# Optimization

Wenxiao Yang[*]

[*]Department of Mathematics, University of Illinois at Urbana-Champaign

2022

# 目录

# 1 Unconstrained Optimization

## 1.1 Conditions for Optimality

Function: $f : \mathbb{R}^n \to \mathbb{R}^n$, $x \in \&$, $\& \subseteq \mathbb{R}^n$.

Terminology: $x^*$ will always be the optimal input at some function.

## 1.2 Global minimizer, Local minimizer

**Definition 1.**

*Say $x^*$ is a <u>global minimizer(minimum)</u> of $f$ if $f(x^*) \leq f(x), \forall x \in \&$.*

*Say $x^*$ is a <u>unique global minimizer(minimum)</u> of $f$ if $f(x^*) < f(x), \forall x \neq x^*$.*

*Say $x^*$ is a <u>local minimizer(minimum)</u> of $f$ if $\exists r > 0$ so that $f(x^*) \leq f(x)$ when $\|x - x^*\| < r$.*

A minimizer is <u>strict</u> if $f(x^*) < f(x)$ for all relevant $x$.

## 1.3 Optimization in $\mathbb{R}$

### 1.3.1 Theorem: local minimizer $\Rightarrow f'(x^*) = 0$

**Theorem 1.** *If $f(x)$ is differentiable function and $x^*$ is a local minimizer, then $f'(x^*) = 0$.*

证明.

Def of $f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$

Def of local minimizer: $f(x^*) - f(x) \geq 0, |x^* - x| < r$

when $0 < h < r$, $\frac{f(x+h) - f(x)}{h} \geq 0$; when $-r < h < 0$, $\frac{f(x+h) - f(x)}{h} \leq 0$. Then $f'(x) = 0$. □

### 1.3.2 Theorem: $f'(x^*) = 0, f''(x^*) \geq 0 \Rightarrow$ local minimizer

**Theorem 2.** *If $f : \mathbb{R} \to \mathbb{R}$ is a function with a continuous second derivative and $x^*$ is a critical point of $f$ (i.e. $f'(x) = 0$), then:*

*(1): If $f''(x) \geq 0$, $\forall x \in \mathbb{R}$, then $x^*$ is a global minimizer on $\mathbb{R}$.*

*(2): If $f''(x) \geq 0$, $\forall x \in [a, b]$, then $x^*$ is a global minimizer on $[a, b]$.*

*(3): If we only know $f''(x^*) \geq 0$, $x^*$ is a local minimizer.*

证明.

$(1) f(x) = f(x^*) + f'(x^*)(x - x^*) + \frac{1}{2} f''(\xi)(x - x^*)^2 = f(x^*) + 0 + something\ non\ negative \geq f(x^*)\ \forall x$

(2) Similar to (1)

$(3) f''(x^*) \geq 0$, $f''$ continuous $\Rightarrow \exists r$ s.t. $f''(x) \geq 0\ \forall x \in [x^* - \frac{r}{2}, x^* + \frac{r}{2}]$, then $x$ is a local minimizer. □

## 1.4 Optimization in $\mathbb{R}^n$

### 1.4.1 Necessary Conditions for Optimality: Local Extremum $\Rightarrow \nabla f(x^*) = 0$

A base point $x$, we consider an arbitrary direction $u$. $\{x + tu | t \in \mathbb{R}\}$

For $\alpha > 0$ sufficiently small:

1. $f(x^*) \leq f(x^* + \alpha u)$

2. $g(\alpha) = f(x^* + tu) - f(x^*) \geq 0$

3. $g(\beta)$ is continuously differentiable for $\beta \in [0, \alpha]$

By chain rule,

$$g'(\beta) = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(x^* + \beta u)u_i$$

By Mean Value Theorem,

$$g(\alpha) = g(0) + g'(\beta)\alpha \text{ for some } \beta \in [0, \alpha]$$

Thus

$$g(\alpha) = \alpha \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(x^* + \beta u)u_i \geq 0$$

$$\Rightarrow \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(x^* + \beta u)u_i \geq 0$$

Letting $\alpha \to 0$ and hence $\beta \to 0$, we get

$$\sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(x^*)u_i \geq 0 \text{ for all } u \in \mathbb{R}^n$$

By choosing $u = [1, 0, ..., 0]^T$, $u = [-1, 0, ..., 0]^T$, we get

$$\frac{\partial f(x^*)}{\partial x_1} \geq 0, \ \frac{\partial f(x^*)}{\partial x_1} \leq 0 \Rightarrow \frac{\partial f(x^*)}{\partial x_1} = 0$$

Similarly, we can get

$$\nabla f(x^*) = [\frac{\partial f(x^*)}{\partial x_1}, \frac{\partial f(x^*)}{\partial x_2}, ..., \frac{\partial f(x^*)}{\partial x_n}]^T = 0$$

**Theorem 3.** *If $f$ is continuously differentiable and $x^*$ is a local extremum. Then $\nabla f(x^*) = 0$.*

### 1.4.2 Stationary Point, Saddle Point

All points $x^*$ s.t. $\nabla f(x^*) = 0$ are called stationary points.

