Notes for Optimization Algorithms
Spring 2023

Xiangxiong Zhang
Department of Mathematics, Purdue University

Contents

Preface

These notes are supposed to be self-content. The main focus is currently the classical analysis of popular algorithms for large-scale optimization. Typos are inevitable. Use with caution.

Notation

Unless specified otherwise:

1. x denotes a single variable, and $\mathbf{x}$ denotes a column vector.

2. $\mathbf{x}^T$ is the transpose of $\mathbf{x}$, thus a row vector.

3. $f(\mathbf{x})$ is a scalar-valued multi-variable function.

4. $\nabla f(\mathbf{x})$ is a column vector.

5. For a matrix $A \in \mathbb{R}^{n\times n}$, $\|A\|$ is the spectral norm; $\sigma_i(A)$ and $\lambda_i(A)$ denote its singular values and eigenvalues respectively.

6. $\forall$ means *for any*, and $\exists$ means *there exists*.

7. C^k functions: the partial derivatives up to k-th order exist and are continuous.

8. $\langle \mathbf{a}, b \rangle$ denotes the dot product of two vectors.

1

Taylor's Theorems, Lipschitz continuity and convexity

In this chapter, we first introduce some tools that will be needed for analyzing the simplest gradient descent method.

1.1 Multi-variable Taylor's Theorems

We first start with the well-known mean value theorem in calculus without proof:

Theorem 1.1. *If a function $f(\mathbf{x})$ is continuous on an interval $[a, b]$ and $f'(\mathbf{x})$ exists, then there exists $c \in (a, b)$ s.t.*

$$f(b) - f(a) = f'(c)(b - a).$$

Remark 1.1. *The geometrical meaning of this theorem is simply saying that there is a point c where the tangent line (with slope $f'(c)$) is parallel to the secant line passing two end points at a and b (with slope $\frac{f(b)-f(a)}{b-a}$).*

Theorem 1.2 (Single variable Taylor's Theorem)**.** *Suppose that $I \subset \mathbb{R}$ is an open interval and that $f(\mathbf{x})$ is a function of class C^2 ($f''(\mathbf{x})$ exists and is continuous) on I. For any $a \in I$ and h such that $a + h \in I$, there exists some $\theta \in (0, 1)$ such that*

$$f(a + h) = f(a) + hf'(a) + \frac{h^2}{2} f''(a + \theta h).$$

Proof. Consider

$$g_1(\mathbf{x}) = f(\mathbf{x}) - f(a) - (x - a)f'(a)$$

then $g_1(a) = g_1'(a) = 0$. Define

$$g(\mathbf{x}) = g_1(\mathbf{x}) - \left(\frac{x - a}{h}\right)^2 g_1(a + h),$$

then $g(a) = g'(a) = g(a+h) = 0$. By Mean Value Theorem on $g(\mathbf{x})$, we have

$$g(a) = g(a+h) = 0 \Longrightarrow g'(a + \alpha h) = 0, \quad \alpha \in (0,1).$$

Use Mean Value Theorem again on $g'(\mathbf{x})$:

$$g'(a) = g'(a+\alpha h) = 0 \Longrightarrow g''(a + \theta h) = 0, \quad \theta \in (0, \alpha).$$

Since $g''(\mathbf{x}) = f''(\mathbf{x}) - \frac{2}{h^2} g_1(a+h)$, $g''(a+\theta h) = 0$ implies that we get the explicit remainder for the second order Taylor expansion as $g_1(a+h) = \frac{h^2}{2} f''(a+\theta h)$. □

Theorem 1.3 (Multivariate First Order Taylor's Theorem)**.** *Suppose that $S \subset \mathbb{R}^n$ is an open set and that $f : S \longrightarrow \mathbb{R}$ is a function of class C^1 on S (first order partial derivatives exist and are continuous). Then for any $\mathbf{a} \in S$ and $\mathbf{h} \in \mathbb{R}^n$ such that the line segment connecting $\mathbf{a}$ and $\mathbf{a}+\mathbf{h}$ is contained in S, there exists $\theta \in (0,1)$ such that*

$$f(\mathbf{a}+\mathbf{h}) = f(\mathbf{a}) + \nabla f(\mathbf{a}+\theta\mathbf{h}) \cdot \mathbf{h}.$$

Proof. Define $g(t) = f(\mathbf{a} + t\mathbf{h})$. By Mean Value Theorem on $g(t)$, there is $\theta \in (0,1)$ s.t.

$$g(1) = g(0) + g'(\theta).$$

By chain rule, we have $g'(\theta) = \nabla f(\mathbf{a}+\theta\mathbf{h}) \cdot \mathbf{h}$, which completes the proof. □

Theorem 1.4 (Multivariate Quadratic Taylor's Theorem)**.** *Suppose that $S \subset \mathbb{R}^n$ is an open set and that $f : S \longrightarrow \mathbb{R}$ is a function of class C^2 on S (second order partial derivatives exist and are continuous). Then for any $\mathbf{a} \in S$ and $\mathbf{h} \in \mathbb{R}^n$ such that the line segment connecting $\mathbf{a}$ and $\mathbf{a}+\mathbf{h}$ is contained in S, there exists $\theta \in (0,1)$ such that*

$$f(\mathbf{a}+\mathbf{h}) = f(\mathbf{a}) + \nabla f(\mathbf{a}) \cdot \mathbf{h} + \frac{1}{2}\mathbf{h}^T \nabla^2 f(\mathbf{a}+\theta\mathbf{h})\mathbf{h}.$$

Proof. Define $g(t) = f(\mathbf{a} + t\mathbf{h})$. By Theorem 1.2 on $g(t)$, there is $\theta \in (0,1)$ s.t.

$$g(1) = g(0) + g'(0) + \frac{1}{2} g''(\theta).$$

By chain rule, we have $g'(0) = \nabla f(\mathbf{a}) \cdot \mathbf{h}$ and $g''(\theta) = h^T \nabla^2 f(\mathbf{a}+\theta\mathbf{h})\mathbf{h}$, which completes the proof. □

We need to be careful that these Taylor's Theorems may not hold for a vector-valued function. For instance, consider a smooth scalar-valued function

$$f : \mathbb{R}^n \longrightarrow \mathbb{R},$$

its gradient is a vector-valued function

$$\nabla f : \mathbb{R}^n \longrightarrow \mathbb{R}^n.$$

One might presume a formula like $\nabla f(\mathbf{a}+\mathbf{h}) = \nabla f(\mathbf{a}) + \nabla^2 f(\mathbf{a}+\theta\mathbf{h})\mathbf{h}$, which could be wrong!

1.2 Convex functions

1.2.1 Definition

Definition 1.1. *Consider a function $f : \mathbb{R}^n \to \mathbb{R}$ and any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ and any $\lambda \in (0, 1)$.*

1. *$f(\mathbf{x})$ is called convex if $f(\lambda \mathbf{x} + (1-\lambda)\mathbf{y}) \leq \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y})$.*
2. *$f(\mathbf{x})$ is called strictly convex if $f(\lambda \mathbf{x} + (1-\lambda)\mathbf{y}) < \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y})$.*
3. *$f(\mathbf{x})$ is called strongly convex with a constant parameter $\mu > 0$ if*
$$f(\lambda \mathbf{x} + (1-\lambda)\mathbf{y}) \leq \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y}) - \frac{\mu}{2}\lambda(1-\lambda)\|\mathbf{x} - \mathbf{y}\|^2.$$
4. *$f(\mathbf{x})$ is (strictly or strongly) concave if $-f(\mathbf{x})$ is (strictly or strongly) convex.*
5. *East to verify that $f(\mathbf{x})$ is strongly convex with $\mu > 0$ if and only if $f(\mathbf{x}) - \frac{\mu}{2}\|x\|^2$ is convex. Strong convexity with $\mu = 0$ is convexity.*
6. *It is easy to see that*

   strong convexity $\Rightarrow$ strict convexity $\Rightarrow$ convexity.

A convex function does not need to be differentiable, e.g., the single variable absolute value function $f(x) = |x|$ is convex.

Example 1.1. *Any norm of a matrix $X \in \mathbb{R}^{n \times n}$ is convex due to the triangle inequality of norms:*

$$\|\lambda X + (1-\lambda)Y\| \leq \|\lambda X\| + \|(1-\lambda)Y\| = \lambda\|X\| + (1-\lambda)\|Y\|.$$

See Appendix A.5 for examples of matrix norms.

It is straightforward to verify the following from the definition:

Theorem 1.5. *Let $f(\mathbf{x})$ and $g(\mathbf{x})$ be two convex functions. Then*

1. *$f(\mathbf{x}) + g(\mathbf{x})$ is convex;*
2. *If $g(\mathbf{x})$ is strictly convex, so is $f(\mathbf{x}) + g(\mathbf{x})$;*
3. *If $g(\mathbf{x})$ is strongly convex, so is $f(\mathbf{x}) + g(\mathbf{x})$.*

If a single variable function is continuously differentiable, then being convex simply means that the derivative $f'(x)$ is increasing, i.e., $[f'(y) - f'(x)](y - x) \geq 0$. If twice continuously differentiable, then convexity simply means $f''(x) \geq 0$, and strong convexity means $f''(x) \geq \mu > 0$. The following subsections provide justifications.

1.2.2 Equivalent conditions

Geometrically convexity also means that function graph is always above any tangent line: $f(x) \geq f(y) + f'(y)(x-y)$.

