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 global minimizer(minimum) of f if $f(x^*) \leq f(x), \forall x \in \&$.
Say x^* is a unique global minimizer(minimum) of f if $f(x^*) < f(x), \forall x \neq x^*$.
Say x^* is a local minimizer(minimum) of f if $\exists r > 0$ so that $f(x^*) \leq f(x)$ when $\|x - x^*\| < r$.

A minimizer is strict 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)$f(x) = f(x^*) + f'(x^*)(x - x^*) + \frac{1}{2}f''(\xi)(x - x^*)^2 = f(x^*) + 0 + \textit{something non negative} \geq f(x^*)\ \forall x$
(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 minimum. Then*

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

证明.
$\nabla f(x^*) = 0$ already proved before.
Let α be small enough so that $g(\alpha) = f(x^* + \alpha u) - f(x^*) \geq 0$.
By Taylor series expansion,

$$\begin{aligned}
g(\alpha) &= g(0) + \alpha g'(0) + \frac{\alpha^2}{2} g''(0) + O(\alpha^2) \\
g'(\alpha) &= \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(x^* + \beta u) u_i = \nabla f(x^* + \alpha u)^T u \\
g''(\alpha) &= \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{\partial^2 f}{\partial x_i \partial x_j}(x^* + \beta u) u_i u_j = u^T \nabla^2 f(x^* + \alpha u) u
\end{aligned}$$

$$\begin{aligned}
g'(0) = \nabla f(x^*)^T u = 0;\ g''(0) &= u^T \nabla^2 f(x^*) u \\
g(\alpha) &= \frac{\alpha^2}{2} u^T \nabla^2 f(x^*) u + O(\alpha^2) \geq 0 \\
\text{When } \alpha \to 0, \text{ we get } u^T \nabla^2 f(x^*) u &\geq 0,\ \forall u \in \mathbb{R}^n \\
\Rightarrow \nabla^2 f(x^*) &\succeq 0
\end{aligned}$$

□

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

$$\begin{aligned}
g(\alpha) &= f(x^* + \alpha u) - f(x^*) \\
&= \frac{\alpha^2}{2} u^T \nabla^2 f(x^*) u + O(\alpha^2) \geq 0 \\
&= \frac{\alpha^2}{2} [u^T \nabla^2 f(x^*) u + 2 \frac{O(\alpha^2)}{\alpha^2}] \\
u^T \nabla^2 f(x^*) u > 0;\ & \frac{O(\alpha^2)}{\alpha^2} \to 0 \\
&\Rightarrow g(\alpha) > 0 \text{ for } \alpha \text{ sufficiently small for all } u \neq 0 \\
&\Rightarrow x^* \text{ is local minimum.}
\end{aligned}$$

(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 $\nabla^2 f(x^*)$.) $\square$

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

$$
\begin{aligned}
f'(x) &= 4x - 4x^3 = 0 \\
&\Rightarrow x = 0, x = 1, x = -1 \text{ are stationary points} \\
f''(x) &= 4 - 12x^2 = \begin{cases} 4 & \text{if } x = 0 \\ -8 & \text{if } x = 1, -1 \end{cases} \\
&\Rightarrow x = 0 \text{ is the only local min, and it is strict}
\end{aligned}
$$

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^*. $\square$

1.5 Optimization in a Set

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \in X \end{array}$$

- 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. □

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 α.

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) "$\Rightarrow$"

$$
\begin{aligned}
f(x+\alpha(y-x)) &\leq (1-\alpha)f(x) + \alpha f(y), \forall \alpha \in (0,1) \\
\Rightarrow \frac{f(x+\alpha(y-x)) - f(x)}{\alpha} &\leq f(y) - f(x) \\
\text{Limit as } \alpha \to 0 \Rightarrow (y-x)^T \nabla f(x) &\leq f(y) - f(x)
\end{aligned}
$$

(ii) "$\Leftarrow$" Let $g = \alpha x + (1-\alpha)y$

$$
\begin{aligned}
f(g) + (x-g)^T \nabla f(g) &\leq f(x) \\
f(g) + (y-g)^T \nabla f(g) &\leq f(y) \\
\Rightarrow f(g) &\leq \alpha f(x) + (1-\alpha) f(y) \\
f(\alpha x + (1-\alpha)y) &\leq \alpha f(x) + (1-\alpha) f(y)
\end{aligned}
$$

□

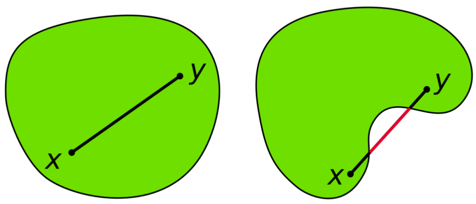
Property (2nd order): If f is twice differentiable, then f is convex iff

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

Strictly convex: $\nabla^2 f(x) \succ 0, \ \forall x \in C \Rightarrow f$ is strictly convex.
Note: f is strictly convex $\not\Rightarrow \nabla^2 f(x) \succ 0$.

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)

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

□

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 Unconstrained Quadratic Optimization

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

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

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

Case 1: $\mathbf{Q}\mathbf{w} = \mathbf{b}$ 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 λ_1,λ_2 satisfy $\left(\lambda_i-2\right)^2=\alpha^2, i=1,2$. 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, $\mathbf{Q}=\mathbf{X}\mathbf{X}^T\in\mathbb{R}^{d\times d}$, $\mathbf{b}=\mathbf{X}\mathbf{y}\in\mathbb{R}^{d\times 1}$
- $\mathbf{Q}\succeq 0$; Case 2 never happens!
- First order condition $\mathbf{X}\mathbf{X}^\top\mathbf{w}^*=\mathbf{X}\mathbf{y}$.
  - It always has a solution; Case 1 never happens!

Claim: Linear regression problem is always convex; it has global-min.
First order condition

$$\mathbf{X}\mathbf{X}^\top\mathbf{w}^*=\mathbf{X}\mathbf{y}$$

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.

2.4 Strongly Convexity

2.4.1 μ-Strongly Convex: $\langle \nabla f(w) - \nabla f(v), w - v \rangle \geq \mu \|w - v\|^2$

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

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

2.4.2 μ-strongly convex $\Leftrightarrow \nabla^2 f(x) \succeq \mu I$

If f is twice differentiable, then f is μ-strongly convex iff

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

Definition 3. *A twice continuously differentiable function is* <u>*strongly convex*</u> *if*

$$\exists m > 0 \ s.t. \ \nabla^2 f(x) \succeq mI \quad \forall x$$

which is also called $m-$strongly convex.

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

2.4.3 Lemma: Strongly convexity $\Rightarrow$ Strictly convexity

Lemma 1. *Strongly convexity $\Rightarrow$ Strictly convexity.*

证明.

$$\begin{aligned} \nabla^2 f(x) \succeq mI &\Rightarrow \nabla^2 f(x) - mI \succeq 0 \\ &\Rightarrow \forall z \neq 0 \quad z^T(\nabla^2 f(x) - mI)z \geq 0 \\ &\Rightarrow z^T \nabla^2 f(x) z \geq m z^T z > 0 \end{aligned}$$

□

Note: converse is not true: e.g. $f(x) = x^4$ is strictly convex but $\nabla^2 f(0) = 0$

2.4.4 Lemma: $\nabla^2 f(x) \succeq mI \Rightarrow f(y) \geq f(x) + \nabla f(x)^T(y - x) + \frac{m}{2}\|y - x\|^2$

Lemma 2. $\nabla^2 f(x) \succeq mI \quad \forall x$

$$\Rightarrow f(y) \geq f(x) + \nabla f(x)^T(y - x) + \frac{m}{2}\|y - x\|^2$$

证明. By Taylor's Theorem,

$$\begin{aligned} f(y) &= f(x) + \nabla f(x)^T(y-x) + \frac{1}{2}(y-x)^T \nabla^2 f((1-\beta)x+\beta y)(y-x), \quad \text{for some } \beta \in [0,1] \\ &\geq f(x) + \nabla f(x)^T(y-x) + \frac{1}{2}(y-x)^T m(y-x) \\ &\geq f(x) + \nabla f(x)^T(y-x) + \frac{m}{2}\|y-x\|^2 \end{aligned}$$

□

3 Gradient Methods

Definition 4 (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 5 (**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. □

Definition 6 (**Steepest Descent Algorithm**)**.**

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

α_k is the step size, which need to choose carefully.

3.2 Methods for Choosing α_k

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

Method (2): **Optimal Line Search**: choose α_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 α_k is sufficiently small. But we also don't want α_k to be too small (slow).

3.3 Optimal(Exact) Line Search

Example 7. *(False $\times$) The gradient descent algorithm with an exact line search always finds the minimum of a strictly convex quadratic function in exactly one iteration.*

Note: the moving direction is restricted to the gradient. x 只能朝倒数方向移动不一定能一次到大最优点。

Counterexample: False. It is not necessary that the gradient at x_0 towards the exact solution. For example, let $f(x) = \frac{1}{2}x^\top Qx + x^\top b$ where $Q = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}$ and $b = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$. Clearly we have $x^* = \begin{pmatrix} -1/2 \\ 1 \end{pmatrix}$. If we start with $x_0 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$, by using exact line search, the step size $\alpha = \arg\min f(x_0 - \alpha \nabla f(x_0)) = 10/19$. Hence $x_1 = x_0 - \alpha \nabla f(x_0) = \begin{pmatrix} -11/19 \\ 28/19 \end{pmatrix} \neq x^*$.

3.4 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 α_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 σ, β are smaller, the algorithm is quicker.

3.5 Armijo's Rule for Steepest Descent

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

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

Proposition 3. *Assume* $\inf_x f(x) > -\infty$. *Then every limit point of* $\{x_k\}$ *for steepest descent with Armijo's rule is a stationary point 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 $\sigma < 1$.

□

4 Convergence of GD with Constant Stepsize

4.1 Lipschitz Gradient (L-Smooth)

Definition 7 (Lipschitz Continunity)**.** *A function $g : \mathbb{R}^n \to \mathbb{R}^m$ is called Lipschitz (continuous) if $\exists L > 0$ s.t.*

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

L is Lipschitz constant. g is called L-smooth.

Definition 8 (Lipschitz Gradient)**.** *$\nabla f(x)$ is Lipschitz if $\exists L > 0$ s.t.*

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

Example 8.