Thus, all extrema are stationary points.

But not all stationary points have to be extrema.

Saddle points are the stationary points neither local minimum nor local maximum.

**Example 1.** *$f(x) = x^3$, $x = 0$ is a stationary point but not extrema. (saddle point)*

### 1.4.3 Second Order Necessary Condition

**Definition 2.** *The Hessian of $f$ at point $x$ is an $n \times n$ symmetric matrix denoted by $\nabla^2 f(x)$ with $[\nabla^2 f(x)]_{ij} = \frac{\partial^2 f(x)}{\partial x_i \partial x_j}$*

**Theorem 4.** *Suppose $f$ is twice continuously differentiable and $x^*$ in local <u>minimum</u>. Then*

$$\nabla f(x^*) = 0 \ and \ \nabla^2 f(x^*) \succeq 0$$

证明.

$\nabla f(x^*) = 0$ already proved before.

Let $\alpha$ be small enough so that $g(\alpha) = f(x^* + \alpha u) - f(x^*) \geq 0$.

By Taylor series expansion,

$$g(\alpha) = g(0) + \alpha g'(0) + \frac{\alpha^2}{2} g''(0) + O(\alpha^2)$$

$$g'(\alpha) = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(x^* + \beta u)u_i = \nabla f(x^* + \alpha u)^T u$$

$$g''(\alpha) = \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{\partial^2 f}{\partial x_i \partial x_j}(x^* + \beta u)u_i u_j = u^T \nabla^2 f(x^* + \alpha u)u$$

$$g'(0) = \nabla f(x^*)^T u = 0; \ g''(0) = u^T \nabla^2 f(x^*)u$$

$$g(\alpha) = \frac{\alpha^2}{2} u^T \nabla^2 f(x^*)u + O(\alpha^2) \geq 0$$

$$\text{When } \alpha \to 0, \text{ we get } u^T \nabla^2 f(x^*)u \geq 0, \ \forall u \in \mathbb{R}^n$$

$$\Rightarrow \nabla^2 f(x^*) \succeq 0$$

$\square$

### 1.4.4 Sufficient Conditions for Optimality

**Theorem 5.** *Suppose $f$ is twice continuously differentiable in a neighborhood of $x^*$ and (1) $\nabla f(x^*) = 0$; (2) $\nabla^2 f(x^*) \succ 0$ ($u^T \nabla^2 f(x^*)u > 0, \ \forall u \in \mathbb{R}^n$). Then $x^*$ is local minimum.*

证明.

Consider $u \in \mathbb{R}^n$, $\alpha > 0$ and let

$$g(\alpha) = f(x^* + \alpha u) - f(x^*)$$

$$= \frac{\alpha^2}{2} u^T \nabla^2 f(x^*)u + O(\alpha^2) \geq 0$$

$$= \frac{\alpha^2}{2}[u^T \nabla^2 f(x^*)u + 2\frac{O(\alpha^2)}{\alpha^2}]$$

$$u^T \nabla^2 f(x^*)u > 0; \ \frac{O(\alpha^2)}{\alpha^2} \to 0$$

$$\Rightarrow g(\alpha) > 0 \text{ for } \alpha \text{ sufficiently small for all } u \neq 0$$

$$\Rightarrow x^* \text{ is local minimum.}$$

(specially if $\|u\| = 1$, $u^T \nabla^2 f(x^*) u \geq \lambda_{\min}(\nabla^2 f(x^*))$, $\lambda_{\min}(\nabla^2 f(x^*))$ is the minimal eigenvalues of $\nabla^2 f(x^*)$.) $\qquad\square$

### 1.4.5  Using Optimality Conditions to Find Minimum

1. Find all points satisfying necessary condition $\nabla f(x) = 0$ (all stationary points)

2. Filter out points that don't satisfy $\nabla^2 f(x) \geq 0$

3. Points with $\nabla^2 f(x) > 0$ are strict local minimum.

4. Among all points with $\nabla^2 f(x) \geq 0$, declare a global minimum, one with the smallest value of $f$, assuming that global minimum exists.

**Example 2.** $f(x) = 2x^2 - x^4$

$$f'(x) = 4x - 4x^3 = 0$$

$$\Rightarrow x = 0, x = 1, x = -1 \text{ are stationary points}$$

$$f''(x) = 4 - 12x^2 = \begin{cases} 4 & \text{if } x = 0 \\ -8 & \text{if } x = 1, -1 \end{cases}$$

$$\Rightarrow x = 0 \text{ is the only local min, and it is strict}$$

But $-f(x) \to \infty$ as $|x| \to \infty \Rightarrow$ no global min, but global max exists. $f(1), f(-1)$ are strict local max and both global max.

### 1.4.6  Fix Conditions for Global Optimality

**Claim 1**: Consider a differentiable function $f$. Suppose:
(C1) $f$ has at least one global minimizer;
(C2) The set of stationary points is $S$, and $f(x^*) \leq f(x), \forall x \in S$.
Then $x^*$ is a global minimizer of $f^*$.