Lemma 1.1. *Assume $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable. Then the following are equivalent definitions of $f(\mathbf{x})$ being convex:*

1. $f(\mathbf{x}) \geq f(\mathbf{y}) + \langle \nabla f(\mathbf{y}), \mathbf{x}-\mathbf{y} \rangle, \quad \forall \mathbf{x}, \mathbf{y}.$
2. $\langle \nabla f(\mathbf{y}) - \nabla f(\mathbf{x}), \mathbf{y}-\mathbf{x} \rangle \geq 0, \quad \forall \mathbf{x}, \mathbf{y}.$

If replacing $\geq$ with $>$ above, then we get equivalent definitions for strict convexity. For strong convexity with parameter $\mu > 0$, the following are equivalent definitions:

1. $f(\mathbf{x}) \geq f(\mathbf{y}) + \langle \nabla f(\mathbf{y}), \mathbf{x}-\mathbf{y} \rangle + \frac{\mu}{2}\|\mathbf{x}-\mathbf{y}\|^2, \quad \forall \mathbf{x}, \mathbf{y}.$
2. $\langle \nabla f(\mathbf{y}) - \nabla f(\mathbf{x}), \mathbf{y}-\mathbf{x} \rangle \geq \mu\|\mathbf{x}-\mathbf{y}\|^2, \quad \forall \mathbf{x}, \mathbf{y}.$

Proof. We only prove the equivalency for strong convexity, since convexity is simply strong convexity with $\mu = 0$ and discussion for strict convexity is similar to convexity.

First, assume $f(\mathbf{x})$ is strongly convex, then

$$f(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}) \leq \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y}) - \frac{\mu}{2}\lambda(1-\lambda)\|\mathbf{x}-\mathbf{y}\|^2$$

$$\Rightarrow \frac{f(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}) - f(\mathbf{y})}{\lambda} \leq f(\mathbf{x}) - f(\mathbf{y}) - \frac{\mu}{2}(1-\lambda)\|\mathbf{x}-\mathbf{y}\|^2.$$

Let $g(t) = f(t\mathbf{x} + (1-t)\mathbf{y})$ then $g(0) = f(\mathbf{y})$ and

$$g'(t) = \nabla f(t\mathbf{x} + (1-t)\mathbf{y})^T(\mathbf{x}-\mathbf{y}) = \langle \nabla f(t\mathbf{x} + (1-t)\mathbf{y}), \mathbf{x}-\mathbf{y} \rangle.$$

By the Mean Value Theorem on $g(t)$, there exists $s \in (0, t)$ such that $g'(s) = \frac{g(t)-g(0)}{t}$, thus

$$\frac{f(t\mathbf{x} + (1-t)\mathbf{y}) - f(\mathbf{y})}{t} = \frac{g(t)-g(0)}{t} = g'(s) = \langle \nabla f(s\mathbf{x} + (1-s)\mathbf{y}), \mathbf{x}-\mathbf{y} \rangle,$$

and

$$\langle \nabla f(s\mathbf{x} + (1-s)\mathbf{y}), \mathbf{x}-\mathbf{y} \rangle \leq f(\mathbf{x}) - f(\mathbf{y}) - \frac{\mu}{2}(1-t)\|\mathbf{x}-\mathbf{y}\|^2.$$

Let $t \to 0$ then $s \to 0$, we get $f(\mathbf{x}) \geq f(\mathbf{y}) + \langle \nabla f(\mathbf{y}), \mathbf{x}-\mathbf{y} \rangle + \frac{\mu}{2}\|\mathbf{x}-\mathbf{y}\|^2$.

Second, assume

$$f(\mathbf{x}) \geq f(\mathbf{y}) + \langle \nabla f(\mathbf{y}), \mathbf{x}-\mathbf{y} \rangle + \frac{\mu}{2}\|\mathbf{x}-\mathbf{y}\|^2.$$

Then combining with

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{\mu}{2}\|\mathbf{x} - \mathbf{y}\|^2,$$

we get $\langle \nabla f(\mathbf{y}) - \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \geq \mu\|\mathbf{x} - \mathbf{y}\|^2$.

Third, assume $\langle \nabla f(\mathbf{y}) - \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \geq \mu\|\mathbf{x} - \mathbf{y}\|^2$. Let $\mathbf{x}_t = t\mathbf{x} + (1-t)\mathbf{y}$, then

$$\langle \nabla f(\mathbf{x}_t) - \nabla f(\mathbf{y}), \mathbf{x}_t - \mathbf{y} \rangle \geq \mu\|\mathbf{x}_t - \mathbf{y}\|^2,$$

thus

$$\langle \nabla f(t\mathbf{x} + (1-t)\mathbf{y}) - \nabla f(\mathbf{y}), t(\mathbf{x} - \mathbf{y}) \rangle \geq \mu t^2\|\mathbf{x} - \mathbf{y}\|^2,$$

and

$$\langle \nabla f(t\mathbf{x} + (1-t)\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \geq \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle + \mu t\|\mathbf{x} - \mathbf{y}\|^2.$$

Consider $g(t) = f(t\mathbf{x} + (1-t)\mathbf{y})$, then

$$\int_0^1 g'(t)dt = \int_0^1 \langle \nabla f(t\mathbf{x} + (1-t)\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle dt \geq \int_0^1 (\langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle + \mu t\|\mathbf{x} - \mathbf{y}\|^2)dt$$

$$= \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle + \frac{\mu}{2}\|\mathbf{x} - \mathbf{y}\|^2.$$

So

$$f(\mathbf{x}) - f(\mathbf{y}) = g(1) - g(0) \geq \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle + \frac{\mu}{2}\|\mathbf{x} - \mathbf{y}\|^2.$$

Finally, assume

$$f(\mathbf{x}) \geq f(\mathbf{y}) + \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle + \frac{\mu}{2}\|\mathbf{x} - \mathbf{y}\|^2, \quad \forall \mathbf{x}, \mathbf{y}.$$

Let $\mathbf{x}_t = t\mathbf{x} + (1-t)\mathbf{y}$, then we have

$$f(\mathbf{x}) \geq f(\mathbf{x}_t) + \langle \nabla f(\mathbf{x}_t), \mathbf{x} - \mathbf{x}_t \rangle + \frac{\mu}{2}\|\mathbf{x} - \mathbf{x}_t\|^2,$$

$$f(\mathbf{y}) \geq f(\mathbf{x}_t) + \langle \nabla f(\mathbf{x}_t), \mathbf{y} - \mathbf{x}_t \rangle + \frac{\mu}{2}\|\mathbf{y} - \mathbf{x}_t\|^2,$$

Combining the two inequalities with coefficients t and $1-t$, notice that $\mathbf{x} - \mathbf{x}_t = (1-t)(\mathbf{x} - \mathbf{y})$ and $\mathbf{y} - \mathbf{x}_t = (-t)(\mathbf{x} - \mathbf{y})$,

$$f(t\mathbf{x} + (1-t)\mathbf{y}) \leq tf(\mathbf{x}) + (1-t)f(\mathbf{y}) - \frac{\mu}{2}t(1-t)\|\mathbf{x} - \mathbf{y}\|^2.$$

□

Lemma 1.2. *Assume $f : \mathbb{R}^n \to \mathbb{R}$ is twice continuously differentiable (second order partial derivatives exist and are continuous).*

1. *$f(\mathbf{x})$ is convex if and only if $\nabla^2 f(\mathbf{x}) \geq 0$ (Hessian matrix is positive semi-definite) for all $\mathbf{x}$.*
2. *$f(\mathbf{x})$ is strongly convex if and only if $\nabla^2 f(\mathbf{x}) \geq \mu I$ for all $\mathbf{x}$.*
3. *$f(\mathbf{x})$ is strictly convex if $\nabla^2 f(\mathbf{x}) > 0$ for all $\mathbf{x}$. This is not necessary even for single variable functions: $f(x) = x^4$ is strictly convex but $f''(x) > 0$ is not true at $x = 0$.*

Proof. First, we shown assumptions on the Hessian are sufficient for convexity, strict convexity and strong convexity. Apply Multivariate Quadratic Taylor's Theorem (Theorem 1.4), we get

$$f(\mathbf{x}) = f(\mathbf{y}) + \nabla f(\mathbf{y})^T(\mathbf{x}-\mathbf{y}) + \frac{1}{2}(\mathbf{x}-\mathbf{y})^T\nabla^2 f[\mathbf{y}+\theta(\mathbf{x}-\mathbf{y})](\mathbf{x}-\mathbf{y}), \theta \in (0,1).$$

Strong convexity is proven by Lemma 1.1 and the fact that

$$\nabla^2 f \geq \mu I \Rightarrow \frac{1}{2}(\mathbf{x}-\mathbf{y})^T\nabla^2 f[\mathbf{y}+\theta(\mathbf{x}-\mathbf{y})](\mathbf{x}-\mathbf{y}) \geq \frac{\mu}{2}\|\mathbf{x}-\mathbf{y}\|^2.$$

Convexity and strict convexity are similarly proven.