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

   Test $\|\nabla f(x) - \nabla f(-x)\| \le L\|2x\|$, $8\|x\|^2\|x\| \le 2L\|x\|$ 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 $\|\nabla f(x) - \nabla f(y)\| \le M\|x - y\|, \forall x, y \in \mathbb{R}^n$. ($A \succeq B$ means $A - B \succeq 0$, $A \preceq B$ means $A - B \preceq 0$)*

4.1.1 Theorem: $-MI \preceq \nabla^2 f(x) \preceq MI \Rightarrow \nabla f(x)$ is Lipschitz with constant M

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

证明. For symmetric A,

1. $x^T A x \le \lambda_{\max}(A)\|x\|^2$
2. $\lambda_i(A^2) = \lambda_i^2(A)$
3. $-MI \preceq A \preceq MI \Rightarrow \lambda_{\min}(A) \ge -M, \lambda_{\max}(A) \le M$

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

$$\begin{aligned}
g(1) &= g(0) + \int_0^1 g'(s)ds \\
\Rightarrow \frac{\partial f(y)}{\partial x_i} &= \frac{\partial f(x)}{\partial x_i} + \int_0^1 \sum_{j=1}^n \frac{\partial^2 f(x+s(y-x))}{\partial x_i \partial x_j}(y_j - x_j)ds \\
\nabla f(y) &= \nabla f(x) + \int_0^1 \nabla^2 f(x+s(y-x))(y-x)ds \\
\|\nabla f(y) - \nabla f(x)\| &= \|\int_0^1 \nabla^2 f(x+s(y-x))(y-x)ds\| \\
&\le \int_0^1 \|\nabla^2 f(x+s(y-x))(y-x)\|ds \\
&= \int_0^1 \sqrt{(y-x)^T[\nabla^2 f(x+s(y-x))]^2(y-x)}ds \\
&(\text{Set } H = \nabla^2 f(x+s(y-x))) \\
&\le \int_0^1 \sqrt{\lambda_{\max}(H^2)\|y-x\|^2}ds \\
&\le M\|y-x\|
\end{aligned}$$

□

4.1.2 Descent Lemma: $\nabla f(x)$ is Lipschitz with constant $L \Rightarrow f(y) \le f(x) + \nabla f(x)^T(y-x) + \frac{L}{2}\|y-x\|^2$

Lemma 3 (Descent Lemma)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable with a Lipschitz gradient with Lipschitz constant L. Then*

$$f(y) \le f(x) + \nabla f(x)^T(y-x) + \frac{1}{2}L\|y-x\|^2$$

证明. Let $g(t) = f(x + t(y-x))$. Then $g(0) = f(x)$ and $g(1) = f(y)$, $g(1) = g(0) + \int_0^1 g'(t)dt$. Where $g'(t) = \nabla f(x+t(y-x))^T(y-x)$

$$\begin{aligned}
\Rightarrow f(y) &= f(x) + \int_0^1 \nabla f(x+t(y-x))^T(y-x)dt \\
&= f(x) + \int_0^1 (\nabla f(x+t(y-x)) - \nabla f(x))^T(y-x)dt + \nabla f(x)^T(y-x) \\
&\le f(x) + \int_0^1 \|\nabla f(x+t(y-x)) - \nabla f(x)\|\|y-x\|dt + \nabla f(x)^T(y-x) \\
&\le f(x) + L\int_0^1 \|t(y-x)\|\|y-x\|dt + \nabla f(x)^T(y-x) \\
&= f(x) + \frac{1}{2}L\|y-x\|^2 + \nabla f(x)^T(y-x)
\end{aligned}$$

□

4.1.3 Co-coercivity Condition: $(\nabla f(x)-\nabla f(y))^T(x-y) \geq \frac{1}{L}\|\nabla f(x)-\nabla f(y)\|^2$

Theorem 8 (Co-coercivity Condition)**.** *Let f be convex and continuously differentiable. Let f be L-smooth. Then*

$$(\nabla f(x)-\nabla f(y))^T(x-y) \geq \frac{1}{L}\|\nabla f(x)-\nabla f(y)\|^2$$

证明. Let $y \in \mathbb{R}^n$, and define $g(x) = f(x) - \nabla f(y)^T x$. Then $\nabla g(y) = \nabla f(y) - \nabla f(y) = 0$ and $\nabla^2 g(y) = \nabla^2 f(y) \succeq 0$, i.e. y minimize g. Because $g(y) \leq g(\cdot)$, $g(y) \leq g(x - \frac{1}{L}\nabla g(x))$ According to the descent lemma,

$$\begin{aligned}
g(x-\frac{1}{L}\nabla g(x)) &= f(x-\frac{1}{L}\nabla g(x)) - \nabla f(y)^T(x-\frac{1}{L}\nabla g(x)) \\
&\leq f(x) + \frac{L}{2}\|-\frac{1}{L}\nabla g(x)\|^2 + \nabla f(x)^T(-\frac{1}{L}\nabla g(x)) - \nabla f(y)^T(x-\frac{1}{L}\nabla g(x)) \\
&\leq f(x) + \frac{1}{2L}\|\nabla g(x)\|^2 - (\nabla f(x)-\nabla f(y))^T\frac{1}{L}\nabla g(x) - \nabla f(y)^T x \\
&= f(x) - \frac{1}{2L}\|\nabla f(x)-\nabla f(y)\|^2 - \nabla f(y)^T x \\
&= g(x) - \frac{1}{2L}\|\nabla f(x)-\nabla f(y)\|^2
\end{aligned}$$

Then,

$$\begin{aligned}
g(y) \leq g(x-\frac{1}{L}\nabla g(x)) = g(x) - \frac{1}{2L}\|\nabla f(x)-\nabla f(y)\|^2 \\
\Rightarrow g(y)-g(x) = f(y) - \nabla f(y)^T y - f(x) - \nabla f(y)^T x \leq -\frac{1}{2L}\|\nabla f(x)-\nabla f(y)\|^2
\end{aligned}$$

We can interchange x, y,

$$\begin{cases} f(y) - \nabla f(y)^T y - f(x) - \nabla f(y)^T x \leq -\frac{1}{2L}\|\nabla f(x)-\nabla f(y)\|^2 \\ f(x) - \nabla f(x)^T x - f(y) - \nabla f(x)^T y \leq -\frac{1}{2L}\|\nabla f(x)-\nabla f(y)\|^2 \end{cases}$$

Add these two inequalitys together,

$$(\nabla f(x)-\nabla f(y))^T(x-y) \geq \frac{1}{L}\|\nabla f(x)-\nabla f(y)\|^2$$

□

4.2 Convergence of Steepest Descent with Fixed Stepsize

4.2.1 Theorem: f has Lipschitz gradient $\Rightarrow \{x_k\}$ converges to stationary point

Theorem 9. *Consider the GD algorithm*

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

Assume that f has Lipschitz gradient with a Lipschitz gradient with Lipschitz constant L. Then if α is sufficiently small ($\alpha \in (0, \frac{2}{L})$) and $f(x) \geq f_{\min}$ for all $x \in \mathbb{R}^n$,
(1). $f(x_{k+1}) \leq f(x_k) - \alpha(1-\frac{L\alpha}{2})\|\nabla f(x_k)\|^2$
(2). $\sum_{k=0}^{N}\|\nabla f(x_k)\|^2 \leq \frac{f(x_0)-f_{\min}}{\alpha(1-\frac{L\alpha}{2})}$
(3). every limit point of $\{x_k\}$ is a stationary point of f.

证明. Applying the descent lemma,

$$
\begin{aligned}
f(x_{k+1}) &\leq f(x_k) + \nabla f(x_k)^T (x_{k+1} - x_k) + \frac{L}{2}\|x_{k+1} - x_k\|^2 \\
&= f(x_k) - \alpha \nabla f(x_k)^T \nabla f(x_k) + \frac{L}{2}\alpha^2 \|\nabla f(x_k)\|^2 \\
&= f(x_k) + \alpha(\frac{L\alpha}{2} - 1)\|\nabla f(x_k)\|^2
\end{aligned}
$$

$$
\begin{aligned}
\Rightarrow \alpha(1 - \frac{L\alpha}{2})\|\nabla f(x_k)\|^2 &\leq f(x_k) - f(x_{k+1}) \\
\alpha \sum_{k=0}^{N} (1 - \frac{L\alpha}{2})\|\nabla f(x_k)\|^2 &\leq f(x_0) - f(x_{N+1}) \\
&\leq f(x_0) - f_{\min}
\end{aligned}
$$

If $\alpha \in (0, \frac{2}{L})$, i.e. $\alpha(1 - \frac{L\alpha}{2})$,

$$
\begin{aligned}
&\sum_{k=0}^{N} \|\nabla f(x_k)\|^2 \leq \frac{f(x_0) - f_{\min}}{\alpha(1 - \frac{L\alpha}{2})} < \infty, \forall N \\
&\Rightarrow \lim_{k \to \infty} \nabla f(x_k) = 0
\end{aligned}
$$

If $\bar{x}$ is a limit point of $\{x_k\}$, $\lim_{k\to\infty} x_k = \bar{x}$.
By continunity of ∇f, $\nabla f(\bar{x}) = 0$ □

Example 9. *$f(x) = \frac{1}{2}x^2, x \in \mathbb{R}$, $\nabla f(x) = x$, Lipschitz with $L = 1$.*

$$
\begin{aligned}
x_{k+1} &= x_k - \alpha \nabla f(x_k) \\
&= x_k(1 - \alpha)
\end{aligned}
$$

$0 < \alpha < \frac{2}{L} = 2$ is needed for convergence.

Test (1) $\alpha = 1.5$ Then $x_{k+1} = x_k(-0.5)$,

$$\Rightarrow x_k = x_0(-0.5)^k \to 0 \text{ as } k \to \infty$$

Test (2) $\alpha = 2.5$ Then $x_{k+1} = x_k(-1.5)$.

$$\Rightarrow x_k = x_0(-1.5)^k \Rightarrow |x_k| \to \infty$$

Test (3) $\alpha = 2$ Then $x_{k+1} = -x_k$.

$$\Rightarrow x_k = (-1)^k x_0 \Rightarrow \text{ oscillation between } -x_0, x_0$$

Example 10. *What if gradient is not Lipschitz? e.g. $f(x) = x^4, x \in \mathbb{R}$, $\nabla f(x) = 4x^3$, $x = 0$ is the only stationary point (global-min)*

$$x_{k+1} = x_k - 4\alpha x_k^3 = x_k(1 - 4\alpha x_k^2)$$

- $|x_1| = |x_0|$, then $|x_k| = |x_0|$ for all k, and $\{x_k\}$ stays bounded away from 0, except if $x_0 = 0$

-
$$\begin{aligned}|x_1| < |x_0| &\Leftrightarrow |x_0||1 - 4\alpha x_0^2| < |x_0| \\ &\Leftrightarrow -1 < 1 - 4\alpha x_0^2 < 1 \\ &\Leftrightarrow 0 < x_0^2 < \frac{1}{2\alpha} \Leftrightarrow 0 < |x_0| < \frac{1}{\sqrt{2\alpha}}\end{aligned}$$

- Therefore, if $|x_1| < |x_0|$, then $|x_1| < |x_0| < \frac{1}{\sqrt{2\alpha}} \Rightarrow |x_2| < |x_1|, ..., |x_{k+1}| < |x_k|, \forall k \Rightarrow \{|x_k|\}$convergences

- And if $|x_1| > |x_0|$, then $|x_{k+1}| > |x_k|$ for all k and $\{x_k\}$ stays bounded away from 0.

Claim 3. $0 < |x_0| < \frac{1}{\sqrt{2\alpha}} \Rightarrow |x_k| \to 0$

证明. Suppose $|x_k| \to c > 0$. Then $\frac{|x_{k+1}|}{|x_k|} \to 1$
But $\frac{|x_{k+1}|}{|x_k|} = |1 - 4\alpha x_k^2| \to |1 - 4\alpha c^2|$. Thus $|1 - 4\alpha c^2| = 1 \Rightarrow c = \frac{1}{\sqrt{2\alpha}}$, which contradicts to $c < |x_0| < \frac{1}{\sqrt{2\alpha}}$, hence $c = 0$

□

4.3 Convergence of GD for convex functions

4.3.1 Theorem: f is convex and has Lipschitz gradient $\Rightarrow f(x_k)$ converges to global-min value with rate $\frac{1}{k}$

Theorem 10. *Consider the GD algorithm*

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

Assume that f has Lipschitz gradient with Lipschitz constant L. Further assume that

(a) f is a convex function.

(b) $\exists x^*$ s.t. $f(x^*) = \min f(x)$

Then for sufficiently small α:

(i) $\lim_{k\to\infty} f(x_k) = \min f(x) = f(x^*)$

(ii) $f(x_k)$ converges to $f(x^*)$ at rate $\frac{1}{k}$.