证明.
Suppose $\hat{x}$ is a global minimizer of $f$, i.e.,

$$f(\hat{x}) \leq f(x), \forall x.$$

By the necessary optimality condition, we have $\nabla f(\hat{x}) = 0$, thus $\hat{x} \in S$. By (C2), we have

$$f(x^*) \leq f(\hat{x}).$$

Combining the two inequalities, we have $f(\hat{x}) \leq f(x^*) \leq f(\hat{x})$, thus $f(\hat{x}) = f(x^*)$. Plugging into the second inequality, we have $f(x^*) \leq f(x), \forall x$. Thus $x^*$ is a global minimizer of $f^*$. $\qquad\square$

## 1.5 Optimization in a Set

$$\text{minimize} \quad f(x)$$
$$\text{subject to} \quad x \in X$$

- Objective function $f : \mathbb{R}^n \to \mathbb{R}$ is a continuous function
- Optimization variable $x \in X$
- Local minimum of $f$ on $X$ : $\exists \epsilon > 0$ s.t. $f(x) \geq f(\hat{x})$, for all $x \in X$ such that $\|x - \hat{x}\| \leq \epsilon$;
i.e., $x^*$ is the best in the intersection of a small neighborhood and $X$
- Global minimum of $f$ on $X$ : $f(x) \geq f(x^*)$ for all $x \in X$
"Strict global minimum", "strict local minimum" "local maximum", "global maximum" of $f$ on $X$
are defined accordingly

### 1.5.1 Existence of Global-min

**Theorem 6** (Bolzano-Weierstrass Theorem (compact domain))**.** *Any continuous function $f$ has at least one global minimizer on any **compact set** $X$.*
*That is, there exists an $x^* \in X$ such that $f(x) \geq f(x^*), \forall x \in X$.*

**Corollary 1** (bounded level sets)**.** *Suppose $f : \mathbb{R}^d \to \mathbb{R}$ is a continuous function. If for a certain $c$, the level set*

$$\{x \mid f(x) \leq c\}$$

*is **non-empty** and **compact**, then the global minimizer of $f$ exists, i.e., there exists $x^* \in \mathbb{R}^d$ s.t.*

$$f(x^*) = \inf_{x \in \mathbb{R}^d} f(x)$$

**Example 3.** $f(x) = x^2$. *Level set $\{x|x^2 \leq 1\}$ is $\{x|-1 \leq x \leq 1\}$: non-empty compact. Thus there exists a global minimum.*

**Corollary 2** (coercive)**.** *Suppose $f : \mathbb{R}^d \to \mathbb{R}$ is a continuous function. If $f(x) \to \infty$ as $\|x\| \to \infty$, then the global minimizer of $f$ over $\mathbb{R}^d$ exists.*

证明. Let $\alpha \in \mathbb{R}^d$ be chosen so that the set $S = \{x|f(x) \leq \alpha\}$ is non-empty. By coercivity, this set is compact. $\qquad \square$

Coercive $\Rightarrow$ one non-empty bounded level set; but not the other way.
Claim (all level sets bounded $\Leftrightarrow$ coercive): Let $f$ be a continuous function, then $f$ is coercive iff $\{x|f(x) \leq \alpha\}$ is compact for any $\alpha$.

## 1.6 Method of finding-global-min-among-stationary-points (FGMSP)

Method of finding-global-min-among-stationary-points (FGMSP):

Step 0: Verify coercive or bounded level set:

- Case 1: success, go to Step 1.

- Case 2: otherwise, try to show non-existence of global-min. If success, exit and report "no global-min exists".

- Case 3: cannot verify coercive or bounded level set; cannot show non-existence of global-min. Exit and report "cannot decide".

Step 1: Find all stationary points (candidates) by solving $\nabla f(\mathbf{x}) = 0$;

Step 2 (optional): Find all candidates s.t. $\nabla^2 f(\mathbf{x}) \succeq 0$.

Step 3: Among all candidates, find one candidate with the minimal value. Output this candidate, and report "find a global min ".

# 2 Convexity

## 2.1 Definition

**Convex set** $C: x, y \in C$ implies $\lambda x + (1 - \lambda)y \in C$, for any $\lambda \in [0, 1]$.

**Convex function** (0-th order): $f$ is convex in a convex set $C$ iff $f(\alpha x + (1 - \alpha)y) \leq \alpha f(x) + (1 - \alpha)f(y), \forall x, y \in C, \forall \alpha \in [0, 1]$.

Property (1st order) If $f$ is differentiable, then $f$ is convex iff $f(z) \geq f(x) + (z - x)^T \nabla f(x), \ \forall x, z \in C$. The inequality is strict for strict convexity.

证明.