Second, assume $f(\mathbf{x})$ is strongly convex. By Lemma 1.1, we have

$$\forall t > 0, \forall \mathbf{p}, \mathbf{x} \in \mathbb{R}^n, f(\mathbf{x}+t\mathbf{p}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), t\mathbf{p}\rangle + \frac{\mu}{2}\|t\mathbf{p}\|^2.$$

With the Quadratic Taylor's Theorem we get

$$\exists \theta \in (0,t), f(\mathbf{x}+t\mathbf{p}) = f(\mathbf{x}) + t\nabla f(\mathbf{x})^T\mathbf{p} + \frac{1}{2}t^2\mathbf{p}^T\nabla^2 f[\mathbf{x}+\theta\mathbf{p}]\mathbf{p}$$

thus

$$\frac{1}{2}t^2\mathbf{p}^T\nabla^2 f[\mathbf{x}+\theta\mathbf{p}]\mathbf{p} \geq \frac{\mu}{2}\|t\mathbf{p}\|^2 \Rightarrow \frac{\mathbf{p}^T\nabla^2 f[\mathbf{x}+\theta\mathbf{p}]}{\|\mathbf{p}\|^2} \geq \mu.$$

Let $t \to 0$, then $\theta \to 0$, we get

$$\frac{\mathbf{p}^T\nabla^2 f[\mathbf{x}]\mathbf{p}}{\|\mathbf{p}\|^2} \geq \mu, \quad \forall \mathbf{p} \in \mathbb{R}^n, \mathbf{p} \neq \mathbf{0}.$$

By the Courant-Fischer-Weyl min- max principle in Appendix A.1, we get $\nabla^2 f[\mathbf{x}] \geq \mu I$. Repeat the same argument for $\mu = 0$, we prove the Hessian condition is sufficient for the convexity. □

Problem 1.1. *In gas dynamics, governing hydrodynamics equations are defined by conservation of mass ρ, momentum $\mathbf{m} = (m_x, m_y, m_z)$ and total energy E. The pressure is defined as $p = (\gamma - 1)(E - \frac{1}{2}\frac{\|\mathbf{m}\|^2}{\rho})$ in equation of state for for ideal gas where $\gamma > 1$ is a constant parameter, e.g., $\gamma = 1.4$ for air. Regard p as a function of conservative variables ρ, m_x, m_y, m_z, E, verify* *that $p(\rho, \mathbf{m}, E)$ is a concave function for $\rho > 0$ thus satisfies the Jensen's inequity:*

$$p\left(a_1\begin{bmatrix}\rho\\ \mathbf{m}\\ E\end{bmatrix}+a_2\begin{bmatrix}\rho\\ \mathbf{m}\\ E\end{bmatrix}\right)\leq a_1 p\left(\begin{bmatrix}\rho\\ \mathbf{m}\\ E\end{bmatrix}\right)+a_2 p\left(\begin{bmatrix}\rho\\ \mathbf{m}\\ E\end{bmatrix}\right),\quad a_1,a_2>0, a_1+a_2=1.$$

Hint: *show the Hessian matrix is negative definite. Start with an easier problem by considering 1D case: $p = (\gamma-1)(E - \frac{1}{2}\frac{m^2}{\rho})$ where m is scalar.*

1.2.3 Jensen's inequality

A convex function by definition satisfies the **Jensen's inequality**:

$$\forall \mathbf{x}, \mathbf{y}, \quad f(a_1\mathbf{x}+a_2\mathbf{y}) \leq a_1 f(\mathbf{x}) + a_2 f(\mathbf{y}), \quad \forall a_1, a_2 \geq 0, a_1+a_2=1.$$

It is straightforward to extend it to n terms by induction, i.e., **Jensen's inequality** also implies

$$\forall \mathbf{x}_i, \quad f\left(\sum_{i=1}^n a_i \mathbf{x}_i\right) \leq \sum_{i=1}^n a_i f(\mathbf{x}_i), \quad \forall a_i \geq 0, \sum_{i=1}^n a_i = 1.$$

Theorem 1.6 (Jensen's inequality in integral form)**.** *If a single variable function $\phi : \mathbb{R} \longrightarrow \mathbb{R}$ is convex, and $\int_a^b g(x)dx$ exists, then*

$$\phi\left(\frac{1}{b-a}\int_a^b g(x)dx\right) \leq \frac{1}{b-a}\int_a^b \phi[g(x)]dx.$$

Proof. First of all, this result can be proven without assuming the differentiability of the convex function. But for convenience, assume $\phi'(x)$ exists, then Lemma 1.1 implies

$$\phi(t) \geq \phi(t_0) + \phi'(t_0)(t-t_0). \tag{1.1}$$

Plug in $t_0 = \frac{1}{b-a}\int_a^b g(x)dx$ and $t = g(x)$ we get

$$\phi[g(x)] \geq \phi\left(\frac{1}{b-a}\int_a^b g(x)dx\right) + \phi'(t_0)\left(g(x) - \frac{1}{b-a}\int_a^b g(x)dx\right).$$

Integrate both sides for variable x, we get

$$\frac{1}{b-a}\int_a^b \phi[g(x)]dx \geq \phi\left(\frac{1}{b-a}\int_a^b g(x)dx\right).$$

□

Remark 1.2. *The proof above can be easily extended to a nondifferentiable convex function which is bounded from below by a linear function, e.g., the proof still holds if assuming there is a slope S_{t_0} for any $t_0 \in \mathbb{R}$ such that*

$$\phi(t) \geq \phi(t_0) + S_{t_0}(t - t_0).$$

For instance, $\phi(t) = |t|$ is not differentiable at $t_0 = 0$, but we have

$$|t| \geq |t_0| + S_{t_0}(t - t_0)$$

with $S_{t_0} = \begin{cases} 1, & t_0 \geq 0 \\ -1, & t_0 < 0 \end{cases}.$

Recall that the spectral norm of a matrix X is a convex function due to the triangle inequality. Next, we prove a Jensen's inequality about the spectral norm.

Lemma 1.3. *Let $\mathbf{g} : \mathbb{R} \to \mathbb{R}^n$ be a single variable vector-valued function, which is integrable on [a, b]. Then*

$$\left\| \int_a^b \mathbf{g}(x) dx \right\| \leq \int_a^b \|\mathbf{g}(x)\| \, dx.$$

Proof. Let $\mathbf{v} = \int_a^b \mathbf{g}(x)dx$, then

$$\|\mathbf{v}\|^2 = \sum_{i=1}^n v_i \int_a^b g_i(x)dx = \int_a^b \left[\sum_{i=1}^n v_i g_i(x) \right] dx = \int_a^b \langle \mathbf{v}, \mathbf{g}(x) \rangle dx.$$

With Cauchy-Schwartz inequality $\langle \mathbf{v}, \mathbf{g}(x) \rangle \leq \|\mathbf{v}\| \|\mathbf{g}(x)\|$, we get

$$\|\mathbf{v}\|^2 \leq \int_a^b \|\mathbf{v}\| \|\mathbf{g}(x)\| dx = \|\mathbf{v}\| \int_a^b \|\mathbf{g}(x)\| dx \Rightarrow \|\mathbf{v}\| \int_a^b \|\mathbf{g}(x)\| dx.$$

□

Theorem 1.7 (Jensen's inequality of the spectral norm). *Let $A(t) : \mathbb{R} \longrightarrow \mathbb{R}^{n \times n}$ be a real symmetric matrix valued function. Assume it is integrable on $[0, 1]$. Then*

$$\left\| \int_0^1 A(t) dt \right\| \leq \int_0^1 \|A(t)\| dt.$$

Remark 1.3. *The integral of a matrix-valued function $A(t)$ is called the Bochner integral (for functions mapping to any Banach space). And the inequality above can be regarded as Jensen's inequality applying to the spectral norm, at least for Hermitian matrices, see [4, 1]*

Proof. For real symmetric matrices, the singular values are the absolute value of eigenvalues. Let $\mathbf{v}$ be the unit eigenvector of the matrix $\int_0^1 A(t)dt$ for the extreme eigenvalue λ such that

$$\int_0^1 A(t)dt\mathbf{v} = \lambda\mathbf{v}, \quad |\lambda| = \left\| \int_0^1 A(t)dt \right\|.$$

Lemma 1.3 and $\|\mathbf{v}\| = 1$ imply

$$\|\lambda\mathbf{v}\| = \left\| \int_0^1 A(t)dt\mathbf{v} \right\| \leq \int_0^1 \|A(t)\mathbf{v}\|dt \leq \int_0^1 \|A(t)\|\|\mathbf{v}\|dt = \int_0^1 \|A(t)\|dt.$$

The left hand side is

$$\|\lambda\mathbf{v}\| = |\lambda|\|\mathbf{v}\| = |\lambda| = \left\| \int_0^1 A(t)dt \right\|.$$

□

1.3 Lipschitz continuous functions

Definition 1.2. *A function $f : \mathbb{R}^n \to \mathbb{R}$ is called Lipschitz continuous with Lipschitz constant L if*

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

We can easily verify that $f(x) = |x|$ is Lipschitz continuous with $L = 1$.

