{x} = \boldsymbol{x}^* + \sum_{i=1}^{n-k} z_i \boldsymbol{y}_i = \boldsymbol{x}^* + \boldsymbol{F}\boldsymbol{z}$$

where $\boldsymbol{F} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_{n-k})$. Let $g(\boldsymbol{z}) = f(\boldsymbol{x}^* + \boldsymbol{F}\boldsymbol{z})$. Note $\boldsymbol{x}^*$ is optimal for the constrained problem (EC) iff $\boldsymbol{0}$ is an unconstrained minimum of g. By the chain rule, the optimality condition is

$$\nabla g(\boldsymbol{0}) = \boldsymbol{F}^T \nabla f(\boldsymbol{x}^*) = \boldsymbol{0}$$

or

$$\frac{\partial g(\boldsymbol{0})}{\partial z_i} = \boldsymbol{y}_i^T \nabla f(\boldsymbol{x}^*) = 0, \quad i = 1, \ldots, n-k$$

Since $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_{n-k}$ is a basis of $\mathrm{Null}(\boldsymbol{A})$,

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

Optimality Condition (cont'd)

Theorem. Assume f is differentiable. $\boldsymbol{x}^* \in X$ 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

$$\mathrm{Null}(\boldsymbol{A})^\perp = \mathrm{Range}(\boldsymbol{A}^T) \triangleq \{\boldsymbol{A}^T \boldsymbol{v} : \boldsymbol{v} \in \mathbb{R}^k\},$$

$\boldsymbol{x}^*$ 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 $\boldsymbol{\lambda}^* = -\boldsymbol{v}^*$.