(i) "⇒"
$$f(x + \alpha(y - x)) \leq (1 - \alpha)f(x) + \alpha f(y), \forall \alpha \in (0, 1)$$
$$\Rightarrow \frac{f(x + \alpha(y - x)) - f(x)}{\alpha} \leq f(y) - f(x)$$
$$\text{Limit as } \alpha \to 0 \Rightarrow (y - x)^T \nabla f(x) \leq f(y) - f(x)$$

(ii) "⇐" Let $g = \alpha x + (1 - \alpha)y$
$$f(g) + (x - g)^T \nabla f(g) \leq f(x)$$
$$f(g) + (y - g)^T \nabla f(g) \leq f(y)$$
$$\Rightarrow f(g) \leq \alpha f(x) + (1 - \alpha)f(y)$$
$$f(\alpha x + (1 - \alpha)y) \leq \alpha f(x) + (1 - \alpha)f(y)$$

□

Property (2nd order): If $f$ is twice differentiable, then $f$ is convex iff

$$\nabla^2 f(x) \succeq 0, \ \forall x \in C.$$

Strictly convex: $\nabla^2 f(x) \succ 0, \ \forall x \in C \Rightarrow f$ is strictly convex.

**Note:** $f$ is strictly convex $\nRightarrow \nabla^2 f(x) \succ 0$.

**Example 4.** $f(x) = x^4$ *(strictly convex)*, $\frac{d^2 f(x)}{dx^2} = 12x^2 (= 0 \ at \ x = 0)$

A function $f$ is a **concave function** iff $-f$ is a convex function.

**Convex set graph**:



(a) Convex set       (b) Non-convex set
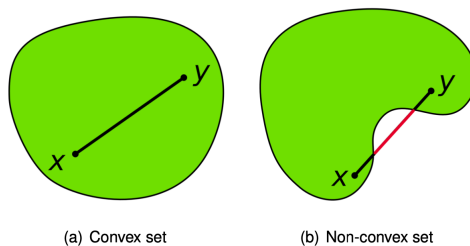
图 1:

**Claim 1.** *Suppose $f$ is a convex function over $\mathbb{R}^n$ and define the set*

$$C = \{x \in \mathbb{R}^n | f(x) \leq a\}, a \in \mathbb{R}$$

*then $C$ is a convex set.*

**Claim 2.** *If $f_1, f_2, ..., f_k$ are convex functions over convex set &,*

1. *$f_{sum}(x) = \sum_{i=1}^{k} f_i(x)$ is convex over &*

2. *$f_{max}(x) = \max_{i=1,...,k} f_i(x)$ is convex over &*

证明.

(2)

$$\begin{aligned}
f_{max}(\alpha x + (1-\alpha)y) &= \max_{i=1,...,k} f_i(\alpha x + (1-\alpha)y) \\
&\leq \max_{i=1,...,k} [\alpha f_i(x) + (1-\alpha)f_i(y)] \\
&\leq \max_{i=1,...,k} \alpha f_i(x) + \max_{i=1,...,k} (1-\alpha)f_i(y) \\
&= \alpha f_{max}(x) + (1-\alpha) f_{max}(y)
\end{aligned}$$

$\square$

9

## 2.2 Convex⇒Stationary point is global-min

**Proposition 1.** *Let $f : X \longmapsto \mathbb{R}$ be a convex function over the convex set $X$.*
*(a) A local-min of $f$ over $X$ is also a global-min over $X$. If $f$ is strictly convex,then min is unique.*
*(b) If $X$ is open (e.g. $\mathbb{R}^n$ ), then $\nabla f(x^*) = 0$ is a necessary and sufficient condition for $x^*$ to be a global minimum.*

证明.
Proof based on a property: If $f$ is differentiable over $C$ (open), then $f$ is convex iff

$$f(z) \geq f(x) + (z - x)'\nabla f(x), \quad \forall x, z \in C.$$

□

**Corollary 3.** *Let $f : X \longmapsto \mathbb{R}$ be a concave function over the convex set $X$.*
*(a) A local-max of $f$ over $X$ is also a global-max over $X$.*
*(b) If $X$ is open (e.g. $\mathbb{R}^n$ ), then $\nabla f(x^*) = 0$ is a necessary and sufficient condition for $x^*$ to be a global maximum.*

## 2.3 $\mu-$strongly convexity

**Definition**: We say $f : C \to \mathbb{R}$ is a $\mu$-strongly convex function in a convex set $C$ if $f$ is differentiable and

$$\langle \nabla f(w) - \nabla f(v), w - v \rangle \geq \mu \|w - v\|^2, \quad \forall w, v \in C.$$

If $f$ is twice differentiable, then $f$ is $\mu$-strongly convex iff

$$\nabla^2 f(x) \succeq \mu I, \quad \forall x \in C.$$

Namely, all eigenvalues of the Hessian at any point is at least $\mu$.
if $f(w)$ is convex, then $f(w) + \frac{\mu}{2}\|w\|^2$ is $\mu$-strongly convex.
- In machine learning, easy to change a convex function to a strongly convex function: just add a regularizer

## 2.4 Unconstrained Quadratic Optimization

$$\begin{aligned} \text{minimize} \quad & f(\mathbf{w}) = \tfrac{1}{2}\mathbf{w}^T\mathbf{Q}\mathbf{w} - \mathbf{b}^T\mathbf{w} \\ \text{subject to} \quad & \mathbf{w} \in \mathbb{R}^d \end{aligned}$$

where $\mathbf{Q}$ is a symmetric $d \times d$ matrix. (what if non-symmetric?)

$$\nabla f(\mathbf{w}) = \mathbf{Q}\mathbf{w} - \mathbf{b}, \ \nabla^2 f(\mathbf{w}) = \mathbf{Q}$$

(i) $\mathbf{Q} \succeq 0 \Leftrightarrow f$ is convex.

