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

(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* <u>*minimum*</u>*. 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 $g = \alpha x + (1-\alpha)y$

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

□

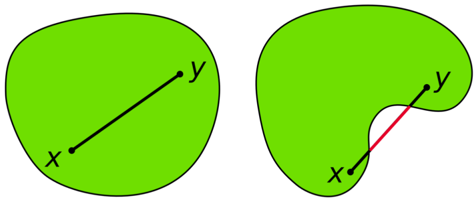
Property (2nd order): If f is twice differentiable, then f is convex iff

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

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

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

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

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

2.4.2 Lemma: Strongly convexity $\Rightarrow$ Strictly convexity

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

证明.

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

□

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

2.4.3 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,

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

□

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 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.

3.4 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 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.*

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

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

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

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

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* $A - B \preceq 0$*)*

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

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

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

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

□

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

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

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

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 $g'(t) = \nabla f(x + t(y - x))^T (y - x)$

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

□

4.2 Convergence of Steepest Descent with Fixed Stepsize

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

Theorem 8. *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,

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

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, $\nabla f(\bar{x}) = 0$ $\square$

Example 8. *$f(x) = \frac{1}{2}x^2, x \in \mathbb{R}$, $\nabla f(x) = x$, 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 9. *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 $c = 0$

$\square$

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 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 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,

$$\begin{aligned} f(x^*) &\geq f(x_k) + \nabla f(x_k)^T (x^* - x_k) \\ \Rightarrow \nabla f(x_k)^T (x^* - x_k) &\leq f(x^*) - f(x_k)\end{aligned}$$

Thus,

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

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 10. *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

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

Then,

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

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. $\alpha = \frac{1}{M} \Rightarrow 1 + \alpha^2 M^2 - 2m\alpha = 2(1 - \frac{m}{M}) \geq 1$ if $\frac{m}{M} \leq 0.5$
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 $\alpha = \alpha^*$,

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

$\frac{M}{m}$ 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 10. $f(x) = \frac{1}{2}x^T Qx + b^T x + c, \quad Q \succ 0,\ \nabla^2 f(x) = Q.$

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

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

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

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

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

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

Theorem 11. *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

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

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 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*

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

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

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

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

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

Expand the inequality,

$$\begin{bmatrix} \xi_k \\ u_k \end{bmatrix}^\top \begin{bmatrix} I & -\alpha I \\ -\alpha I & \alpha^2 I \end{bmatrix} \begin{bmatrix} \xi_k \\ u_k \end{bmatrix} + \begin{bmatrix} \xi_k \\ u_k \end{bmatrix}^\top \begin{bmatrix} -\rho^2 I & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \xi_k \\ u_k \end{bmatrix} + \lambda \begin{bmatrix} \xi_k \\ u_k \end{bmatrix}^\top M \begin{bmatrix} \xi_k \\ u_k \end{bmatrix} \leq 0$$

Applying the key relation

$$\|\xi_{k+1}\|^2 - \rho^2\|\xi_k\|^2 + \lambda \begin{bmatrix} \xi_k \\ u_k \end{bmatrix}^\top M \begin{bmatrix} \xi_k \\ u_k \end{bmatrix} \leq 0$$

□