Remark 1.4. *For a continuously differentiable function $f(x)$, by the Mean Value Theomem, we have $\frac{|f(x)-f(y)|}{|x-y|} = |f'(x+\theta(y-x))|$ for some $\theta \in (0,1)$. Assume $|f'(x)|$ is bounded by L for any x, we obtain Lipschitz continuity. Assume Lipschitz continuity, and take the limit $y \to x$, we get $|f'(x)| \leq L$. Thus for a continuously differentible function, Lipschitz continuity is equivalent to boundedness of first order derivative.*

Example 1.2. *Assume $\|\nabla f(\mathbf{x})\| \leq L, \forall \mathbf{x}$, then $f(\mathbf{x})$ is Lipschitz continuous with Lipschitz constant L. Apply the Mean Value Theorem to $g(t) = f(\mathbf{y} + t(\mathbf{x} - \mathbf{y}))$, we get*

$$|g(1)-g(0)| = |g'(\theta)|, \quad \theta \in (0,1) \Rightarrow |f(\mathbf{x})-f(\mathbf{y})| = |\langle \nabla f(\mathbf{y}+\theta(\mathbf{x}-\mathbf{y})), \mathbf{x}-\mathbf{y}\rangle|.$$

With the Cauchy-Schwartz inequality for two vectors $\langle \mathbf{a}, \mathbf{b}\rangle \leq \|\mathbf{a}\|\|\mathbf{b}\|$, we get

$$|f(\mathbf{x})-f(\mathbf{y})| = |\langle \nabla f(\mathbf{y}+\theta(\mathbf{x}-\mathbf{y})), \mathbf{x}-\mathbf{y}\rangle| \leq \|\nabla f(\mathbf{y}+\theta(\mathbf{x}-\mathbf{y}))\|\|\mathbf{x}-\mathbf{y}\| \leq L\|\mathbf{x}-\mathbf{y}\|.$$

Theorem 1.8. *For a twice continuously differentiable function (second-order derivatives exist and are continuous) $f : \mathbb{R}^n \to \mathbb{R}$, if*

$$\|\nabla^2 f(\mathbf{x})\| \leq L, \quad \forall \mathbf{x},$$

where $\|\nabla^2 f(\mathbf{x})\|$ denotes the spectral norm, then $\nabla f(\mathbf{x})$ is Lipschitz continuous with Lipschitz constant L.

Example 1.3. *Let $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T K \mathbf{x} - \mathbf{x}^T \mathbf{b}$ where $\mathbf{b}$ is a given vector and $-K$ is the discrete Laplacian matrix as in Appendix B. Then $\nabla^2 f = K$ and we have $\|\nabla^2 f\| < (n+1)^2$. See Appendix B.*

Proof. By Fundamental Theorem of Calculus on a vector-valued single variable function $g(t) = \nabla f(\mathbf{x} + t\mathbf{h})$, $g(1) - g(0) = \int_0^1 g'(t)dt$ gives

$$\nabla f(\mathbf{x} + \mathbf{h}) - \nabla f(\mathbf{x}) = \int_0^1 \nabla^2 f(\mathbf{x} + t\mathbf{h})\mathbf{h} dt.$$

The definition of spectral norm (See Appendix A.5) gives $\|A\mathbf{x}\| \leq \|A\|\|\mathbf{x}\|$. With Lemma 1.3, we have

$$\begin{aligned}
\|\nabla f(\mathbf{x} + \mathbf{h}) - \nabla f(\mathbf{x})\| &= \left\| \int_0^1 \nabla^2 f(\mathbf{x} + t\mathbf{h})\mathbf{h} dt \right\| \\
&\leq \int_0^1 \left\| \nabla^2 f(\mathbf{x} + t\mathbf{h})\mathbf{h} \right\| dt \\
&\leq \int_0^1 \left\| \nabla^2 f(\mathbf{x} + t\mathbf{h}) \right\| \|\mathbf{h}\| dt \\
&= \int_0^1 \left\| \nabla^2 f(\mathbf{x} + t\mathbf{h}) \right\| dt \|\mathbf{h}\| = L\|\mathbf{h}\|.
\end{aligned}$$

Finally, let $\mathbf{h} = \mathbf{y} - \mathbf{x}$, we get the Lipschitz continuity. □

Remark 1.5. *The proof above can be also be done as the following by Theorem 1.7:*

$$\begin{aligned}
\nabla f(\mathbf{x} + \mathbf{h}) - \nabla f(\mathbf{x}) &= \int_0^1 \nabla^2 f(\mathbf{x} + t\mathbf{h})\mathbf{h} dt \\
&= \left(\int_0^1 \nabla^2 f(\mathbf{x} + t\mathbf{h}) dt \right) \mathbf{h}.
\end{aligned}$$

thus

$$\begin{aligned}
\|\nabla f(\mathbf{x} + \mathbf{h}) - \nabla f(\mathbf{x})\| &\leq \left\| \int_0^1 \nabla^2 f(\mathbf{x} + t\mathbf{h}) dt \right\| \|\mathbf{h}\| \\
&\leq \int_0^1 \left\| \nabla^2 f(\mathbf{x} + t\mathbf{h}) \right\| dt \|\mathbf{h}\| \\
&\leq \int_0^1 L dt \|\mathbf{h}\| = L\|\mathbf{h}\|.
\end{aligned}$$

2

Unconstrained smooth optimization algorithms

In this chapter, we consider the unconstrained smooth optimization, i.e., minimizing $f(\mathbf{x})$ for $\mathbf{x} \in \mathbb{R}^n$.

2.1 Optimality conditions

Definition 2.1. *For $f : \mathbb{R}^n \longrightarrow \mathbb{R}$, $\mathbf{x}^*$ is a global minimizer if $f(\mathbf{x}^*) \leq f(\mathbf{x}), \forall \mathbf{x} \in S$. $\mathbf{x}^*$ is a local minimizer of $f(\mathbf{x})$ if there is a ball $B \subseteq \mathbb{R}^n$ centered at $\mathbf{x}^*$ on which $\mathbf{x}^*$ is the global minimizer of $f(\mathbf{x})$ restricted on B.*

We review the well-known optimality conditions.

Theorem 2.1 (First Order Necessary Conditions)**.** *For a C^1 function (first order derivatives exist and are continuous) $f(\mathbf{x}) : \mathbb{R}^n \longrightarrow \mathbb{R}$, if $\mathbf{x}^*$ is a local minimizer, then $\nabla f(\mathbf{x}^*) = \mathbf{0}$.*

Proof. Assume $\nabla f(\mathbf{x}^*) \neq \mathbf{0}$. Let $\mathbf{p} = -\nabla f(\mathbf{x}^*)$, then $g(t) = \mathbf{p}^T \nabla f(\mathbf{x}^* + t\mathbf{p})$ is a continuous function, thus

$$g(0) = -\|\nabla f(\mathbf{x}^*)\|^2 < 0 \Rightarrow \exists T > 0, \forall t \in [0, T], g(t) < 0.$$

For any fixed $t \in (0, T]$, by Theorem 1.3, there is $\theta \in (0, t)$ s.t.

$$f(\mathbf{x}^* + t\mathbf{p}) = f(\mathbf{x}^*) + t\mathbf{p}^T \nabla f(\mathbf{x}^* + \theta\mathbf{p}) < f(\mathbf{x}^*).$$

So along the line segment connecting $\mathbf{x}^*$ and $\mathbf{x}^* + t\mathbf{p}$ for arbitrarily small t, $f(\mathbf{x}^*)$ is not the smallest function value, which is a contradiction to the fact that $f(\mathbf{x}^*)$ is a local minimizer. □

Definition 2.2. $\mathbf{x}^*$ *is called a stationary point or a critical point of the function $f(\mathbf{x})$ if $\nabla f(\mathbf{x}^*) = \mathbf{0}$.*

Theorem 2.2 (Second Order Necessary Conditions)**.** *For a C^2 function (second order derivatives exist and are continuous) $f(\mathbf{x}) : R^n \longrightarrow \mathbb{R}$, if $\mathbf{x}^*$ is a local minimizer, then $\nabla f(\mathbf{x}^*) = \mathbf{0}$ and $\nabla^2 f(\mathbf{x}^*) \geq 0$ (Hessian matrix is positive semi-definite).*

Proof. Assume $\nabla^2 f(\mathbf{x}^*)$ is not positive semi-definite, then there exists $\mathbf{p} \in \mathbb{R}^n$ s.t. $\mathbf{p}^T \nabla^2 f(\mathbf{x}^*)\mathbf{p} < 0$. The continuity of the function $g(t) = \mathbf{p}^T \nabla^2 f(\mathbf{x}^* + t\mathbf{p})\mathbf{p}$ implies that

$$\exists T > 0, \forall t \in [0, T], \mathbf{p}^T \nabla^2 f(\mathbf{x}^* + t\mathbf{p})\mathbf{p} < 0.$$

For any fixed $t \in (0, T]$, by Theorem 1.4, there is $\theta \in (0, t)$ s.t.

$$f(\mathbf{x}^* + t\mathbf{p}) = f(\mathbf{x}^*) + t\mathbf{p}^T \nabla f(\mathbf{x}^*) + \frac{1}{2} t^2 \mathbf{p}^T \nabla^2 f(\mathbf{x}^* + \theta\mathbf{p})\mathbf{p} < f(\mathbf{x}^*),$$

where we have used Theorem 2.1. So along the line segment connecting $\mathbf{x}^*$ and $\mathbf{x}^* + t\mathbf{p}$ for arbitrarily small t, $f(\mathbf{x}^*)$ is not the smallest function value, which is a contradiction to the fact that $f(\mathbf{x}^*)$ is a local minimizer. □

Theorem 2.3 (Second Order Sufficient Conditions)**.** *For a C^2 function (second order derivatives exist and are continuous) $f(\mathbf{x}) : \mathbb{R}^n \longrightarrow \mathbb{R}$, if $\nabla f(\mathbf{x}^*) = \mathbf{0}$ and $\nabla^2 f(\mathbf{x}^*) > 0$ (Hessian matrix is positive definite), then $\mathbf{x}^*$ is a strict local minimizer.*

Proof. First of all, for the real symmetric Hessian matrix $\nabla^2 f(\mathbf{x})$, positive definiteness means that all eigenvalues are positive.