证明.
$$\begin{aligned}\|x_{k+1} - x^*\|^2 &= \|x_k - \alpha\nabla f(x_k) - x^*\|^2 \\ &= \|x_k - x^*\|^2 + \alpha^2\|\nabla f(x_k)\|^2 - 2\alpha\nabla f(x)^T(x_k - x^*)\end{aligned}$$

By convexity,
$$f(x^*) \geq f(x_k) + \nabla f(x_k)^T(x^* - x_k)$$
$$\Rightarrow \nabla f(x_k)^T(x^* - x_k) \leq f(x^*) - f(x_k)$$

Thus,

$$\begin{aligned}
\|x_{k+1} - x^*\|^2 &\le \|x_k - x^*\|^2 + \alpha^2\|\nabla f(x_k)\|^2 + 2\alpha(f(x^*) - f(x_k)) \\
\Rightarrow 2\alpha(f(x_k) - f(x^*)) &\le \|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2 + \alpha^2\|\nabla f(x_k)\|^2 \\
2\alpha\sum_{k=0}^{N}(f(x_k) - f(x^*)) &\le \|x_0 - x^*\|^2 - \|x_{N+1} - x^*\|^2 + \alpha^2\sum_{k=0}^{N}\|\nabla f(x_k)\|^2 \\
&\le \|x_0 - x^*\|^2 + \alpha^2\sum_{k=0}^{N}\|\nabla f(x_k)\|^2
\end{aligned}$$

According to previous theorm, if $\alpha \in (0, \frac{2}{L})$, $\sum_{k=0}^{N}\|\nabla f(x_k)\|^2 \le \frac{f(x_0)-f(x^*)}{\alpha(1-\frac{L\alpha}{2})}$ and

$$\begin{aligned}
f(x_{k+1}) - f(x_k) &\le -\alpha(1 - \frac{L\alpha}{2})\|\nabla f(x_k)\|^2 \le 0 \\
\Rightarrow f(x_N) &\le f(x_k), \quad \forall k = 0, 1..., N \\
\Rightarrow \sum_{k=0}^{N}(f(x_k) - f(x^*)) &\ge (N+1)(f(x_N) - f(x^*)) \\
f(x_N) - f(x^*) &\le \frac{1}{N+1}\sum_{k=0}^{N}(f(x_k) - f(x^*)) \\
&\le \frac{1}{2\alpha(N+1)}(\|x_0 - x^*\|^2 + \alpha^2\frac{f(x_0) - f(x^*)}{\alpha(1 - \frac{L\alpha}{2})}) \\
&\to 0 \text{ as } N \to \infty
\end{aligned}$$

The rate of convergence is $\frac{1}{N}$.
To make $f(x_N) - f(x^*) < \varepsilon$, we need $N \sim O(\frac{1}{\varepsilon})$. □

Note: Armijo's rule also convergencesat rate $\frac{1}{N}$ if ∇f is Lipschitz, without priot knowledge of L. But need $r \in [\frac{1}{2}, 1)$

4.4 Convergence of GD for strongly convex functions

Strong convexity with parameter m, along with $M-$Lipschitz gradient assumption (with $M \ge m$) According to the lemmas we proved before

$$\frac{m}{2}\|y - x\|^2 \le f(y) - f(x) - \nabla^T f(x)(y - x) \le \frac{M}{2}\|y - x\|^2$$

4.4.1 Theorem: Strongly convex, Lipschitz gradient $\Rightarrow$ $\{x_k\}$ converges to global-min geometrically

Theorem 11. *If f has Lipschitz gradient with Lipschitz constant L and strongly convex with parameter m, $\{x_k\}$ converges to x^** ***geometrically****.*

$$
\begin{aligned}
\|x_{k+1} - x^*\|^2 &= \|x_k - \alpha\nabla f(x_k) - x^*\|^2 \\
(\nabla f(x^*) = 0) \qquad &= \|(x_k - x^*) - \alpha(\nabla f(x_k) - \nabla f(x^*))\|^2 \\
&= \|x_k - x^*\|^2 + \alpha^2\|\nabla f(x_k) - \nabla f(x^*)\|^2 - 2\alpha(x_k - x^*)^T(\nabla f(x_k) - 0) \\
(\nabla f \text{ is M-Lipschitz}) \qquad &\le \|x_k - x^*\|^2 + \alpha^2 M^2\|x_k - x^*\|^2 + 2\alpha(x^* - x_k)^T\nabla f(x_k) \\
(\text{Strong convexity with } m) \qquad &\le \|x_k - x^*\|^2 + \alpha^2 M^2\|x_k - x^*\|^2 + 2\alpha(f(x^*) - f(x_k) - \frac{m}{2}\|x^* - x_k\|^2) \\
&= (1 + \alpha^2 M^2 - \alpha m)\|x_k - x^*\|^2 + 2\alpha(f(x^*) - f(x_k))
\end{aligned}
$$

By strong convexity of f

$$
\begin{aligned}
f(x_k) &\ge f(x^*) + \nabla^T f(x^*)(x_k - x^*) + \frac{m}{2}\|x_k - x^*\|^2 \\
&= f(x^*) + \frac{m}{2}\|x_k - x^*\|^2 \\
\Rightarrow f(x^*) - f(x_k) &\le -\frac{m}{2}\|x_k - x^*\|^2
\end{aligned}
$$

Then,

$$
\begin{aligned}
\|x_{k+1} - x^*\|^2 &\le (1 + \alpha^2 M^2 - \alpha m)\|x_k - x^*\|^2 + 2\alpha(-\frac{m}{2}\|x_k - x^*\|^2) \\
&\le (1 + \alpha^2 M^2 - 2\alpha m)\|x_k - x^*\|^2 \\
&\le (1 + \alpha^2 M^2 - 2\alpha m)^{k+1}\|x_0 - x^*\|^2 \\
\Rightarrow \|x_N - x^*\|^2 &\le (1 + \alpha^2 M^2 - 2\alpha m)^N\|x_0 - x^*\|^2
\end{aligned}
$$

If $\alpha \in (0, \frac{2m}{M^2})$, $1 + \alpha^2 M^2 - 2\alpha m < 1$. Then $x_N \to x^*$ **geometrically** as $N \to \infty$.
Note: Just having $0 < \alpha < \frac{2}{M}$ doesn't guarantee geometric convergence to x^*. e.g. $\alpha = \frac{1}{M} \Rightarrow 1 + \alpha^2 M^2 - 2m\alpha = 2(1 - \frac{m}{M}) \ge 1$ if $\frac{m}{M} \le 0.5$
To get the highest convergence rate:

$$
\begin{aligned}
1 + \alpha^2 M^2 - 2m\alpha &= (\alpha M)^2 - 2\alpha M\frac{m}{M} + 1 \\
&= (\alpha M - \frac{m}{M})^2 + 1 - \frac{m^2}{M^2}
\end{aligned}
$$

Which is minimized by setting

$$
\alpha = \alpha^* = \frac{m}{M^2}
$$

$$
\min_{\alpha > 0} 1 + \alpha^2 M^2 - 2m\alpha = 1 - \frac{m^2}{M^2} \in [0, 1)
$$

Since $M > m$, $\alpha^* = \frac{m}{M^2} < \frac{1}{M} < \frac{2}{M}$.
With $\alpha = \alpha^*$,

$$
\|x_N - x^*\|^2 \le (1 - \frac{m^2}{M^2})^N\|x_0 - x^*\|^2
$$

$\frac{M}{m}$ is called the **<u>condition number</u>**.

- If $\frac{M}{m} >> 1$, then $1 - \frac{m^2}{M^2}$ is close to 1 and convergence is slow.

- If $\frac{M}{m} = 1$, $\alpha^* = \frac{1}{M}$, and $x_N = x^*, \forall N \geq 1$. (Convergence in one step.)

Note that since $\nabla f(x^*) = 0$,

$$\begin{aligned} f(x_N) - f(x^*) &\leq \frac{M}{2}\|x_N - x^*\|^2 \\ &\leq (1 - \frac{m^2}{M^2})^N \frac{M}{2}\|x_0 - x^*\|^2 \end{aligned}$$

To make $f(x_N) - f(x^*) < \varepsilon$, we only need $N \sim O(log\frac{1}{\varepsilon})$ - called "linear" convergence.

4.4.2 Example

Example 11. $f(x) = \frac{1}{2}x^T Qx + b^T x + c, \quad Q \succ 0,\ \nabla^2 f(x) = Q.$

Let $\lambda_{\min}$ and $\lambda_{\max}$ be the min and max eigenvalue of Q. Then we know

$$\lambda_{\min}\|z\|^2 \leq z^T Qz \leq \lambda_{\max}\|z\|^2$$

Thus for all $z \in \mathbb{R}^n$

$$z^T(Q - \lambda_{\min}I)z \geq 0 \Rightarrow Q \succeq \lambda_{\min}I$$

Similarly, $Q \preceq \lambda_{\max}I$. Thus

$$\lambda_{\min}I \preceq \nabla^2 f(x) \preceq \lambda_{\max}I$$

$\lambda_{\min}I \preceq \nabla^2 f(x) \Leftrightarrow f$ is $\lambda_{\min}$-strongly convex; $\nabla^2 f(x) \preceq \lambda_{\max}I$ is a sufficient condition for f is $\lambda_{\max}$-smooth.
The condition number $= \frac{\lambda_{\max}}{\lambda_{\min}}$
Special Case: $Q = \mu I, \quad \mu > 0,\ \lambda_{\min} = \lambda_{\max} = \mu = m = M.$
$f(x) = \frac{\mu}{2}\|x\|^2 + b^T x + c$, $\nabla f(x) = \mu x + b$, $x^* = -\frac{b}{\mu}$, $\alpha^* = \frac{m}{M^2} = \frac{1}{\mu}$,

$$x_1 = x_0 - \alpha^* \nabla f(x_0) = x_0 - \frac{1}{\mu}(\mu x_0 + b) = -\frac{b}{\mu} = x^*$$

Convergence in one step!

4.5 Convergence of Gradient Descent on Smooth Strongly-Convex Functions

Still consider the constant stepsize gradient method

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

Lemma 4. *Suppose the sequences $\{\xi_k \in \mathbb{R}^p : k = 0, 1, \ldots\}$ and $\{u_k \in \mathbb{R}^p : k = 0, 1, 2, \ldots\}$ satisfy $\xi_{k+1} = \xi_k - \alpha u_k$. In addition, assume there is a martix M, the following inequality holds for all k*

$$\begin{bmatrix} \xi_k \\ u_k \end{bmatrix}^\top M \begin{bmatrix} \xi_k \\ u_k \end{bmatrix} \geq 0$$

If there exist $0<\rho<1$ *and* $\lambda \geq 0$ *such that*

$$\left[\begin{array}{cc}\left(1-\rho^2\right) I & -\alpha I \\ -\alpha I & \alpha^2 I\end{array}\right]+\lambda M$$

is a negative semidefinite matrix, then the sequence $\{\xi_k : k=0,1,\ldots\}$ *satisfies* $\|\xi_k\| \leq \rho^k \|\xi_0\|$.