(ii) $\mathbf{Q} \succ 0 \Leftrightarrow f$ is strictly convex.

(iii) $\mathbf{Q} \preceq 0 \Leftrightarrow f$ is concave.

(iv) $\mathbf{Q} \prec 0 \Leftrightarrow f$ is strictly concave.

- Necessary condition for (local) optimality

$$\mathbf{Qw} = \mathbf{b}, \quad \mathbf{Q} \succeq 0$$

Case 1: $\mathbf{Qw} = \mathbf{b}$ has no solution, i.e. $\mathbf{b} \notin R(\mathbf{Q})$. No stationary point, no lower bound ($f$ can achieve $-\infty$).

Case 2: $\mathbf{Q}$ is not PSD ($f$ is non-convex) No local-min, no lower bound ($f$ can achieve $-\infty$).

Case 3: $\mathbf{Q} \succeq 0$ (PSD) and $\mathbf{b} \in R(\mathbf{Q})$. Convex, has global-min, any stationary point is a global optimal solution.

**Example 5.** *Toy Problem 1:* $\min_{x,y \in \mathbb{R}} f(x,y) \triangleq x^2 + y^2 + \alpha xy$.

1. Step 1: First order condition: $2x^* + \alpha y^* = 0, 2y^* + \alpha x^* = 0$.

   - We get $4x^* = -2\alpha y^* = \alpha^2 x^*$. So $\left(4 - \alpha^2\right) x^* = 0$.

   - Case 1: $\alpha^2 = 4$. If $x^* = -\alpha y^*/2$, then $(x^*, y^*)$ is a stationary point.

   - Case 2: $\alpha^2 \neq 4$. Then $x^* = 0; y^* = -\alpha x^*/2 = 0$. So $(0,0)$ is stat-pt.

2. Step 2: Check convexity. Hessian $\nabla^2 f(x,y) = \begin{pmatrix} 2 & \alpha \\ \alpha & 2 \end{pmatrix}$.

   Eigenvalues $\lambda_1, \lambda_2$ satisfy $(\lambda_i - 2)^2 = \alpha^2, i = 1, 2$. Thus $\lambda_{1,2} = 2 \pm |\alpha|$.

   - If $|\alpha| \leq 2$, then $\lambda_i \geq 0, \forall i$. Thus $f$ is convex. Any stat-pt is global-min.

   - If $|\alpha| > 2$, at least one $\lambda_i < 0$, thus $f$ is not convex.

3. Step 3 (can be skipped now): For non-convex case ($|\alpha| > 2$), prove no lower bound.

   $f(x,y) = (x + \alpha y/2) + \left(1 - \alpha^2/4\right) y^2$. Pick $y = M, x = -\alpha M/2$, then $f(x,y) = \left(1 - \alpha^2/4\right) M^2 \to -\infty$ as $M \to \infty$.

Summary:

If $|\alpha| > 2$, no global-min, $(0,0)$ is stat-pt;

if $|\alpha| = 2$, any $(-0.5\alpha t, t), t \in \mathbb{R}$ is a stat-pt and global-min;

if $|\alpha| < 2, (0,0)$ is the unique stat-pt and global-min.

**Example 6.** *Linear Regression*

minimize $f(\mathbf{w}) = \frac{1}{2}\left\|\mathbf{X}^T\mathbf{w} - \mathbf{y}\right\|^2$ subject to $\mathbf{w} \in \mathbb{R}^d$

$n$ data points, $d$ features

- $\mathbf{X}$ may be wide (under-determined), tall (over-determined), or rank-deficient

- Note that comparing with the previous case, $\mathbf{Q} = \mathbf{XX}^T \in \mathbb{R}^{d \times d}$, $\mathbf{b} = \mathbf{Xy} \in \mathbb{R}^{d \times 1}$

- $\mathbf{Q} \succeq 0$; Case 2 never happens!

- First order condition $\mathbf{XX}^\top\mathbf{w}^* = \mathbf{Xy}$.

    - It always has a solution; Case 1 never happens!

**Claim**: Linear regression problem is always convex; it has global-min.

First order condition

$$\mathbf{XX}^\top\mathbf{w}^* = \mathbf{Xy}$$

which always has a solution.

If $XX^\top \in \mathbb{R}^{d \times d}$ is invertible (only happen when $n \geq d$ ), then there is a unique stationary point $x = \left(A^\top A\right)^{-1} A^\top b$. It is also a global minimum.

If $XX^\top \in \mathbb{R}^{d \times d}$ is not invertible, then there can be infinitely many stationary points, which are the solutions to the linear equation. All of them are global minima, giving the same function value.