Second, eigenvalues are continuous functions of matrix entries because polynomial roots are continuous functions of coefficients, thus the smallest eigenvalue of $\nabla^2 f(\mathbf{x})$ is a continuous function of $\mathbf{x}$. Thus, $\nabla^2 f(\mathbf{x}^*) > 0$ implies that there is an open ball centered at $\mathbf{x}^*$ with radius $r > 0$:

$$B = \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x} - \mathbf{x}^*\| < r\}$$

such that $\nabla^2 f(\mathbf{x}) > 0, \forall \mathbf{x} \in B$.

For any $\mathbf{y} \in B$, we have $\mathbf{y} = \mathbf{x}^* + \mathbf{p}$ where $\mathbf{p} \in \mathbb{R}^n$ with $\|\mathbf{p}\| < r$. By Theorem 1.4, there is $\theta \in (0, t)$ s.t.

$$f(\mathbf{x}^* + \mathbf{p}) = f(\mathbf{x}^*) + \mathbf{p}^T \nabla f(\mathbf{x}^*) + \frac{1}{2} \mathbf{p}^T \nabla^2 f(\mathbf{x}^* + \theta\mathbf{p})\mathbf{p} > f(\mathbf{x}^*),$$

which is due to the positive definiteness of $\nabla^2 f(\mathbf{x}^* + \theta\mathbf{p})$ (because $\mathbf{x}^* + \theta\mathbf{p} \in B$). It implies $\mathbf{x}^*$ is a strict local minimizer on the ball B. □

Theorem 2.4. *Assume $f(\mathbf{x}) : \mathbb{R}^n \longrightarrow \mathbb{R}$ is convex.*

1. *Any local minimizer is also a global minimizer.*

2. *If $f(\mathbf{x})$ is also continuously differentiable (the same as C^1 functions), then $\mathbf{x}^*$ is a global minimizer if and only if $\nabla f(\mathbf{x}^*) = \mathbf{0}$.*

Remark 2.1. *A convex function may not have any minimizer at all, e.g., $f(x) = x$.*

Proof. Let $\mathbf{x}^*$ be a local minimizer. For any $\mathbf{y}$, there exists $T > 0$ s.t.

$$\forall t \in (0, T], \quad f(\mathbf{x}^* + t(\mathbf{y} - \mathbf{x}^*)) \geq f(\mathbf{x}^*),$$

because $\mathbf{x}^*$ is a local minimizer. The convexity implies

$$f(\mathbf{x}^* + t(\mathbf{y} - \mathbf{x}^*)) = f((1-t)f\mathbf{x}^* + t\mathbf{y}) \leq (1-t)f(\mathbf{x}^*) + tf(\mathbf{y})$$

thus we get $f(\mathbf{x}^*) \leq f(\mathbf{y})$.

Next, assume $\mathbf{x}^*$ is a global minimizer thus also a local one, then Theorem 2.1 implies $\nabla f(\mathbf{x}^*) = \mathbf{0}$. If assuming $\nabla f(\mathbf{x}^*) = \mathbf{0}$, then Lemma 1.1 implies

$$f(\mathbf{x}) \geq f(\mathbf{x}^*) + \langle \nabla f(\mathbf{x}^*), \mathbf{x} - \mathbf{x}^* \rangle \geq f(\mathbf{x}^*).$$

□

Theorem 2.5. *Assume $f(\mathbf{x}) : \mathbb{R}^n \longrightarrow \mathbb{R}$ is strongly convex and also continuously differentiable (the same as C^1 functions). Then $f(\mathbf{x})$ has a unique global minimizer $\mathbf{x}^*$, which is the only critical point of the function.*

Proof. By Theorem 2.4, we only need to show $f(\mathbf{x})$ has a global minimum and the minimizer is unique.

By Theorem 1.1, we have

$$f(\mathbf{x}) \geq f(\mathbf{y}) + \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle + \frac{\mu}{2}\|\mathbf{x} - \mathbf{y}\|^2, \quad \forall \mathbf{x}, \mathbf{y}.$$

Plug in $\mathbf{y} = \mathbf{0}$, we get

$$f(\mathbf{x}) \geq f(\mathbf{0}) + \langle \nabla f(\mathbf{0}), \mathbf{x} \rangle + \frac{\mu}{2}\|\mathbf{x}\|^2,$$

which implies $f(\mathbf{x}) \to +\infty$ as $\|\mathbf{x}\| \to \infty$. Thus for any fixed number M, there is $R > 0$ s.t.,

$$f(\mathbf{x}) > M, \quad \forall \mathbf{x} \text{ satisfying } \|\mathbf{x}\| > R.$$

In particular, consider the $R > 0$ for $M = f(\mathbf{0})$, and the close ball

$$B = \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\| \leq R\}.$$

The closed ball B is a compact set thus $f(\mathbf{x})$ attains its minimum on B, see Appendix C. Let $\mathbf{x}^*$ be one minimizer of $f(\mathbf{x})$ on B, then $\mathbf{x}^*$ is the global minimizer because $f(\mathbf{x}^*) \leq f(\mathbf{0}) = M$.

Let $\mathbf{x}^*, \mathbf{y}^*$ be two global minimizers, then

$$f(\mathbf{x}^*) \geq f(\mathbf{y}^*) + \langle \nabla f(\mathbf{y}^*), \mathbf{x}^* - \mathbf{y}^* \rangle + \frac{\mu}{2}\|\mathbf{x}^* - \mathbf{y}^*\|^2 \Rightarrow \frac{\mu}{2}\|\mathbf{x}^* - \mathbf{y}^*\|^2 \leq 0 \Rightarrow \mathbf{x}^* = \mathbf{y}^*,$$

where we have used $\nabla f(\mathbf{y}^*) = 0$ and $f(\mathbf{x}^*) = f(\mathbf{y}^*)$. □

Similar proof also gives

Theorem 2.6. *Assume $f(\mathbf{x}) : \mathbb{R}^n \longrightarrow \mathbb{R}$ is strictly convex and also continuously differentiable. If $f(\mathbf{x})$ has a global minimizer $\mathbf{x}^*$, then it is unique and also the only critical point.*

Remark 2.2. *Strict convexity is not enough to ensure the existence of a minimizer. For instance, $f(x) = e^x$ is strictly convex.*

2.2 Gradient descent

The gradient descent method with a constant step size $\eta > 0$ is the most popular and also the simplest algorithm for minimizing $f(\mathbf{x})$:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \eta \nabla f(\mathbf{x}_k), \quad \eta > 0. \tag{2.1}$$

In this section, we need to assume the gradient $\nabla f(\mathbf{x})$ is Lipschitz continuous, which however does not necessarily imply $f(\mathbf{x})$ is Lipschitz continuous. For example, $f(x) = x^2$ is not Lipschitz continuous because $f'(x) = 2x$ is not a bounded function (see Remark 1.4), but $f'(2x) = 2x$ is Lipschitz continuous because its derivative is a constant.

2.2.1 Stable step sizes

Lemma 2.1 (Descent Lemma)**.** *Assume $\nabla f(\mathbf{x})$ is Lipschitz-continuous with Lipschitz constant L, then*

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{L}{2} \|\mathbf{x} - \mathbf{y}\|^2.$$

Remark 2.3. *Notice that there is no assumption on the existence of Hessian. But if assuming $\|\nabla^2 f\| \leq L$, then by Theorem 1.4,*

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{1}{2} (\mathbf{x} - \mathbf{y})^T \nabla^2 f(\mathbf{z}) (\mathbf{x} - \mathbf{y})$$

which implies

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{L}{2} \|\mathbf{x} - \mathbf{y}\|^2,$$
$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle - \frac{L}{2} \|\mathbf{x} - \mathbf{y}\|^2, .$$

where the spectral $\|\nabla^2 f\|$ is the largest singular value thus also the largest magnitude of eigenvalue for a real symmetric matrix, and we have used the Courant-Fischer-Weyl min-max inequality, see Appendix A.1.

Remark 2.4. *Notice that there is no assumption on convexity. But if assuming strong convexity of $f(\mathbf{x})$, by Theorem 1.1,*

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{\mu}{2}\|\mathbf{x} - \mathbf{y}\|^2.$$

Proof. Let $g(t) = f(\mathbf{x} + t(\mathbf{y} - \mathbf{x}))$. The fundamental theorem of calculus gives

$$g(1) - g(0) = \int_0^1 g'(t)dt,$$

thus

$$f(\mathbf{y}) - f(\mathbf{x}) = \int_0^1 \langle \nabla f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})), \mathbf{y} - \mathbf{x} \rangle dt.$$

Let $\mathbf{z}(t) = \mathbf{x} + t(\mathbf{y} - \mathbf{x})$. Then by subtracting $\langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle$ from both sides, we get

$$\begin{aligned}
|f(\mathbf{y}) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle| &= \left| \int_0^1 \langle \nabla f(\mathbf{z}(t)) - f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle dt \right| \\
&\leq \int_0^1 \left| \langle \nabla f(\mathbf{z}(t)) - f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \right| dt \\
\text{(Cauchy-Schwart inequality)} &\leq \int_0^1 \|\nabla f(\mathbf{z}(t)) - f(\mathbf{x})\| \|\mathbf{y} - \mathbf{x}\| dt \\
&= \int_0^1 \|\nabla f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) - f(\mathbf{x})\| dt \|\mathbf{y} - \mathbf{x}\| \\
&\leq \left(\int_0^1 Lt \|\mathbf{y} - \mathbf{x}\| dt \right) \|\mathbf{y} - \mathbf{x}\| = \frac{L}{2}\|\mathbf{y} - \mathbf{x}\|^2.
\end{aligned}$$

□

Remark 2.5. *The proof also implies*

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle - \frac{L}{2}\|\mathbf{x} - \mathbf{y}\|^2.$$

Lemma 2.2 (Sufficient Decrease Lemma). *Assume $\nabla f(\mathbf{x})$ is Lipschitz-continuous with Lipschitz constant L, then the gradient descent method* (2.1) *satisfies*

$$f(\mathbf{x}) - f(\mathbf{x} - \eta \nabla f(\mathbf{x})) \geq \eta(1 - \frac{L}{2}\eta)\|\nabla f(\mathbf{x})\|^2, \quad \forall \mathbf{x}, \forall \eta > 0.$$

Proof. Lemma 2.1 gives

$$f(\mathbf{x} - \eta \nabla f(\mathbf{x})) \leq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), -\eta \nabla f(\mathbf{x}) \rangle + \frac{L}{2}\|\eta \nabla f(\mathbf{x})\|^2.$$

□

Lemma 2.2 implies that the gradient descent method (2.1) decreases the cost function, i.e., $f(\mathbf{x}_{k+1}) < f(\mathbf{x}_k)$ for any $\eta \in (0, \frac{L}{2})$.