证明. The key relation is

$$\|\xi_{k+1}\|^2=\|\xi_k-\alpha u_k\|^2=\|\xi_k\|^2-2\alpha(\xi_k)^T u_k+\alpha^2\|u_k\|^2=\left[\begin{array}{c}\xi_k \\ u_k\end{array}\right]^{\top}\begin{bmatrix}I & -\alpha I \\ -\alpha I & \alpha^2 I\end{bmatrix}\left[\begin{array}{c}\xi_k \\ u_k\end{array}\right]$$

Since $\left[\begin{array}{cc}\left(1-\rho^2\right) I & -\alpha I \\ -\alpha I & \alpha^2 I\end{array}\right]+\lambda M$ is negative semidefinite, we have

$$\left[\begin{array}{c}\xi_k \\ u_k\end{array}\right]^{\top}\left(\left[\begin{array}{cc}\left(1-\rho^2\right) I & -\alpha I \\ -\alpha I & \alpha^2 I\end{array}\right]+\lambda M\right)\left[\begin{array}{c}\xi_k \\ u_k\end{array}\right] \leq 0$$

Expand the inequality,

$$\left[\begin{array}{c}\xi_k \\ u_k\end{array}\right]^{\top}\left[\begin{array}{cc}I & -\alpha I \\ -\alpha I & \alpha^2 I\end{array}\right]\left[\begin{array}{c}\xi_k \\ u_k\end{array}\right]+\left[\begin{array}{c}\xi_k \\ u_k\end{array}\right]^{\top}\left[\begin{array}{cc}-\rho^2 I & 0 \\ 0 & 0\end{array}\right]\left[\begin{array}{c}\xi_k \\ u_k\end{array}\right]+\lambda\left[\begin{array}{c}\xi_k \\ u_k\end{array}\right]^{\top} M\left[\begin{array}{c}\xi_k \\ u_k\end{array}\right] \leq 0$$

Applying the key relation

$$\|\xi_{k+1}\|^2-\rho^2\|\xi_k\|^2+\lambda\left[\begin{array}{c}\xi_k \\ u_k\end{array}\right]^{\top} M\left[\begin{array}{c}\xi_k \\ u_k\end{array}\right] \leq 0$$

$$\|\xi_{k+1}\|^2-\rho^2\|\xi_k\|^2 \leq -\lambda\left[\begin{array}{c}\xi_k \\ u_k\end{array}\right]^{\top} M\left[\begin{array}{c}\xi_k \\ u_k\end{array}\right] \leq 0$$

Hence, $\|\xi_{k+1}\| \leq \rho\|\xi_k\|$ for all k. Therefore, we have $\|\xi_k\| \leq \rho^k \|\xi_0\|$. □

Theorem 12. *Suppose* f *is* L*-smooth and* m*-strongly convex. Let* x^* *be the unique global min. Given a stepsize* α*,if there exists* $0<\rho<1$ *and* $\lambda \geq 0$ *such that*

$$\begin{bmatrix}1-\rho^2 & -\alpha \\ -\alpha & \alpha^2\end{bmatrix}+\lambda\begin{bmatrix}-2mL & m+L \\ m+L & -2\end{bmatrix}$$

is a negative semidefinite matrix, then the gradient method satisfies

$$\|x_k-x^*\| \leq \rho^k\|x_0-x^*\|$$

证明. We set f is L-smooth and m-strongly convex,
According to the definition of m-strongly convex

$$(\nabla f(x)-\nabla f(y))^T(x-y) \geq m\|x-y\|^2$$

And the co-coercivity condition, if f is $L-$smooth,

$$(\nabla f(x) - \nabla f(y))^T(x-y) \geq \frac{1}{L}\|\nabla f(x) - \nabla f(y)\|^2$$

Set $g(x) = f(x) - \frac{m}{2}\|x\|^2$, $\nabla g(x) = \nabla f(x) - mx$.

$$\begin{aligned} f \text{ is } L-\text{smooth} &\Leftrightarrow \|\nabla f(x) - \nabla f(y)\| \leq L\|x-y\| \\ &\Leftrightarrow \|\nabla g(x) - \nabla g(y)\| \leq (L-m)\|x-y\| \\ &\Leftrightarrow g \text{ is } L-m\text{-smooth} \end{aligned}$$

Hence,

$$\begin{gathered} (\nabla g(x) - \nabla g(y))^T(x-y) \geq \frac{1}{L-m}\|\nabla g(x) - \nabla g(y)\|^2 \\ (\nabla f(x) - \nabla f(y) - m(x-y))^T(x-y) \geq \frac{1}{L-m}\|\nabla f(x) - \nabla f(y) - m(x-y)\|^2 \\ (L-m)[(\nabla f(x) - \nabla f(y))^T(x-y) - m\|x-y\|^2] \\ \geq \|\nabla f(x) - \nabla f(y)\|^2 + m^2\|(x-y)\|^2 - 2m(\nabla f(x) - \nabla f(y))^T(x-y) \\ (L+m)(\nabla f(x) - \nabla f(y))^T(x-y) \geq mL\|x-y\|^2 + \|\nabla f(x) - \nabla f(y)\|^2 \\ \Rightarrow (\nabla f(x) - \nabla f(y))^T(x-y) \geq \frac{mL}{m+L}\|x-y\|^2 + \frac{1}{m+L}\|\nabla f(x) - \nabla f(y)\|^2 \end{gathered}$$

Which can be rewritten as

$$\begin{bmatrix} x-y \\ \nabla f(x) - \nabla f(y) \end{bmatrix}^T \begin{bmatrix} -2mLI & (m+L)I \\ (m+L)I & -2I \end{bmatrix} \begin{bmatrix} x-y \\ \nabla f(x) - \nabla f(y) \end{bmatrix} \geq 0$$

Let $y = x^*$ and $\nabla f(y) = \nabla f(x^*) = 0$

$$\begin{bmatrix} x-x^* \\ \nabla f(x) \end{bmatrix}^T \begin{bmatrix} -2mLI & (m+L)I \\ (m+L)I & -2I \end{bmatrix} \begin{bmatrix} x-x^* \\ \nabla f(x) \end{bmatrix} \geq 0$$

Set $\xi_k = x_k - x^*$ and $u_k = \nabla f(x_k)$. And $\xi_{k+1} = x_{k+1} - x^* = x_k - \alpha\nabla f(x_k) - x^* = \xi_k - \alpha u_k$

$$\begin{bmatrix} \xi_k \\ u_k \end{bmatrix}^T \begin{bmatrix} -2mLI & (m+L)I \\ (m+L)I & -2I \end{bmatrix} \begin{bmatrix} \xi_k \\ u_k \end{bmatrix} \geq 0$$

Choose $M = \begin{bmatrix} -2mLI & (m+L)I \\ (m+L)I & -2I \end{bmatrix}$. Then prove by previous lemma. $\square$

Now we apply the theorem to obtain the convergence rate ρ for the gradient method with various stepsize choices.

- Case 1: If we choose $\alpha = \frac{1}{L}, \rho = 1 - \frac{m}{L}$, and $\lambda = \frac{1}{L^2}$, we have

$$\begin{bmatrix} 1-\rho^2 & -\alpha \\ -\alpha & \alpha^2 \end{bmatrix} + \lambda \begin{bmatrix} -2mL & m+L \\ m+L & -2 \end{bmatrix} = \begin{bmatrix} -\frac{m^2}{L^2} & \frac{m}{L^2} \\ \frac{m^2}{L^2} & -\frac{1}{L^2} \end{bmatrix} = \frac{1}{L^2} \begin{bmatrix} -m^2 & m \\ m & -1 \end{bmatrix}$$

The right side is clearly negative semidefinite due to the fact that $\begin{bmatrix} a \\ b \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} -m^2 & m \\ m & -1 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = -(ma-b)^2 \leq 0$. Therefore, the gradient method with $\alpha = \frac{1}{L}$ converges as

$$\|x_k - x^*\| \leq \left(1 - \frac{m}{L}\right)^k \|x_0 - x^*\|$$

- Case 2: If we choose $\alpha = \frac{2}{m+L}, \rho = \frac{L-m}{L+m}$, and $\lambda = \frac{2}{(m+L)^2}$, we have

$$\begin{bmatrix} 1-\rho^2 & -\alpha \\ -\alpha & \alpha^2 \end{bmatrix} + \lambda \begin{bmatrix} -2mL & m+L \\ m+L & -2 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$$

  The zero matrix is clearly negative semidefinite. Therefore, the gradient method with $\alpha = \frac{2}{m+L}$ converges as

$$\|x_k - x^*\| \leq \left(\frac{L-m}{L+m}\right)^k \|x_0 - x^*\|$$

Notice $L \geq m > 0$ and hence $1 - \frac{m}{L} \geq \frac{L-m}{L+m}$. This means the gradient method with $\alpha = \frac{2}{m+L}$ converges slightly faster than the case with $\alpha = \frac{1}{L}$. However, m is typically unknown in practice. The step choice of $\alpha = \frac{1}{L}$ is also more robust. The most popular choice for α is still $\frac{1}{L}$.
We can further express ρ as a function of α. To do this, we need to choose λ carefully for a given α. If we choose λ reasonably, we can show the best value for ρ that we can find is $\max\{|1-m\alpha|, |L\alpha - 1|\}$.

4.6 From convergence rate to iteration complexity

The convergence rate ρ naturally leads to an iteration number T guaranteeing the algorithm to achieve the so-called **ε-optimality**, i.e. $\|x_T - x^*\| \leq \varepsilon$.
To guarantee $\|x_T - x^*\| \leq \varepsilon$, we can use the bound $\|x_T - x^*\| \leq \rho^T \|x_0 - x^*\|$. If we choose T such that $\rho^T \|x_0 - x^*\| \leq \varepsilon$, then we guarantee $\|x_T - x^*\| \leq \varepsilon$. Denote $c = \|x_0 - x^*\|$. Then $c\rho^k \leq \varepsilon$ is equivalent to

$$\log c + k \log \rho \leq \log(\varepsilon)$$

Notice $\rho < 1$ and $\log \rho < 0$. The above inequality is equivalent to

$$k \geq \log\left(\frac{\varepsilon}{c}\right) / \log \rho = \log\left(\frac{c}{\varepsilon}\right) / (-\log \rho)$$

So if we choose $T = \log\left(\frac{c}{\varepsilon}\right)/(-\log \rho)$, we guarantee $\|x_T - x^*\| \leq \varepsilon$. Notice $\log \rho \leq \rho - 1 < 0$ (this can be proved using the concavity of log function and we will talk about concavity in later lectures), so $\frac{1}{1-\rho} \geq -\frac{1}{\log \rho}$ and we can also choose $T = \log\left(\frac{c}{\varepsilon}\right)/(1-\rho) \geq \log\left(\frac{c}{\varepsilon}\right)/(-\log \rho)$ to guarantee $\|x_T - x^*\| \leq \varepsilon$.
Another interpretation for $T = \log\left(\frac{c}{\varepsilon}\right)/(1-\rho)$ is that a first-order Taylor expansion of $-\log \rho$ at $\rho = 1$ leads to $-\log \rho \approx 1 - \rho$. So $\log\left(\frac{c}{\varepsilon}\right)/(-\log \rho)$ is roughly equal to $\log\left(\frac{c}{\varepsilon}\right)/(1-\rho)$ when ρ is close to 1 .

Clearly the smaller T is, the more efficient the optimization method is. The iteration number T describes the "ε-optimal iteration complexity" of the gradient method for smooth strongly-convex objective functions.

- For the gradient method with $\alpha = \frac{1}{L}$, we have $\rho = 1 - \frac{m}{L} = 1 - \frac{1}{\kappa}$ and hence $T = \log\left(\frac{c}{\varepsilon}\right)/(1-\rho) = \kappa \log\left(\frac{c}{\varepsilon}\right) = O\left(\kappa \log\left(\frac{1}{\varepsilon}\right)\right)$ .2 Here we use the big O notation to highlight the dependence on κ and ε and hide the dependence on the constant c.
- For the gradient method with $\alpha = \frac{2}{L+m}$, we have $\rho = \frac{\kappa-1}{\kappa+1} = 1 - \frac{2}{\kappa+1}$ and hence $T = \log\left(\frac{c}{\varepsilon}\right)/(1-\rho) = \frac{\kappa+1}{2}\log\left(\frac{c}{\varepsilon}\right)$. Although $\frac{\kappa+1}{2} \leq \kappa$, we still have $\frac{\kappa+1}{2}\log\left(\frac{c}{\varepsilon}\right) = O\left(\kappa \log\left(\frac{1}{\varepsilon}\right)\right)$. Therefore, the stepsize $\alpha = \frac{2}{m+L}$ can only improve the constant C hidden in the big O notation of the iteration complexity. People call this "improvement of a constant factor".
- In general, when ρ has the form $\rho = 1 - 1/(a\kappa + b)$, the resultant iteration complexity is always $O\left(\kappa \log\left(\frac{1}{\varepsilon}\right)\right)$.