# 3 Gradient Methods

**Definition 3** (Iterative Descent)**.** *Start at some point $x_0$, and successively generate $x_1, x_2, ..$ s.t.*

$$f(x_{k+1}) < f(x_k) \quad k = 0, 1, ...$$

**Definition 4** (**General Gradient Descent Algorithm**)**.** *Assume that $\nabla f(x_k) \neq 0$. Then*

$$x_{k+1} = x_k + \alpha_k d_k$$

*where $d_k$ is s.t. $d_k$ has a positive projection along $-\nabla f(x_k)$,*

$$\nabla f(x_k)^T d_k < 0 \equiv -\nabla f(x_k)^T d_k > 0$$

- If $d_k = -\nabla f(x_k)$ we get **steepest descent**.

- Often $d_k$ is constructed using matrix $D_k \succ 0$

$$d_k = -D_k \nabla f(x_k)$$


## 3.1 Steepest Descent

We want the $x_k$ that decreases the function most.

**Proposition 2.** $-\nabla f(x_k)$ *is the direction deceases the function most.*

证明. Suppose the direction is $v \in \mathbb{R}^n, v \neq 0$.

$$f(x + \alpha v) = f(x) + \alpha v^T \nabla f(x) + O(\alpha)$$

The rate of change of $f$ along direction $v$:

$$\lim_{\alpha \to 0} \frac{f(x + \alpha v) - f(x)}{\alpha} = v^T \nabla f(x)$$

By Cauchy-schwarz inequality,

$$|v^T \nabla f(x)| \leq \|v\| \|\nabla f(x)\|$$

Equation holds when $v = \beta \nabla f(x)$. Hence, $-\nabla f(x)$ is the direction decreases the function most. $\square$

**Definition 5** (**Steepest Descent Algorithm**).

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k)$$

$\alpha_k$ *is the step size, which need to choose carefully.*

## 3.2 Methods for Choosing $\alpha_k$

Method (1): Fixed step size: $\alpha_k = \alpha$ (can have issue with *convergence*)

Method (2): **Optimal Line Search**: choose $\alpha_k$ to optimize the value of next iteration, i.e. solve

$$\min_{\alpha \geq 0} f(x_k + \alpha d_k)$$

(may be *difficult in practice*)

Method (3): **Armijo's Rule** (successive step size reduction):

$$f(x_k + \alpha_k d_k) = f(x_k) + \alpha_k \nabla f(x_k)^T d_k + O(\alpha_k)$$

Since $\nabla f(x_k)^T d_k < 0$, $f$ decreases when $\alpha_k$ is sufficiently small. But we also don't want $\alpha_k$ to be too small (slow).

## 3.3 Armijo's Rule

(i) Initialize $\alpha_k = \tilde{\alpha}$. Let $\sigma, \beta \in (0, 1)$ be prespecified paramenters.

(ii) If $f(x_k) - f(x_k + \alpha_k d_k) \geq -\sigma \alpha_k \nabla f(x_k)^T d_k$, stop.

   (Which shows $f(x_k + \alpha_k d_k)$ is at least smaller than $f(x_k)$ in a degree that correlated with $\nabla f(x_k)^T d_k$)

(iii) Else, set $\alpha_k = \beta \alpha_k$ and go back to step 2. (use a smaller $\alpha_k$)

   Termination at smallest integer $m$ s.t.

$$f(x_k) - f(x_k + \beta^m \tilde{\alpha} d_k) \geq -\sigma \beta^m \tilde{\alpha} \nabla f(x)^T d_k$$

In Bersekas's book: $\sigma \in [10^{-5}, 10^{-1}], \beta \in [\frac{1}{10}, \frac{1}{2}]$.

As $\sigma, \beta$ are smaller, the algorithm is quicker.

## 3.4 Armijo's Rule for Steepest Descent

$\alpha_k = \tilde{\alpha}\beta^{m_k}$, where $m_k$ is smallest $m$ s.t.

$$f(x_k) - f(x_k - \tilde{\alpha}\beta^m \nabla f(x_k)) \geq \sigma\tilde{\alpha}\beta^m \|\nabla f(x)\|^2$$

**Proposition 3.** *Assume $\inf_x f(x) > -\infty$. Then every limit point of $\{x_k\}$ for steepest descent with Armijo's rule is a <u>stationary point</u> of $f$.*

证明. Assume that $\bar{x}$ is a limit point of $\{x_k\}$ s.t. $\nabla f(\bar{x}) \neq 0$.

- Since $\{f(x_k)\}$ is monotonically non-increasing and bounded below, $\{f(x_k)\}$ converges.

- $f$ is continuous $\Rightarrow f(\bar{x})$ is a limit point of $\{f(x_k)\} \Rightarrow \lim_{k\to\infty} f(x_k) = f(\bar{x}) \Rightarrow f(x_k) - f(x_{k+1}) \to 0$

- By definition of Armijo's rule:

$$f(x_k) - f(x_{k+1}) \geq \sigma\alpha_k \|\nabla f(x_k)\|^2$$

Hence, $\sigma\alpha_k \|\nabla f(x_k)\|^2 \to 0$.