In practice, it is difficult to obtain the exact value of L. But any small enough positive step size η can make the iteration (2.1) stable in the sense of not blowing up, e.g., $f(\mathbf{x}_{k+1}) < f(\mathbf{x}_k)$.

Consider an ordinary differential equation (ODE) system:

$$\frac{d}{dt}\mathbf{u}(t) = F(\mathbf{u}(t)), \quad \mathbf{u}(0) = \mathbf{u}_0,$$

where $\mathbf{u} = \begin{bmatrix} u_1(t) & u_2(t) & \cdots & u_n(t) \end{bmatrix}^T$. The simplest forward Euler scheme for this ODE system is

$$\mathbf{u}_{k+1} = \mathbf{u}_k + \Delta t F(\mathbf{u}_k). \tag{2.2}$$

If setting $F = -\nabla f$ and $\Delta t = \eta$, then the gradient descent method (2.1) can be regarded as the forward Euler scheme above. However, usually (2.2) is used for approximating the time-dependent solution $\mathbf{u}(t)$, whereas the (2.1) is used for finding the minimizer (the steady state ODE solution $F(\mathbf{u}) = 0$).

Nonetheless, since (2.2) is exactly the same as (2.1), the stability requirement from numerically solving ODE should give the same result as $\eta \leq \frac{L}{2}$.

Example 2.1. *Consider solving the initial boundary value problem for the one-dimensional heat equation*

$$\begin{cases} u_t(x,t) = u_{xx}(x,t), & x \in (0,1) \\ u(x,0) = u_0(x), & x \in (0,1) \\ u(x,0) = u(x,1) = 0 \end{cases}.$$

With the second order discrete Laplacian in Appendix B, a semi-discrete scheme defined on a uniform grid $x_i = i\Delta x$ with $\Delta x = \frac{1}{n+1}$ can be written as an ordinary differential equation (ODE) system:

$$\frac{d}{dt}\mathbf{u}(t) = K\mathbf{u}(t), \quad \mathbf{u}(0) = \mathbf{u}_0,$$

where $\mathbf{u} = \begin{bmatrix} u_1(t) & u_2(t) & \cdots & u_n(t) \end{bmatrix}^T$ and $u_i(t)$ approximates $u(x_i, t)$. The simplest forward Euler scheme for this ODE system is

$$\mathbf{u}_{k+1} = \mathbf{u}_k + \Delta t K \mathbf{u}_k \tag{2.3}$$

The linear ODE solver stability requirement $\|\mathbf{u}_{k+1}\| \leq \|\mathbf{u}_k\|$ gives $\Delta t \leq \frac{1}{2}\Delta x^2$ by using eigenvalues of K given in Appendix B.

If regarding (2.3) as the gradient descent method, then $f(\mathbf{u}) = \frac{1}{2}\mathbf{u}^T K \mathbf{u}$, and $\|\nabla^2 f\| = \|K\| < \frac{1}{\Delta x^2}$ as in Appendix B. This implies the gradient ∇f is Lipschitz-continuous with $L = \frac{1}{\Delta x^2}$, thus $\eta < \frac{2}{L}$ gives $\eta < \frac{1}{2}\Delta x^2$.

2.2.2 Convergence for Lipschitz continuous ∇f

Theorem 2.7. *Assume $\nabla f(\mathbf{x})$ is Lipschitz-continuous with Lipschitz constant L, and assume $f(\mathbf{x})$ has a global minimizer: $f(\mathbf{x}) \geq f(\mathbf{x}_*), \forall \mathbf{x}$. Then for the gradient descent method* (2.1) *with a constant step size $\eta \in (0, \frac{2}{L})$, the following holds:*

1.
$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) \leq -\eta(1 - \frac{L}{2}\eta)\|\nabla f(\mathbf{x}_k)\|^2 \leq 0. \tag{2.4}$$

2. *The sequence $\{f(\mathbf{x}_k)\}$ converges.*

3. $\lim\limits_{k\to\infty} \|\nabla f(\mathbf{x}_k)\| = 0.$

4.
$$\max_{0\leq k\leq n} \|\nabla f(\mathbf{x}_k)\| \leq \frac{1}{\sqrt{n+1}}\sqrt{\frac{1}{\eta(1-\frac{L}{2}\eta)}\left[f(\mathbf{x}_0) - f(\mathbf{x}_*)\right]}.$$

Remark 2.6. *Notice that none of the conclusions can imply the sequence $\{\mathbf{x}_k\}$ converges to a critical point. As a matter of fact, $\{\mathbf{x}_k\}$* **may not have a limit***. See an example below.*

Proof. First of all, by plugging $\mathbf{y} = \mathbf{x}_k - \eta\nabla f(\mathbf{x}_k)$ into Lemma 2.2, we get

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) \leq -\eta(1 - \frac{L}{2}\eta)\|\nabla f(\mathbf{x}_k)\|^2.$$

Second, since $\eta \in (0, \frac{2}{L})$, we have $f(\mathbf{x}_{k+1}) < f(\mathbf{x}_k)$ thus $\{f(\mathbf{x}_k)\}$ is a decreasing sequence. Moreover, $f(\mathbf{x}_k)$ has a lower bound $f(\mathbf{x}_k) \geq f(\mathbf{x}_*)$. Thus, the sequence $\{f(\mathbf{x}_k)\}$ is bounded from below and decreasing, thus it has a limit (a bounded monotone sequence has a limit, see Appendix C).

Let $\omega = \eta(1 - \frac{L}{2}\eta)$, then $\omega > 0$. By summing up (2.4), we get

$$\sum_{k=0}^{N} \|\nabla f(\mathbf{k})\|^2 \leq \frac{1}{\omega}\left[f(\mathbf{x}_0) - f(\mathbf{x}_{N+1})\right] \leq \frac{1}{\omega}\left[f(\mathbf{x}_0) - \lim_{k\to\infty} f(\mathbf{x}_k)\right],$$

because $\{-f(\mathbf{x}_k)\}$ is an increasing sequence.

So $\sum\limits_{k=0}^{N} \|\nabla f(\mathbf{k})\|^2$ is an increasing and bounded above sequence, thus it converges, which implies the convergence of the infinite series

$$\sum_{k=0}^{\infty} \|\nabla f(\mathbf{k})\|^2 = \lim_{N\to\infty} \sum_{k=0}^{N} \|\nabla f(\mathbf{k})\|^2.$$

The convergence of the series further implies (see Appendix C.4)

$$\lim_{k\to\infty} \|\nabla f(\mathbf{k})\|^2 = 0 \Rightarrow \lim_{k\to\infty} \|\nabla f(\mathbf{k})\| = 0.$$

Let $g_n = \max_{0\leq k\leq n} \|\nabla f(\mathbf{x}_k)\|$, then

$$(n+1)g_n^2 \leq \sum_{k=0}^{n} \|\nabla f(\mathbf{k})\|^2 \leq \frac{1}{\omega}\left[f(\mathbf{x}_0) - f(\mathbf{x}_{n+1})\right] \leq \frac{1}{\omega}\left[f(\mathbf{x}_0) - f(\mathbf{x}_*)\right],$$

□

Next, in order to understand the convergence of $\{\mathbf{x}_k\}$, we discuss sufficient conditions for its convergence. For example, assume $\sum\limits_{k=0}^{\infty} \|\nabla f(\mathbf{x}_k)\|$ converges, then we can prove the convergence of $\{\mathbf{x}_k\}$ as the following.