There are algorithms which can significantly decrease the iteration complexity for unconstrained optimization problems with smooth strongly-convex objective functions. For example, Nesterov's method can decrease the iteration complexity from $O\left(\kappa \log\left(\frac{1}{\varepsilon}\right)\right)$ to $O\left(\sqrt{\kappa} \log\left(\frac{1}{\varepsilon}\right)\right)$. Momentum is used to accelerate optimization as:

$$x_{k+1} = x_k - \alpha \nabla f\left((1+\beta)x_k - \beta x_{k-1}\right) + \beta\left(x_k - x_{k-1}\right).$$

5 Newton's Method

One dimential:

Finding solution to non-linear equation:

$$g(x^*) = 0$$

with $g : \mathbb{R} \to \mathbb{R}$. Given x_k, find x_{k+1} to solve x^*.

$$0 = g(x_{k+1}) \approx g(x_k) + g'(x_k)(x_{k+1} - x_k)$$

Assuming $g'(x_k) \neq 0$, set

$$x_{k+1} = x_k - (g'(x_k))^{-1} g(x_k)$$

5.1 Generalization to Optimization

In optimization, the goal is to get to x s.t. $\nabla f(x) = 0$.

Given x_k, we want to find x_{k+1} s.t. $\nabla f(x_{k+1}) = 0$.

Taylor's Approx:

$$\nabla f(x_{k+1}) \approx \nabla f(x_k) + \nabla^2 f(x_k)(x_{k+1} - x_k)$$

Set

$$x_{k+1} = x_k - (\nabla^2 f(x_k))^{-1}\nabla f(x_k)$$

, which can be viewed as GD with $\alpha_k = 1$ and $d_k = -(\nabla^2 f(x_k))^{-1}\nabla f(x_k)$
If $\nabla^2 f(x_k) \succeq 0$, then $\nabla f(x_k)^T d_k \geq 0$.

5.2 A New Interpretation of Newton's Method

Since $f(x) \approx f(x_k) + \nabla^T f(x_k)(x - x_k) + \frac{1}{2}(x - x_k)^T \nabla^2 f(x_k)(x - x_k)$, at each step k, we can solve a quadratic minimization problem,

$$x_{k+1} = \underset{x\in\mathbb{R}^p}{\operatorname{argmin}}\{f(x_k) + \nabla^T f(x_k)(x - x_k) + \frac{1}{2}(x - x_k)^T \nabla^2 f(x_k)(x - x_k)\}$$

5.3 Convergence of Newton's Method

Let x^* be s.t. $\nabla f(x^*) = 0$, then

$$\begin{aligned}\|x_{k+1} - x^*\| &= \|x_k - x^* - (\nabla^2 f(x_k))^{-1}\nabla f(x_k)\| \\ &= \|x_k - x^* - (\nabla^2 f(x_k))^{-1}\nabla(f(x_k) - \nabla f(x^*))\|\end{aligned}$$

By Taylor's theorem,

$$\nabla f(x_k) = \nabla f(x^*) + \nabla^2 f(x^* + \beta(x_k - x^*))(x_k - x^*) \text{ for some } \beta \in [0, 1]$$

Thus,

$$\begin{aligned}\|x_{k+1} - x^*\| &= \|x_k - x^* - (\nabla^2 f(x_k))^{-1}\nabla^2 f(x^* + \beta(x_k - x^*))(x_k - x^*)\| \\ &= \|(\nabla^2 f(x_k))^{-1}(\nabla^2 f(x^* + \beta(x_k - x^*)) - \nabla^2 f(x_k))(x_k - x^*)\| \\ &\leq \|(\nabla^2 f(x_k))^{-1}\|\|\nabla^2 f(x^* + \beta(x_k - x^*)) - \nabla^2 f(x_k)\|\|x_k - x^*\|\end{aligned}$$

We use 1-norm $\|A\| = \max_{x\neq 0}\frac{\|Ax\|}{\|x\|}$ here, $\|A\| \geq \frac{\|Ax\|}{\|x\|} \Rightarrow \|Ax\| \leq \|A\|\|x\|$.
Easy to prove, for symmetric $A \succeq 0$, $\|A\| = \lambda_{\max}(A)$, $\|A^{-1}\| = \lambda_{\max}(A^{-1}) = \lambda_{\min}^{-1}(A)$

- Now suppose **f is loacl m-strongly convex near x^***, then

$$\begin{aligned}\nabla^2 f(x^*) \succeq mI &\text{ with } m > 0 \\ \Rightarrow \lambda_{\min}(\nabla^2 f(x^*)) &\geq m > 0 \\ \Rightarrow \lambda_{\min}^{-1}(\nabla^2 f(x^*)) &\leq \frac{1}{m}\end{aligned}$$

- When f is not local strongly convex near x^*. Assuming $\nabla^2 f(x)$ is continuous, if $\|x_k - x^*\|$ is small, then $\lambda_{\min}(\nabla^2 f(x_k))$ is close to $\lambda_{\min}(\nabla^2 f(x^*))$ i.e $\lambda_{\min}(\nabla^2 f(x^*))$ should be greater than a constant $\lambda_{\min}(\nabla^2 f(x^*)) \geq \bar{\gamma} > 0$. Then,

$$\|\nabla^2 f(x_k)^{-1}\| = \lambda_{\min}^{-1}(\nabla^2 f(x_k)) \leq \frac{1}{\bar{\gamma}} = \gamma$$

Furthurmore, assume that $\nabla^2 f$ **is L-Lipschitz in a neighborhood** $\&$ **of** x^*, i.e.

$$\|\nabla^2 f(x) - \nabla^2 f(y)\| \leq L\|x - y\| \quad \forall x, y \in \&$$

Thus,

$$\begin{aligned}
\|x_{k+1} - x^*\| &\leq \|(\nabla^2 f(x_k))^{-1}\| \|\nabla^2 f(x^* + \beta(x_k - x^*)) - \nabla^2 f(x_k)\| \|x_k - x^*\| \\
&\leq \gamma L \|x^* + \beta(x_k - x^*) - x_k\| \|x_k - x^*\| \\
&\leq \gamma L \|(\beta - 1)(x_k - x^*)\| \|x_k - x^*\| \\
\text{(Since } \beta \in [0,1]) \quad &\leq \gamma L \|x_k - x^*\|^2
\end{aligned}$$

Hence,

$$\|x_{k+1} - x^*\| \leq \gamma L \|x_k - x^*\|^2$$

Now suppose x_0 **is close enough to** x^* s.t.

$$\gamma L \|x_0 - x^*\| = \sigma < 1$$

Then,

$$\begin{aligned}
\|x_1 - x^*\| &\leq \sigma \|x_0 - x^*\| \\
\|x_2 - x^*\| &\leq \gamma L \|x_1 - x^*\|^2 \\
&\leq \gamma L \sigma^2 \|x_0 - x^*\|^2 = \sigma^3 \|x_0 - x^*\| \\
\|x_3 - x^*\| &\leq \gamma L \|x_2 - x^*\|^2 \\
&\leq \gamma L \sigma^6 \|x_0 - x^*\|^2 = \sigma^7 \|x_0 - x^*\| \\
&\dots \\
\|x_N - x^*\| &\leq \sigma^{2^N - 1} \|x_0 - x^*\|
\end{aligned}$$

Assuming ∇f **is M-Lipschitz in neighborhood of** x^*,

$$\begin{aligned}
f(x_N) - f(x^*) &\leq \nabla f(x^*)(x_N - x^*) + \frac{M}{2}\|x_N - x^*\|^2 \\
&\leq \frac{M}{2}\sigma^{(2^{N+1}-2)}\|x_N - x^*\|^2
\end{aligned}$$

Thus to make $f(x_N) - f(x^*) < \varepsilon$, need $N \sim O(log(log(\frac{1}{\varepsilon})))$
We call it **order-2 or super-linear convergence**.

5.4 Note: Cons and Pros

- Newton's Method is super-fast close to local min if function strongly convex around min.

- If the function is quadratic, Newton's method converges in one step.

$$f(x) = \frac{1}{2}x^T Q x + bx + c, \quad Q \succ 0.$$

$$\nabla f(x) = Qx + b, \nabla^2 f(x) = Q.$$

$$\text{Global min } x^* \text{ satisfies } Qx^* + b = 0 \Rightarrow x^* = -Q^{-1}b$$

Newton's method: for any $x_0 \in \mathbb{R}^n$,

$$\begin{aligned} x_1 &= x_0 - (\nabla^2 f(x_0))^{-1} \nabla f(x_0) \\ &= x_0 - Q^{-1}(Qx_0 + b) = -Q^{-1}b = x^* \end{aligned}$$

Intuition: when f is a quadratic function, $\nabla^3 f(x) = 0, \forall x$. Hence, $f(x) = f(x_k) + \nabla^T f(x_k)(x - x_k) + \frac{1}{2}(x - x_k)^T \nabla^2 f(x_k)(x - x_k)$, the minimization problem will get the min in one step.

- But Newton's method has several **drawbacks**:

(1) Newton's method requires the matrix inversion step, and this is quite expensive. So the per step cost for Newton's method is higher.

(2) Newton's method has faster local convergence but may diverge if initialized from some place far from the optimal point.

(3) $\nabla^2 f(x)^{-1}$ may fail to exist, i.e. $\nabla^2 f(x)$ is singular, e.g. linear f.

(4) It is not necessarily a general GD method since $\nabla^2 f(x_k)$ may not be $\succ 0$.

(5) It is not a descent method, $f(x_{k+1})$ may be $> f(x_k)$.

(6) It may stop at local max or saddlepoints.

5.5 Modifications to ensure global convergence

(a) Try Newton's method. If either $\nabla^2 f(x_k)$ is singular or $f(x_{k+1}) > f(x_k)$ then use (b).

(b) Find δ_k s.t.

$$(\delta_k I + \nabla^2 f(x_k)) \succ 0$$

and

$$\lambda_{\min}(\delta_k I + \nabla^2 f(x_k)) \succeq \Delta > 0$$

so that $\delta_k I + \nabla^2 f(x_k)$ is easily invertible.

Then set $d_k = -(\delta_k I + \nabla^2 f(x_k))^{-1} \nabla f(x_k)$. This ensures that $\nabla^T f(x_k) d_k < 0$.

Then we use $x_{k+1} = x_k + \alpha_k d_k$ with α_k chosen using Armijo's Rule.

If at any point $\nabla^2 f(x_k) \succ 0$, go back to Newton's method and check if $f(x_{k+1}) < f(x_k)$. Continue Newton's method as long as $\nabla^2 f(x_k) \succ 0$ and $f(x_{k+1}) < f(x_k)$.

5.6 Quasi-Newton Methods

Estimating Hessian $\nabla^2 f(x_k)$ is expensive, so we use some simplier matrix H_k instead.
Quasi-Newton method have the iteration form:

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

where H_k is some estimated version of $\nabla^2 f(x_k)$, and the stepsize α_k is typically determined by Armijo rule.
Previously, we approximate $f(x)$ by

$$f(x) \approx f(x_k) + \nabla^T f(x_k)(x - x_k) + \frac{1}{2}(x - x_k)^T \nabla^2 f(x_k)(x - x_k)$$

Now, we define the form by H_k

$$g(x) = f(x_k) + \nabla^T f(x_k)(x - x_k) + \frac{1}{2}(x - x_k)^T H_k(x - x_k)$$

We hope $g(x) \approx f(x)$ and optimize g for this step. We enforce

(1) $\nabla f(x_k) = \nabla g(x_k)$ (Automatically satisfied)

(2) $\nabla f(x_{k-1}) = \nabla g(x_{k-1}) \quad \Leftrightarrow$

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

The condition (2) is called the secant equation.
There are infinitely many H_k satisfying this condition. Various choices of H_k lead to different Quasi-Newton methods. We discuss the BFGS method.