Since $\nabla f(\bar{x}) \neq 0$, $\lim_{k\to\infty} \alpha_k = 0$

$$ln\alpha_k = ln(\tilde{\alpha}\beta^{m_k}) = ln\tilde{\alpha} + m_k ln\beta \Rightarrow m_k = \frac{ln\alpha_k - ln\tilde{\alpha}}{ln\beta} \Rightarrow \lim_{k\to\infty} m_k = \infty$$

Exist $\bar{k}$ s.t. $m_k > 1, \forall k > \bar{k}$

$$f(x_k) - f(x_k - \frac{\alpha_k}{\beta}\nabla f(x_k)) < \sigma\frac{\alpha_k}{\beta}\|\nabla f(x_k)\|^2, \forall k > \bar{k}$$

By Taylor's Theorem,

$$f(x_k - \frac{\alpha_k}{\beta}\nabla f(x_k)) = f(x_k) - \nabla f(x_k - \frac{\bar{\alpha}_k}{\beta}\nabla f(x_k))^T \frac{\alpha_k}{\beta}\nabla f(x_k)$$

for some $\bar{\alpha}_k \in (0, \alpha_k)$

Hence,

$$\nabla f(x_k - \frac{\bar{\alpha}_k}{\beta}\nabla f(x_k))^T \frac{\alpha_k}{\beta}\nabla f(x_k) < \sigma\frac{\alpha_k}{\beta}\|\nabla f(x_k)\|^2$$

$$\nabla f(x_k - \frac{\bar{\alpha}_k}{\beta}\nabla f(x_k))^T \nabla f(x_k) < \sigma\|\nabla f(x_k)\|^2, \forall k > \bar{k}$$

$$\text{As } \alpha_k \to 0 \Rightarrow \bar{\alpha}_k \to 0$$

$$\|\nabla f(x_k)\|^2 < \sigma\|\nabla f(x_k)\|^2$$

Which contradicts to $\sigma < 1$.

$\square$

# 4   Convergence of GD with Constant Stepsize

**Definition 6** (Lipschitz Continunity). *A function* $y : \mathbb{R}^n \to \mathbb{R}^m$ *is called Lipschitz (continuous) if* $\exists L > 0$ *s.t.*

$$\|g(y) - g(x)\| \leq L\|y - x\|, \forall x, y \in \mathbb{R}^n$$

*L is Lipschitz constant.*

**Definition 7** (Lipschitz Gradient). $\nabla f(x)$ *is Lipschitz if* $\exists L > 0$ *s.t.*

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|, \forall x, y \in \mathbb{R}^n$$

**Example 7.**

1. $f(x) = \|x\|^4$, $\nabla f(x) = 4\|x\|^2 x$

   *Test* $\|\nabla f(x) - \nabla f(-x)\| \leq L\|2x\|$, $8\|x\|^2\|x\| \leq 2L\|x\|$ *which doesn't hold when* $\|x\|^2 > \frac{L}{4}$.

2. *If f is twice continuously differentiable with* $\nabla^2 f(x) \succeq -MI$ *and* $\nabla^2 f(x) \preceq MI$ *then* $\|\nabla f(x) - \nabla f(y)\| \leq M\|x - y\|, \forall x, y \in \mathbb{R}^n$. *(* $A \succeq B$ *means* $A - B \succeq 0$, $A \preceq B$ *means* $A - B \preceq 0$*)*

**Theorem 7.** $-MI \preceq \nabla^2 f(x) \preceq MI, \forall x \Rightarrow \|\nabla f(x) - \nabla f(y)\| \leq M\|y - x\|$

证明. For symmetric $A$,

1. $x^T A x \leq \lambda_{\max}(A)\|x\|^2$

2. $\lambda_i(A^2) = \lambda_i^2(A)$

3. $-MI \preceq A \preceq MI \Rightarrow \lambda_{\min}(A) \geq -M, \lambda_{\max}(A) \leq M$

Define $g(t) = \frac{\partial f}{\partial x_i}(x + t(y - x))$. Then $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad$ $\square$

# 5   MATH 484

A base point $x$, we consider an arbitrary direction $u$. $\{x + tu | t \in \mathbb{R}\}$

We define <u>the restriction of $f$ to the line through $x$ in the direction of $u$</u> to be the function:

$$\phi_u(t) = f(x + tu)$$

**Lemma 1.** $x^*$ *is a global minimizer of $f$ iff for all $u$, $t = 0$ is the global minimizer of $\phi_u(t)$*

证明.
$(\Rightarrow)$ $\phi_u(0) = f(x^*) \leq f(x^* + tu) = \phi_u(t)$
$(\Leftarrow)$ Let $X \in \mathbb{R}^n$, $u = X - x^*$. $\phi_u(0) \leq \phi_u(1) \Rightarrow f(x^*) \leq f(x^* + u) = f(x)$ $\qquad\qquad$ $\square$

### 5.0.1 The first-derivative test in $\mathbb{R}^n$: $\phi_u'(t) = \nabla f(x+tu) \cdot u$

First derivative of $f : \mathbb{R}^n \to \mathbb{R}$, Easier: $\phi_u'(t)$?