Define $\mathbf{y}_n = \sum\limits_{k=0}^{n} (\mathbf{x}_{k+1} - \mathbf{x}_k) = \eta \sum\limits_{k=0}^{n} \nabla f(\mathbf{x}_k)$, then for any $m \geq n$

$$\begin{aligned}\|\mathbf{y}_n - \mathbf{y}_m\| &= \eta \left\| \sum_{k=n+1}^{m} \nabla f(\mathbf{x}_k) \right\| \\ &\leq \eta \sum_{k=n+1}^{m} \|\nabla f(\mathbf{x}_k)\|\end{aligned}$$

We need to use the notion of Cauchy sequence (see Appendix C.3). The convergence of $\sum\limits_{k=0}^{\infty} \|\nabla f(\mathbf{x}_k)\|$ implies $a_n = \sum\limits_{k=0}^{n} \|\nabla f(\mathbf{x}_k)\|$ is a Cauchy sequence, thus

$$\forall \varepsilon > 0, \exists N, \forall m, n \geq N, |a_m - a_n| < \varepsilon.$$

So $\mathbf{y}_n$ is also a Cauchy sequence, because

$$\forall \varepsilon > 0, \exists N, \forall m, n \geq N, \|\mathbf{y}_n - \mathbf{y}_m\| \leq \eta |a_m - a_n| < \eta\varepsilon.$$

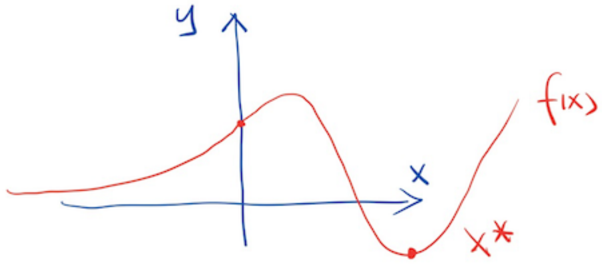
Therefore, $\mathbf{y}_n$ has a limit, which further implies the convergence of $\mathbf{x}_k$. **However, the assumption of convergence of $\sum\limits_{k=0}^{\infty} \|\nabla f(\mathbf{x}_k)\|$ is in general not true**. By the proof of the theorem above, we only have the convergence $\sum\limits_{k=0}^{\infty} \|\nabla f(\mathbf{x}_k)\|^2$, which does not implies the convergence of $\sum\limits_{k=0}^{\infty} \|\nabla f(\mathbf{x}_k)\|$. A quick counter-example would be $\|\nabla f(\mathbf{x}_k)\| = \frac{1}{k}$ (see Appendix C on why $\sum\limits_{k=0}^{\infty} \frac{1}{k^2}$ converges but $\sum\limits_{k=0}^{\infty} \frac{1}{k}$ diverges).

Example 2.2. *We construct an example for which the gradient descent method produces almost $\|\nabla f(\mathbf{x}_k)\| = \frac{1}{k}$. Consider the following function*

$$f(x) = \begin{cases} e^x, & x \leq 0 \\ g(x), & x > 0 \end{cases},$$

where we pick a function $g(x)$ such that

1. *$f(x)$ is very smooth;*
2. *$|f''(x)| \leq 1$ for any x, which implies $f'(x)$ is L-continuous with $L = 1$;*
3. *$f(x)$ has a global minimizer x_*.*



For instance, see the plotted function $f(x)$, which can satisfy all the assumptions of Theorem 2.7, with Lipschitz constant $L = 1$ for the derivative function $f'(x)$.

So a stable step size can be chosen as any positive $\eta < 2$. We consider the following gradient descent iteration with $\eta = 1$:

$$\begin{cases} x_{k+1} = x_k - f'(x_k) \\ x_0 = 0 \end{cases} .$$

Notice that all iterates x_k stays non-positive, it can also be written as

$$x_{k+1} = x_k - e^{x_k}, \quad x_0 = 0.$$

One can easily implement this on MATLAB to verify that numerically we have $|f'(x_k)| \approx \frac{1}{k}$ for this iteration.

```
% A MATLAB code of an example for Gradient Descent
% producing non-convergent x_k, which goes to infinity.
% The cost fuction f(x)=e^x if x≤0.
% Must use zero initial guess and step size eta=1.
x=0;
eta=1;
figure;
for k=0:10000000
    x=x-eta*exp(x); % simple Gradient Descent
    if (mod(k,10000)==0 | k≤100)
        % plot the iterates (x_k, f(x_k)) the first 100
        % then every 10,000 iterations

        semilogy(x,exp(x),'o');
        xlabel('x_k')
        ylabel('log[f(x_k)]')
        hold all
        drawnow

    end
    % print values of [|f'(x_k)|-1/k](1/k): an indicator
    % of how close |f'(x_k)| is to 1/k
    fprintf('%d %d \n', k, abs(exp(x)-1/k)*k)
end
```

More importantly, Theorem 2.7 implies that $e^{x_k} = |f'(x_k)| \to 0$ thus $x_k \to -\infty$. Even though we can informally write it as $x_k \to -\infty$, the sequence $\{x_k\}$ diverges because it is not Cauchy (see Appendix C.3), e.g., it does not have any cluster point.

So in the example above, we can see that Lipschitz-continuity of ∇f may not ensure the convergence of the gradient descent to even a critical point!

2.2.3 Convergence for convex functions

Theorem 2.8. *Assume $\nabla f(\mathbf{x})$ is Lipschitz-continuous with Lipschitz constant L and $f(\mathbf{x}) : \mathbb{R}^n \longrightarrow \mathbb{R}$ is convex. Then for any $\mathbf{x}, \mathbf{y}$:*

1. $f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{1}{2L} \|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|^2$
2. $\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|^2 \leq L \langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle$.

Remark 2.7. *Without convexity, by the proof of Lemma 2.1, we only have*

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{L}{2} \|\mathbf{x} - \mathbf{y}\|^2,$$

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle - \frac{L}{2} \|\mathbf{x} - \mathbf{y}\|^2.$$

With strong convexity, we can have

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{\mu}{2} \|\mathbf{x} - \mathbf{y}\|^2.$$

Proof. Define $\phi(\mathbf{x}) = f(\mathbf{x}) - \langle \nabla f(\mathbf{x}_0), \mathbf{x} \rangle$. Then $\phi(\mathbf{x})$ also has Lipschitz continuous gradient:

$$\|\nabla \phi(\mathbf{x}) - \nabla \phi(\mathbf{y})\| = \|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L \|\mathbf{x} - \mathbf{y}\|.$$

Apply Lemma 2.1 to $\phi(\mathbf{x})$:

$$\begin{aligned} \phi(\mathbf{x}) &\leq \phi(\mathbf{y}) + \langle \nabla \phi(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle + \frac{L}{2} \|\mathbf{x} - \mathbf{y}\|^2 \\ (|\langle \mathbf{a}, \mathbf{b} \rangle| \leq \|\mathbf{a}\| \|\mathbf{b}\|) \quad &\leq \phi(\mathbf{y}) + \|\nabla \phi(\mathbf{y})\| \|\mathbf{x} - \mathbf{y}\| + \frac{L}{2} \|\mathbf{x} - \mathbf{y}\|^2 \end{aligned}$$

By Theorem 1.5, $\phi(\mathbf{x})$ is also convex because $-\langle \nabla f(\mathbf{x}_0), \mathbf{x} \rangle$ is convex. Moreover, $\nabla \phi(\mathbf{x}_0) = \mathbf{0}$, thus by Theorem 2.4, $\mathbf{x}_0$ is a global minimizer of $\nabla \phi(\mathbf{x})$. So we get

$$\begin{aligned}
\phi(\mathbf{x}_0) = \min_{\mathbf{x}} \phi(\mathbf{x}) &\leq \min_{\mathbf{x}} \left[\phi(\mathbf{y}) + \|\nabla \phi(\mathbf{y})\| \|\mathbf{x} - \mathbf{y}\| + \frac{L}{2} \|\mathbf{x} - \mathbf{y}\|^2 \right] \\
&\leq \min_{r \geq 0} \left[\phi(\mathbf{y}) + \|\nabla \phi(\mathbf{y})\| r + \frac{L}{2} r^2 \right] \\
&= \phi(\mathbf{y}) - \frac{1}{2L} \|\nabla \phi(\mathbf{y})\|^2.
\end{aligned}$$

Thus $\phi(\mathbf{x}_0) \leq \phi(\mathbf{y}) - \frac{1}{2L} \|\nabla \phi(\mathbf{y})\|^2$ implies

$$f(\mathbf{x}_0) - \langle \nabla f(\mathbf{x}_0), \mathbf{x}_0 \rangle \leq f(\mathbf{y}) - \langle \nabla f(\mathbf{x}_0), \mathbf{y} \rangle - \frac{1}{2L} \|\nabla f(\mathbf{y}) - \nabla f(\mathbf{x}_0)\|^2.$$

Since $\mathbf{x}_0, \mathbf{y}$ are arbitrary, we can also write is as

$$f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{x} \rangle \leq f(\mathbf{y}) - \langle \nabla f(\mathbf{x}), \mathbf{y} \rangle - \frac{1}{2L} \|\nabla f(\mathbf{y}) - \nabla f(\mathbf{x})\|^2,$$

which implies

$$f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{1}{2L} \|\nabla f(\mathbf{y}) - \nabla f(\mathbf{x})\|^2 \leq f(\mathbf{y}).$$

Switching $\mathbf{x}$ and $\mathbf{y}$, we get

$$f(\mathbf{y}) + \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle + \frac{1}{2L} \|\nabla f(\mathbf{y}) - \nabla f(\mathbf{x})\|^2 \leq f(\mathbf{x}),$$

and adding two we get

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|^2 \leq L \langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle.$$

□

Theorem 2.9. *Assume $f(\mathbf{x}) : \mathbb{R}^n \longrightarrow \mathbb{R}$ is convex and $\nabla f(\mathbf{x})$ is Lipschitz-continuous with Lipschitz constant L, and assume $f(\mathbf{x})$ has a global minimizer: $f(\mathbf{x}) \geq f(\mathbf{x}_*), \forall \mathbf{x}$. Then for the gradient descent method* (2.1) *with a constant step size $\eta \in (0, \frac{2}{L})$, in addition to conclusions in Theorem 2.7, the following holds:*

$$f(\mathbf{x}_k) - f(\mathbf{x}_*) \leq \frac{1}{\frac{1}{f(\mathbf{x}_0) - f(\mathbf{x}_*)} + k\omega \frac{1}{\|\mathbf{x}_0 - \mathbf{x}_*\|^2}} < \frac{1}{k\omega} \|\mathbf{x}_0 - \mathbf{x}_*\|^2,$$

where $\omega = \eta(\frac{2}{L} - \eta)$.