5.6.1 BFGS Method

We need H_k to be constructed in a way that it can be efficiently computed.
We want H_k to have two properties:

(1) H_k can be computed by some iterative formula

$$H_k = H_{k-1} + M_{k-1}$$

(2) H_k is positive definite (at least guarantee that the BFGS method is a descent method, i.e. $f(x_{k+1}) \leq f(x_k)$).

We can choose $H_0 > 0$ and then guarantee $M_k \geq 0$.
Rank-2 BFGS Method:

$$H_{k+1} = H_k + a_k v_k v_k^T + b_k u_k u_k^T$$

where $v_k \in \mathbb{R}^p$ and $u_k \in \mathbb{R}^p$ are some vectors. If $H_0 > 0$, the above iterative formula can guarantee H_k to be positive definite.
How can we choose v_k and u_k to guarantee the secant equation $H_{k+1}(x_{k+1} - x_k) = \nabla f(x_{k+1}) - \nabla f(x_k)$?
Let's denote $s_k = x_{k+1} - x_k$ and $y_k = \nabla f(x_{k+1}) - \nabla f(x_k)$. The secant equation: $H_{k+1} s_k = y_k$, then substitute it into the above formula,

$$y_k = H_{k+1} s_k = H_k s_k + a_k v_k v_k^T s_k + b_k u_k u_k^T s_k$$
$$\Leftrightarrow y_k - H_k s_k = a_k (v_k^T s_k) v_k + b_k (u_k^T s_k) u_k$$

To let the above equation be satisfied. We let $v_k = y_k$, $u_k = H_k s_k$, $a_k = \frac{1}{y_k^T s_k}$, and $b_k = -\frac{1}{s_k^T H_k s_k}$. Then, the iteration formula becomes

$$H_{k+1} = H_k + \frac{y_k y_k^T}{y_k^T s_k} - \frac{H_k s_k s_k^T H_k}{s_k^T H_k s_k}$$

where $s_k = x_{k+1} - x_k$ and $y_k = \nabla f(x_{k+1}) - \nabla f(x_k)$.

This is exactly the BFGS method.

Since we implement the BFGS method as

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

It will be better to compute H_k^{-1} directly instead of H_k.

$$\begin{aligned}
H_{k+1}^{-1} &= \left(H_k + \frac{y_k y_k^T}{y_k^T s_k} - \frac{H_k s_k s_k^T H_k}{s_k^T H_k s_k} \right)^{-1} \\
&= \left(H_k + [H_k s_k \ y_k] \begin{bmatrix} -\frac{1}{s_k^T H_k s_k} & 0 \\ 0 & \frac{1}{y_k^T s_k} \end{bmatrix} \begin{bmatrix} s_k^T H_k \\ y_k^T \end{bmatrix} \right)^{-1} \\
&\quad \text{(by woodbury formula)} \\
&= H_k^{-1} - H_k^{-1} [H_k s_k \ y_k] \left(\begin{bmatrix} -\frac{1}{s_k^T H_k s_k} & 0 \\ 0 & \frac{1}{y_k^T s_k} \end{bmatrix}^{-1} + \begin{bmatrix} s_k^T H_k \\ y_k^T \end{bmatrix} H_k^{-1} [H_k s_k \ y_k] \right)^{-1} \begin{bmatrix} s_k^T H_k \\ y_k^T \end{bmatrix} H_k^{-1} \\
&= H_k^{-1} - [s_k \ H_k^{-1} y_k] \begin{bmatrix} 0 & s_k^T y_k \\ y_k^T s_k & y_k^T (s_k + H_k^{-1} y_k) \end{bmatrix}^{-1} \begin{bmatrix} s_k^T \\ y_k^T H_k^{-1} \end{bmatrix} \\
&= H_k^{-1} - [s_k \ H_k^{-1} y_k] \begin{bmatrix} -\frac{y_k^T s_k + y_k^T H_k^{-1} y_k}{y_k^T s_k s_k^T y_k} & \frac{1}{y_k^T s_k} \\ \frac{1}{y_k^T s_k} & 0 \end{bmatrix} \begin{bmatrix} s_k^T \\ y_k^T H_k^{-1} \end{bmatrix} \\
&= H_k^{-1} - \frac{H_k^{-1} y_k s_k^T}{y_k^T s_K} - \frac{s_k y_k^T H_k^{-1}}{y_k^T s_K} + \frac{s_k s_k^T}{y_k^T s_K} + \frac{s_k y_k^T H_k^{-1} y_k s_k^T}{(y_k^T s_k)^2} \\
&= \left(I - \frac{s_k y_k^T}{y_k^T s_k} \right) H_k^{-1} \left(I - \frac{y_k s_k^T}{y_k^T s_k} \right) + \frac{s_k s_k^T}{y_k^T s_k}
\end{aligned}$$

$$H_{k+1}^{-1} = \left(I - \frac{s_k y_k^T}{y_k^T s_k} \right) H_k^{-1} \left(I - \frac{y_k s_k^T}{y_k^T s_k} \right) + \frac{s_k s_k^T}{y_k^T s_k}$$

is the iteration computation H_k^{-1} of BFGS method. where $s_k = x_{k+1} - x_k$ and $y_k = \nabla f(x_{k+1}) - \nabla f(x_k)$.

5.7 Trust-Region Method

$$x_{k+1} = \underset{\|x - x_k\| \le \Delta_k}{\operatorname{argmin}} \left\{ f(x_k) + \nabla^T f(x_k)(x - x_k) + \frac{1}{2}(x - x_k)^T \nabla^2 f(x_k)(x - x_k) \right\}$$

This method can escape addle points under some assumptions.

5.8 Cubic Regularization

Contain higher order term $\|x - x_k\|^3$ to the quadratic estimation.

6 Neural Networks

6.1 Neuron

Neuron Neuron is a non-linear function, which takes $\mathfrak{J} \in \mathbb{R}$ as input and produce $\sigma(\mathfrak{J}) \in \mathbb{R}$ as output.

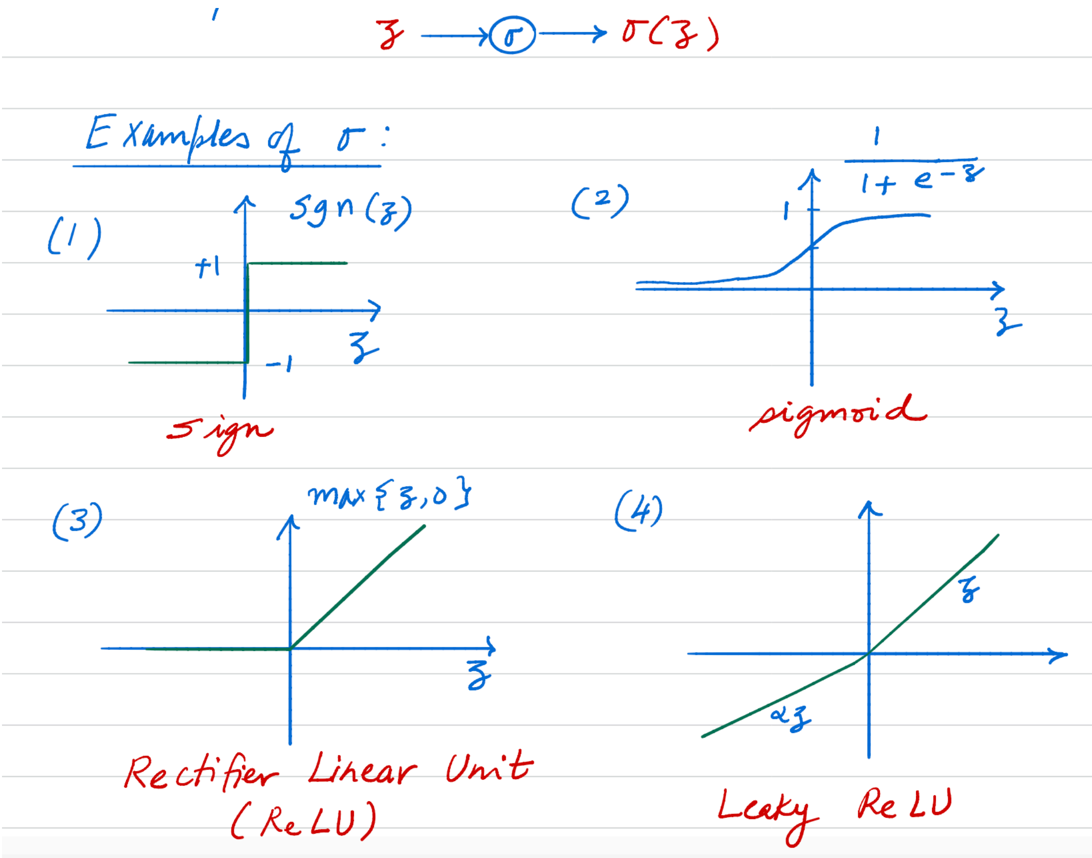


图 2: Neuron Examples

Vector Input

The output is

$$\sigma(\omega^T x + b)$$

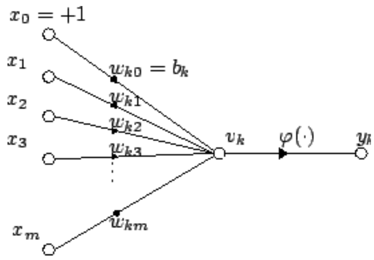


图 3: Vector Input

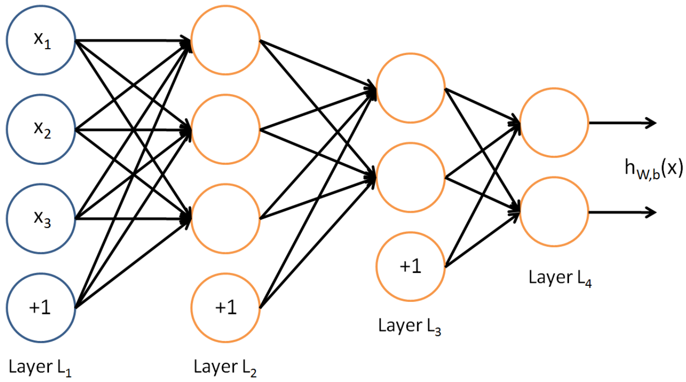


图 4: Multilayer Neural Network

6.2 Multilayer Neural Network

- Number of neurons in each layer can be different.
- All weights on edge connecting layers $m-1$ and m is matrix $W^{(m)}$, with $w_{ij}^{(m)}$ being the weight connecting output j of layer $m-1$ with neuron i of layer m.
- Input to network is vector x; output of layer m is vector $y^{(m)}$

$$y_i^{(1)} = \sigma(x_i^{(1)}), \text{ with } x_i^{(1)} = \sum_j w_{ij}^{(1)} x_j + b_i^{(1)}$$

$$y^{(1)} = \sigma(x^{(1)}), \text{ with } x^{(1)} = W^{(1)} x + b^{(1)}$$

$$y^{(2)} = \sigma(x^{(2)}), \text{ with } x^{(2)} = W^{(2)} y^{(1)} + b^{(2)}$$

$$\vdots$$

$$y^{(M)} = \sigma(x^{(M)}), \text{ with } x^{(M)} = W^{(M)} y^{(M-1)} + b^{(M)}$$

We want to find the weights $W^{(1)}, \cdots, W^{(M)}, b^{(1)}, \cdots, b^{(M)}$ so that the output of last layer

$$\hat{y} = y^{(M)} \approx f^*(x) = y$$

$f^*(x)$ is the unknown thing we need to predict.