Suppose $f : \mathbb{R}^n \to \mathbb{R}$ and $\mathbf{g} : \mathbb{R} \to \mathbb{R}^n$:

$$\frac{\partial f(\mathbf{g}(t))}{\partial t} = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(\mathbf{g}(t)) \frac{d}{dt} g_i(t)$$

$$\frac{\partial \phi_u(t)}{\partial t} = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(x+tu)u_i$$

The gradient of $f$: $\nabla f(x) = (\frac{\partial f}{\partial x_1}, ..., \frac{\partial f}{\partial x_d})^T \Rightarrow \phi_u'(t) = \nabla f(x+tu) \cdot u$

<u>Fine print</u>: Chain rule only works when all $\frac{\partial f}{\partial x_k}$ exists and are continuous.

**Example 8.** $f(x,y) = x^2 + 3xy - 1$, $x^* = (0,0)$, $u = (3,2)$

$\phi_u(t) = f(x^* + tu) = f(3t, 2t) = 27t^2 - 1$

$\phi_u'(t) = 54t$

$\nabla f(x,y) = (2x+3y, 3x)$

$\phi_u'(t) = \nabla f(x+tu) \cdot u = 54t$

### 5.0.2 Theorem 4: $\nabla f$ is continuous, $x^*$ is a global minimizer of $f \Rightarrow \nabla f(x^*) = 0$

**Theorem 8** (Theorem 2.1). *Given a function $f : \mathbb{R}^n \to \mathbb{R}$, if $\nabla f$ is continuous and $x^*$ is a global minimizer of $f$, then $\nabla f(x^*) = 0$. (When $\nabla f(x^*) = 0$, we call $x^*$ a <u>critical point</u> of $f$.)*

$x^*$ is a global minimizer $\Rightarrow x^*$ is a critical point, inverse may not true.

### 5.0.3 The second-derivative test in $\mathbb{R}^n$

$$\phi_u'(t) = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(x+tu)u_i$$

$$\phi_u''(t) = \sum_{i=1}^{n} \sum_{j=1}^{n} u_i u_j \frac{\partial^2 f}{\partial x_i \partial x_j}(x+tu)$$

### 5.0.4 Hessian matrix

Define Hessian matrix of $f$ and write $Hf$. That is,

$$\phi_u''(t) = u^T Hf(x+tu)u$$

<u>Fine print</u>: Chain rule only works when all $\frac{\partial^2 f}{\partial x_i \partial x_j}$ exists and are continuous. ($\Rightarrow Hf$ is continuous)

**5.0.5** **Theorem 5:** $Hf$ **is continuous,** $\nabla f(x^*) = 0$**,** $u^T H f(x^*)u \geq 0, \forall u \Rightarrow x^*$ **is a global**
      **minimizer of** $f$

**Theorem 9.** *Given a function* $f : \mathbb{R}^n \to \mathbb{R}$*, if* $Hf$ *is continuous and* $x^*$ *is a critical point of* $f$*. If*
*for any u, that* $u^T H f(x^*)u \geq 0$*. Then* $x^*$ *is a global minimizer of* $f$*.*

proved by Taylor

**Theorem 10** (Taylor)**.** *Given a function* $f : \mathbb{R}^n \to \mathbb{R}$*, if* $Hf$ *is continuous and* $x^*$ *is a critical point*
*of* $f$*, then*

$$f(x) = f(x^*) = \nabla f(x^*)(x - x^*) + \frac{1}{2}(x - x^*)^T H f(z)(x - x^*)$$

*for some z on the line between x and* $x^*$

## 5.1 Minimizing over other sets

What if the domain of $f$: $D \subset \mathbb{R}^n$
(1): want $x^*$ to be in the interior of $D$, not on the boundary (want to be able to "look" from $x^*$ in
any direction.)
(2): want $x^*$ to "see" all other points in $D$ using straight line $u$.
Convexity
good domain e.g. Ball: $B(x^*, r) = \{x | ||x - x^*|| < r\}$

**5.1.1** **Theorem 7:** $\nabla f$ **is continuous,** $x^*$**(interior of** $D$**) is a local minimizer of** $f \Rightarrow \nabla f(x^*) =$
      $0$

**Theorem 11** (Theorem 4.1, 类似 Theorem 2.1)**.** *Suppose* $f : D \to \mathbb{R}$ *has continuous* $\nabla f$ *and* $x^*$ *is*
*not on the boundary of D. If* $x^*$ *is a local minimizer of* $f$*, then* $x^*$ *is a critical point of* $f$*:* $\nabla f(x^*) = 0$

**5.1.2** **Theorem 8:** $Hf$ **is continuous,** $x^*$**(interior of** $D$**)** $\nabla f(x^*) = 0$**,** $\exists r$ **s.t.** $u^T H f(x^*)u \geq$
      $0, \forall x \in B(x^*, r), \forall u \Rightarrow x^*$ **is a local minimizer of** $f$

**Theorem 12.** *Given a function* $f : D \to \mathbb{R}$ *, if* $Hf$ *is continuous and* $x^*$ *is a critical point of* $f$ *in*
*the interior of D. Suppose* $\exists r$ *s.t. for any u, that* $u^T H f(x^*)u \geq 0$ *whenever* $x \in B(x^*, r) \subset D$*. Then*
$x^*$ *is a local minimizer of* $f$*.*