We use labelled training data, i.e.

$$(x[1], y[1]), (x[2], y[2]), \cdots (x[N], y[N])$$

Minimize the "empirical" loss on training data.

$$J = \sum_{i=1}^{N} L(y[i], \hat{y}[i])$$

where $\bar{y}[i]$ is the output of NN whose input is $x[i]$.

- L is the function of $W^{(1)}, \cdots, W^{(M)}, b^{(1)}, \cdots, b^{(M)}$ to measure the loss. e.g. the square loss

$$L(y, \hat{y}) = (y - \hat{y})^2$$

- We wish to minimize J using a gradient descent procedure.

- To compute gradient we need:

$$\frac{\partial L}{\partial w_{ij}^{(l)}} \text{ for each } l, i, j; \quad \frac{\partial L}{\partial b_i^{(l)}} \text{ for each } l, i.$$

6.3 Back Propagation Algorithm

Recall $y_i^{(m)} = \sigma(x_i^{(m)})$, $x_i^{(m)} = \sum_j w_{ij}^{(m)} y_j^{(m-1)} + b_i^{(m)}$

$$\frac{\partial L}{\partial w_{ij}^{(m)}} = \frac{\partial L}{\partial y_i^{(m)}} \cdot \frac{\partial y_i^{(m)}}{\partial w_{ij}^{(m)}} = \frac{\partial L}{\partial y_i^{(m)}} \cdot \frac{\partial y_i^{(m)}}{\partial x_i^{(m)}} \cdot \frac{\partial x_i^{(m)}}{\partial w_{ij}^{(m)}}$$

$$\frac{\partial L}{\partial b_i^{(m)}} = \frac{\partial L}{\partial y_i^{(m)}} \cdot \frac{\partial y_i^{(m)}}{\partial x_i^{(m)}} \cdot \frac{\partial x_i^{(m)}}{\partial b_i^{(m)}}$$

For large M,

- $\frac{\partial L}{\partial y_i^{(M)}}$ is easy to compute.

- $\frac{\partial y_i^{(M)}}{\partial x_i^{(M)}} = \frac{\partial \sigma(x_i^{(M)})}{\partial x_i^{(M)}} = \sigma'(x_i^{(M)})$, (assuming σ differentiable).

- $\frac{\partial x_i^{(M)}}{\partial w_{ij}^{(M)}} = y_j^{(M-1)}$

Thus,

$$\frac{\partial L}{\partial w_{ij}^{(M)}} = \frac{\partial L}{\partial y_i^{(M)}} \cdot \sigma'(x_i^{(M)}) \cdot y_j^{(M-1)}$$

Similarly,

$$\begin{aligned} \frac{\partial L}{\partial b_i^{(M)}} &= \frac{\partial L}{\partial y_i^{(M)}} \cdot \frac{\partial y_i^{(M)}}{\partial x_i^{(M)}} \cdot \frac{\partial x_i^{(M)}}{\partial b_i^{(M)}} \\ &= \frac{\partial L}{\partial y_i^{(M)}} \cdot \sigma'(x_i^{(M)}) \end{aligned}$$

<u>**For** $1 \le m < M$</u>, in this situation $\frac{\partial L}{\partial y_i^{(m)}}$ is not easy to compute. Note that $x^{(m+1)} = W^{(m+1)} y^{(m)} + b^{(m+1)}$.

$$\begin{aligned}
\frac{\partial L}{\partial y_i^{(m)}} &= \sum_k \frac{\partial L}{\partial x_k^{(m+1)}} \cdot \frac{\partial x_k^{(m+1)}}{\partial y_i^{(m)}} \\
&= \sum_k \frac{\partial L}{\partial y_k^{(m+1)}} \cdot \frac{\partial y_k^{(m+1)}}{\partial x_k^{(m+1)}} \cdot \frac{\partial x_k^{(m+1)}}{\partial y_i^{(m)}} \\
&= \sum_k \frac{\partial L}{\partial y_k^{(m+1)}} \cdot \sigma'(x_k^{(m+1)}) \cdot w_{ki}^{(m+1)}
\end{aligned}$$

Then use this form to compute,

$$\begin{aligned}
\frac{\partial L}{\partial w_{ij}^{(m)}} &= \frac{\partial L}{\partial y_i^{(m)}} \cdot \frac{\partial y_i^{(m)}}{\partial x_i^{(m)}} \cdot \frac{\partial x_i^{(m)}}{\partial w_{ij}^{(m)}} \\
&= \frac{\partial L}{\partial y_i^{(m)}} \cdot \sigma'(x_i^{(m)}) \cdot y_j^{(m-1)}
\end{aligned}$$

Similarly,

$$\begin{aligned}
\frac{\partial L}{\partial b_i^{(m)}} &= \frac{\partial L}{\partial y_i^{(m)}} \cdot \frac{\partial y_i^{(m)}}{\partial x_i^{(m)}} \cdot \frac{\partial x_i^{(m)}}{\partial b_i^{(m)}} \\
&= \frac{\partial L}{\partial y_i^{(m)}} \cdot \sigma'(x_i^{(m)})
\end{aligned}$$

Summary

1. Compute $\frac{\partial L}{\partial y_i^{(M)}}$.

2. Use
$$\frac{\partial L}{\partial y_i^{(m)}} = \sum_k \frac{\partial L}{\partial y_k^{(m+1)}} \cdot \sigma'(x_k^{(m+1)}) \cdot w_{ki}^{(m+1)}$$
compute $\frac{\partial L}{\partial y_i^{(m)}}$ for $m = 1, 2..., M-1$.

3. Compute
$$\frac{\partial L}{\partial w_{ij}^{(m)}} = \frac{\partial L}{\partial y_i^{(m)}} \cdot \sigma'(x_i^{(m)}) \cdot y_j^{(m-1)}$$
for $m = 1, 2..., M$.

4. Compute
$$\frac{\partial L}{\partial b_i^{(m)}} = \frac{\partial L}{\partial y_i^{(m)}} \cdot \sigma'(x_i^{(m)})$$
for $m = 1, 2..., M$.

6.4 Other Methods

Stochastic Gradient Descent (SGD)

Subgradient Method

7 Constrained Optimization and Gradient Projection

7.1 Constrained Optimization: Basic

7.1.1 Def: Optimality

$$\min_{x \in \&} f(x)$$

where & is a non-empty closed and convex subset of $\mathbb{R}^n$.
Assume f is continuously differentaible on &.

Definition 9. *x^* is a local min of f over & if $\exists \varepsilon > 0$ s.t. $f(x^*) \leq f(x) \quad \forall x \in \&$ with $\|x - x^*\| < \varepsilon$. x^* is global min of f over & if $f(x^*) \leq f(x) \quad \forall x \in \&$.*

7.1.2 Prop: local-min$\Rightarrow \nabla f(x^*)^T(x - x^*) \geq 0, \forall x \in \& \Leftrightarrow$ global-min in convex

Proposition 4 (optimality conditions)**.**

(a) (Necessary Conditions for local-min) If x^* is a local min of f over &, then

$$\nabla f(x^*)^T(x - x^*) \geq 0 \quad \forall x \in \&$$

(b) (Sufficient and Necessary Condition for global-min of convex f) If f is convex over &, then above condition is also sufficient for x^* to be a global-min.

证明.

(a) Suppose x^* is a local-min, and $\nabla f(x^*)^T(x - x^*) < 0$ for some $x \in \&$.

Let $g(\varepsilon) = f(x^* + \varepsilon(x - x^*))$, then $g'(\varepsilon) = \nabla f(x^* + \varepsilon(x - x^*))^T(x - x^*)$.

By MVT(middle value theorem), $g(\varepsilon) = g(0) + \varepsilon g'(\beta\varepsilon)$ for some $\beta \in [0, 1]$

$$\Rightarrow f(x^* + \varepsilon(x - x^*)) = f(x^*) + \varepsilon \nabla f(x^* + \beta\varepsilon(x - x^*))^T(x - x^*) \quad \text{for some } \beta \in [0, 1]$$

Since ∇f is continuous, we have that for all sufficient small $\varepsilon > 0$, $\nabla f(x^* + \beta\varepsilon(x - x^*))^T(x - x^*) < 0 \Rightarrow f(x^* + \varepsilon(x - x^*)) = f(x^*)$

Since $x^* + \varepsilon(x - x^*) = \varepsilon x + (1 - \varepsilon)x^* \in \&$, then x^* can't be a local-min over & $\rightarrow$ contradiction.

(b) Convexity of f over & $\Rightarrow f(x) \geq f(x^*) + \nabla f(x^*)^T(x - x^*), \quad \forall x \in \&$.

Thus,

$$\begin{aligned} &\nabla f(x^*)^T(x - x^*), \quad \forall x \in \& \\ \Rightarrow &f(x) \geq f(x^*) \quad \forall x \in \& \\ \Rightarrow &x^* \text{ is a global min of } f \text{ over } \& \end{aligned}$$

□

7.1.3 Def: Interior Point

Definition 10. *y is an interior point of & if $\exists \varepsilon > 0$ s.t.*

$$B_\varepsilon = \{x : \|y - x\| < \varepsilon\} \subset \&$$

<u>Remark:</u> If x^* is an interior point of &, then

$$\text{"}x^* \text{ is local min"} \Rightarrow \text{"}\nabla f(x^*) = 0\text{"}$$

$$\text{If } f \text{ is convex, "}x^* \text{ is global min"} \Leftrightarrow \text{"}\nabla f(x^*) = 0\text{"}$$

7.2 Constrained Optimization Example

$$\begin{array}{c} \max\limits_{x \in \&} \quad x_1^{a_1} x_2^{a_2} \cdots x_n^{a_n} \\ \& = \{x : \sum_{i=1}^n x_i = 1, x_i \geq 0, i = 1, 2, ..., n\} \\ a_i, i = 1, 2, ..., n \text{ are given positive scalars} \end{array}.$$

equivalent to

$$\min_{x \in \&} \quad f(x)$$

with $f(x) = -\sum a_i \ln x_i$

$$\nabla f(x) = \left(-\frac{a_1}{x_1}, -\frac{a_2}{x_2}, ..., -\frac{a_n}{x_n}\right)$$

$$\nabla^2 f(x) = diag\left(\frac{a_1}{x_1^2}, \frac{a_2}{x_2^2}, ..., \frac{a_n}{x_n^2}\right) \succ 0$$

$$\Rightarrow f \text{ is strictly convex.}$$

$$\begin{aligned} x^* \in \& \text{ is (unique) min} &\Leftrightarrow \nabla f(x^*)^T (x - x^*) \geq 0 \quad \forall x \in \&. \\ &\Leftrightarrow -\sum_{i=1}^n \frac{a_i}{x_i^*}(x - x^*) \geq 0 \quad \forall x \in \&. \\ &\Leftrightarrow -\sum_{i=1}^n a_i \frac{x_i}{x_i^*} + \sum_{i=1}^n a_i \geq 0 \quad \forall x \in \&. \end{aligned}$$

Guess: $x_i^* = \frac{a_i}{\sum_{i=1}^n a_i}$. Then,

$$-\sum_{i=1}^n a_i \frac{x_i}{x_i^*} + \sum_{i=1}^n a_i = 0, \quad \forall x \in \&$$

Thus $x^* = \frac{a_i}{\sum_{i=1}^n a_i}$ is unique min.

7.3 Projection onto Closed Convex Set

7.3.1 Def: Projection $[z]^\&$

Definition 11. *Let & be a <u>closed convex</u> subset of $\mathbb{R}^n$. Then, for $z \in \mathbb{R}^n$, the <u>projection</u> of z on & is denoted by $[z]^\&$ and is given by*

$$[z]^\& = \arg\min_{y \in \&} \|z - y\|^2$$

i.e. Find the min distance from & to z

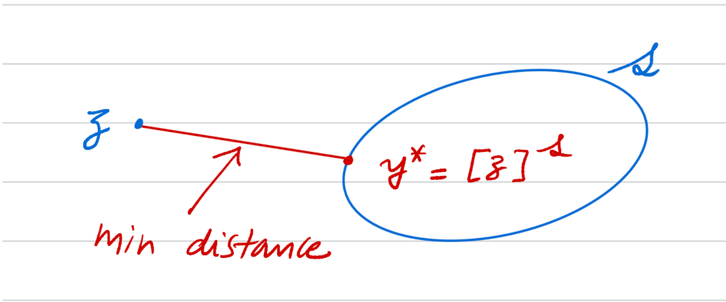


图 5: Projection onto Closed Convex Set

Note: $[z]^{\&}$ exists and is unique in convex &, however, when & is not convex, $[z]^{\&}$ may not be unique.

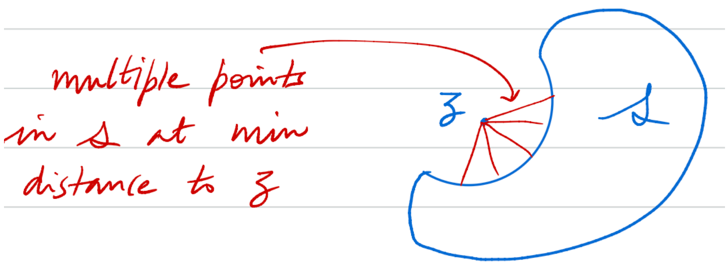


图 6: Projection onto Closed non-Convex Set

7.3.2 Prop: unique projection $[z]^{\&}$ on closed convex subset of $\mathbb{R}^n$

Proposition 5 (Existence and Uniqueness of Projection)**.** *Let & be a closed convex subset of* $\mathbb{R}^n$*. Then, for every* $z \in \mathbb{R}^n$*, there exists a unique* $[z]^{\&}$*.*

证明. Nee to show that $\min_{y\in\&} \|z-y\|^2$ exists and is unique.
Let x be some element of &. Then

$$\begin{aligned} &\text{minimizing } \|z-y\|^2 \text{ over all } y \in \& \\ \equiv &\text{minimizing } \|z-y\|^2 \text{ over the set } A = \{y \in \& : \|z-y\|^2\} \end{aligned}$$

$g(y) = \|z-y\|^2$ is strictly convex on set & $\Rightarrow$ A is a convex set and g is convex on A.
Also g is continuous $\Rightarrow$ A is closed.
Finally, $y \in A \Rightarrow \|y\|^2 = \|y-z+z\|^2 \leq \|y-z\|^2 + \|z\|^2 \leq \|z-x\|^2 + \|z\|^2 \Rightarrow A$ is bounded.
Thus, $g(y) = \|z-y\|^2$ is strictly convex over set A, which is compact.
Therefore, $\min_{y\in\&} \|\& - y\|^2 = \min_{y\in A} \|\& - y\|^2$ exists (Weierstrass'Theorem) and is unique (strict convexity). $\square$

7.3.3 Prop: $[z]^{\&}$ is projection on closed conex subset of $\mathbb{R}^n \Leftrightarrow (z-[z]^{\&})^T(y-[z]^{\&}) \leq 0, \forall y \in \&$

Proposition 6 (Necessary and Sufficient Condition for Projection)**.** *Let & be a closed conex subset of $\mathbb{R}^n$. Then,*

$$
\begin{aligned}
[z]^{\&} = y^* &\Leftrightarrow (y^* - z)^T(y - y^*) \geq 0, \quad \forall y \in \&. \\
&\Leftrightarrow (z - y^*)^T(y - y^*) \leq 0, \quad \forall y \in \&.
\end{aligned}
$$

证明. $[z]^{\&} = \operatorname{argmin}_{y\in\&} g(y)$, with $g(y) = \|z - y\|^2$ (which is strictly convex), $\nabla g(y) = 2(y - z)$. By the optimality conditions,

$$
\begin{aligned}
&y^* \text{ is the unique minimizer of } g(y) \text{ over } \& \\
\Leftrightarrow&\nabla g(y^*)^T(y - y^*) \geq 0 \quad \forall y \in \& \\
\Leftrightarrow&(y^* - z)^T(y - y^*) \geq 0, \quad \forall y \in \&. \\
\Leftrightarrow&(z - y^*)^T(y - y^*) \leq 0, \quad \forall y \in \&.
\end{aligned}
$$

□

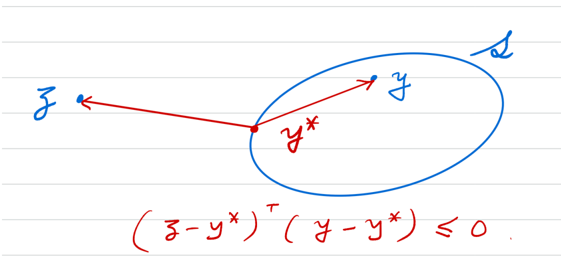


图 7: Necessary and Sufficient Condition for Projection

7.3.4 Prop: Projection is non-expansive $\|[x]^{\&} - [z]^{\&}\| \leq \|x - z\|, \forall x, z \in \mathbb{R}^n$

Proposition 7 (Projection is non-expansive)**.** *Let & be a closed convex subset of $\mathbb{R}^n$. Then for $x, z \in \mathbb{R}^n$*

$$\|[x]^{\&} - [z]^{\&}\| \leq \|x - z\| \quad \forall x, z \in \mathbb{R}^n$$

证明. From previous theorem, we know

$$
\begin{aligned}
(1).\quad ([x]^{\&} - x)^T(y - [x]^{\&}) \geq 0, \quad \forall y \in \&. \\
(2).\quad ([z]^{\&} - z)^T(y - [z]^{\&}) \geq 0, \quad \forall y \in \&.
\end{aligned}
$$

set $y = [z]^{\&}$ in (1) and $y = [x]^{\&}$ in (2), and adding,

$$
\begin{aligned}
&([z]^{\&} - [x]^{\&})^T([x]^{\&} - x + z - [z]^{\&}) \geq 0 \\
\Rightarrow&([z]^{\&} - [x]^{\&})^T(z - x) \geq \|[z]^{\&} - [x]^{\&}\|^2
\end{aligned}
$$

Applying Cauchy-schwary inequality,

$$\|[z]^{\&} - [x]^{\&}\|^2 \leq \|[z]^{\&} - [x]^{\&}\|\|z - x\|$$
$$\|[z]^{\&} - [x]^{\&}\| \leq \|z - x\|$$

□

7.4 Projection on (Linear) Subspaces of $\mathbb{R}^n$

7.4.1 Orthogonality Principle in subspaces of $\mathbb{R}^n$: $(z - y^*)^T x = 0, \forall x \in \&$

Suppose & is a linear subspace of $\mathbb{R}^n$, any linear combination of points in & is also in &. Note that & is closed and convex.

Then, for $z \in \mathbb{R}^n$, $[z]^{\&} = y^*$ satisfies:

$$(z - y^*)^T (y - y^*) \leq 0, \quad \forall y \in \&.$$

According to the property of subsapce, we can infer that

$$(z - y^*)^T x \leq 0, \quad \forall x \in \&.$$

$-x$ also in &, $-x \in \& \Rightarrow$

$$(z - y^*)^T x \geq 0, \quad \forall x \in \&.$$

Then we can infer that

$$(z - y^*)^T x = 0, \quad \forall x \in \&.$$

which is called orthogonality principle.

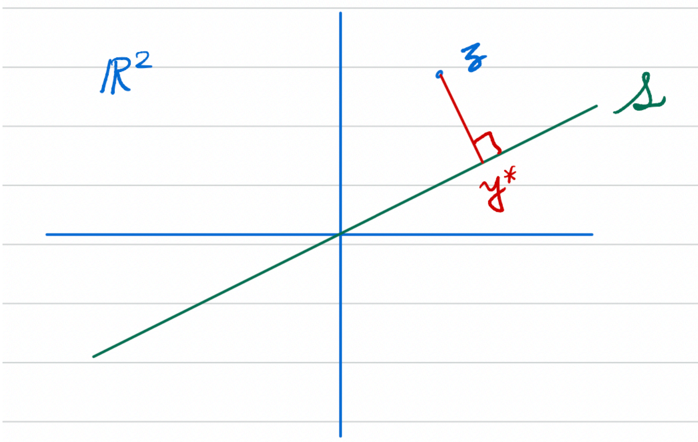


图 8: Point from $\mathbb{R}^2$ to $\mathbb{R}$

7.5 Gradient Projection Method

$\min_{x \in \&} f(x)$, & is convex and closed.

$$x_{k+1} = [x_k + \alpha_k d_k]^{\&}$$

Special Case: Fixed step-size, steepest descent

$$x_{k+1} = [x_k - \alpha \nabla f(x_k)]^{\&} \tag{1}$$

7.5.1 Def: fixed point in fixed step-size steepest descent method, $\tilde{x} = [\tilde{x} - \alpha \nabla f(\tilde{x})]^{\&}$

Definition 12. *$\tilde{x}$ is a fixed (stationary) point of iteration in (1) if*

$$\tilde{x} = [\tilde{x} - \alpha \nabla f(\tilde{x})]^{\&}$$

7.5.2 Prop: $L-$Lipschitz gradient and $0 < \alpha < \frac{2}{L}$ $\Rightarrow$ limit point is a fixed point (in fixed step-size steepest descent method)

Proposition 8. *If f has $L-$Lipschitz gradient and $0 < \alpha < \frac{2}{L}$, every limit point of (1) is a fixed point of (1).*

证明. By the Descent Lemma,

$$f(x_{k+1}) \leq f(x_k) + \nabla f(x_k)^T (x_{k+1} - x_k) + \frac{L}{2} \|x_{k+1} - x_k\|^2 \tag{2}$$

By the necessary and sufficient condition for projection,

$$(x_k - \alpha \nabla f(x_k) - x_{k+1})^T (x - x_{k+1}) \leq 0, \quad \forall x \in \&$$

Set $x = x_k$ above

$$\Rightarrow \alpha \nabla f(x_k)^T (x_{k+1} - x_k) \leq -\|x_k - x_{k+1}\|^2 \tag{3}$$

According to (2) and (3),

$$f(x_{k+1}) - f(x_k) \leq (\frac{L}{2} - \frac{1}{\alpha}) \|x_k - x_{k+1}\|^2$$

where $\frac{L}{2} - \frac{1}{\alpha} < 0$
If $\{x_k\}$ has limit point $\bar{x}$, $LHS \overset{k\to\infty}{\longrightarrow} 0$

$$\|x_{k+1} - x_k\| \overset{k\to\infty}{\longrightarrow} 0 \Rightarrow [\bar{x} - \alpha \nabla f(\bar{x})]^{\&} = \bar{x}$$

□

7.5.3 Prop: x is minimizer in convex func $\Leftrightarrow$ fixed point (in fixed step-size steepest descent method)

Proposition 9. *If f is convex, then x^* is a minimizer of f over & $\Leftrightarrow$ $x^* = [x^* - \alpha \nabla f(x^*)]^{\&}$ (i.e., x^* is a fixed point of (1))*