Remark 2.8. *We obtain convergence rate $\mathcal{O}(\frac{1}{k})$, assuming only convexity of the cost function and Lipschitz-continuity of its gradient. We cannot expect convergence of $\mathbf{x}_k$ to $\mathbf{x}_*$ because a convex function may have multiple global minimizers, e.g., $f(\mathbf{x}) \equiv 0$.*

Proof. Define $r_k = \|\mathbf{x}_k - \mathbf{x}_*\|$. With $\nabla f(\mathbf{x}_*) = \mathbf{0}$, we get

$$\begin{aligned}
r_{k+1}^2 &= \|\mathbf{x}_{k+1} - \mathbf{x}_*\|^2 \\
&= \|\mathbf{x}_k - \eta\|\nabla f(\mathbf{x}_k) - \mathbf{x}_*\|^2 \\
&= \|\mathbf{x}_k - \mathbf{x}_*\|^2 + \|\eta\nabla f(\mathbf{x}_k)\|^2 + 2\langle \mathbf{x}_k - \mathbf{x}_*, -\eta\nabla f(\mathbf{x}_k)\rangle \\
&= \|\mathbf{x}_k - \mathbf{x}_*\|^2 + \eta^2\|\nabla f(\mathbf{x}_k)\|^2 - 2\eta\langle \mathbf{x}_k - \mathbf{x}_*, \nabla f(\mathbf{x}_k) - \nabla f(\mathbf{x}_*)\rangle \\
&\le \|\mathbf{x}_k - \mathbf{x}_*\|^2 + \eta^2\|\nabla f(\mathbf{x}_k)\|^2 - \frac{2}{L}\eta\|\nabla f(\mathbf{x}_k) - \nabla f(\mathbf{x}_*)\|^2 \\
&= r_k^2 + (\eta^2 - \frac{2}{L}\eta)\|\nabla f(\mathbf{x}_k)\|^2,
\end{aligned}$$

where we have used Theorem 2.8 in the last inequality.

Define $R_k = f(\mathbf{x}_k) - f(\mathbf{x}_*)$. By Lemma 1.1, we have

$$f(\mathbf{x}) \ge f(\mathbf{x}_k) + \langle \nabla f(\mathbf{x}_k), \mathbf{x} - \mathbf{x}_k\rangle, \quad \forall \mathbf{x},$$

thus

$$f(\mathbf{x}_*) \ge f(\mathbf{x}_k) + \langle \nabla f(\mathbf{x}_k), \mathbf{x}_* - \mathbf{x}_k\rangle.$$

With Cauchy-Schwartz inequality,

$$f(\mathbf{x}_k) - f(\mathbf{x}_*) \le -\langle \nabla f(\mathbf{x}_k), \mathbf{x}_* - \mathbf{x}_k\rangle \le \|\nabla f(\mathbf{x}_k)\|\|\mathbf{x}_* - \mathbf{x}_k\|,$$

which can be written as

$$R_k \le r_k\|\nabla f(\mathbf{x}_k)\|$$

thus

$$-\|\nabla f(\mathbf{x}_k)\| \le \frac{R_k}{r_k}.$$

Recall Theorem 2.7 gives

$$f(\mathbf{x}_{k+1}) \le f(\mathbf{x}_k) - \omega\|\nabla f(\mathbf{x}_k)\|^2,$$

thus

$$\begin{aligned}
f(\mathbf{x}_{k+1}) - f(\mathbf{x}_*) &\le f(\mathbf{x}_k) - f(\mathbf{x}_*) - \omega\|\nabla f(\mathbf{x}_k)\|^2, \\
0 \le R_{k+1} \le R_k - \omega\|\nabla f(\mathbf{x}_k)\|^2 &\le R_k - \omega\frac{R_k^2}{r_k^2}.
\end{aligned}$$

Multiplying both sides by $\frac{1}{R_k R_{k+1}}$, we get

$$\frac{1}{R_k} \le \frac{1}{R_{k+1}} - \omega\frac{1}{r_k^2}\frac{R_k}{R_{k+1}}$$

$$\frac{1}{R_{k+1}} \ge \frac{1}{R_k} + \omega\frac{1}{r_k^2}\frac{R_k}{R_{k+1}} \ge \frac{1}{R_k} + \omega\frac{1}{r_k^2}.$$

Summing it up for all $k = 0, 1, \cdots, N$, we get

$$\frac{1}{R_{N+1}} \ge \frac{1}{R_0} + \omega\sum_{k=0}^{N}\frac{1}{r_k^2} \ge \frac{1}{R_0} + \omega(N+1)\frac{1}{r_0^2}.$$

□

Example 2.3. *Consider minimizing $f(x) = \frac{1}{4}x^4$. Its derivative $f'(x) = x^3$ is NOT Lipschitz continuous because $f''(x) = 3x^2$ is not bounded. Theorem 2.9 in this section can still apply, because $f'(x) = x^3$ is Lipschitz continuous with $L = 3a^2$ on the interval $x \in [-a, a]$, and the gradient descent with $x_0 = a$ and sufficiently small step size satisfies $x_k \in [-a, a]$.*

2.2.4 Convergence for strongly convex functions

Now we consider a strongly convex function $f(\mathbf{x}) : \mathbb{R}^n \longrightarrow \mathbb{R}$ with parameter $\mu > 0$, and assume $\nabla f(\mathbf{x})$ is Lipschitz continuous with Lipschitz constant L. Then Lemma 1.1 gives

$$\langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \geq \mu \|\mathbf{x} - \mathbf{y}\|^2,$$

and Lipschitz continuity with Cauchy Schwartz inequality gives

$$\langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \leq \|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \|\mathbf{x} - \mathbf{y}\| \leq L \|\mathbf{x} - \mathbf{y}\|^2.$$

Thus $\mu \leq L$ and the $Q_f = \frac{L}{\mu}$ can be called *the condition number* of the function $f(\mathbf{x})$.

Example 2.4. *Consider a quadratic function $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T K \mathbf{x} - \mathbf{x}^T \mathbf{b}$ with the negative discrete Laplacian matrix K, then $\nabla^2 f(\mathbf{x}) = K > 0$. Let σ_1 and σ_n be the largest and the smallest singular values of K, respectively. Then by Appendix B, we have*

$$\sigma_n I \leq K \leq \sigma_1 I,$$

which implies that the Lipschitz constant L for ∇f (see Theorem 1.8) is σ_1. By Lemma 1.2, the strong convexity parameter $\mu = \sigma_n$. The number $\frac{\sigma_1}{\sigma_n}$ is also called the condition number of the matrix K. So the condition number of a strongly convex function with Lipschitz continuous gradient, is also the condition number of the Hessian matrix, if the Hessian matrix is a constant matrix.

Theorem 2.10. *For a function $f(\mathbf{x}) : \mathbb{R}^n \longrightarrow \mathbb{R}$ with continuous gradient $\nabla f(\mathbf{x})$, the assumptions that $f(\mathbf{x})$ is convex and $\nabla f(\mathbf{x})$ is Lipschitz-continuous with Lipschitz constant L are equivalent to the following for any $\mathbf{x}, \mathbf{y}$:*

$$0 \leq f(\mathbf{y}) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \leq \frac{L}{2} \|\mathbf{x} - \mathbf{y}\|^2. \tag{2.5}$$

$$f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{1}{2L} \|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|^2 \leq f(\mathbf{y}). \tag{2.6}$$

$$\frac{1}{L} \|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|^2 \leq \langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle. \tag{2.7}$$

$$0 \leq \langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \leq L \|\mathbf{x} - \mathbf{y}\|^2. \tag{2.8}$$

Proof. The proof is done by the following steps:

1. convexity of $f(\mathbf{x})$ and Lipschitz continuity of $\nabla f(\mathbf{x})$ imply (2.5);
2. (2.5) implies (2.6);
3. (2.6) implies (2.7);
4. (2.7) implies convexity of $f(\mathbf{x})$ and Lipschitz continuity of $\nabla f(\mathbf{x})$;
5. (2.8) is equivalent to (2.5).

First of all, assume $f(\mathbf{x})$ is convex and $\nabla f(\mathbf{x})$ is Lipschitz-continuous with Lipschitz constant L, then (2.5) holds because of the first order condition of convexity (Lemma 1.1) and descent lemma (Lemma 2.1).

Second, assume (2.5) holds, then (2.5) implies $\phi(\mathbf{x}) = f(\mathbf{x}) - \langle \nabla f(\mathbf{x}_0), \mathbf{x}\rangle$ satisfies

$$0 \leq \phi(\mathbf{x}) - \phi(\mathbf{y}) - \langle \nabla\phi(\mathbf{y}), \mathbf{x}-\mathbf{y}\rangle$$

and

$$\begin{aligned} \phi(\mathbf{x}) &\leq \phi(\mathbf{y}) + \langle \nabla\phi(\mathbf{y}), \mathbf{x}-\mathbf{y}\rangle + \frac{L}{2}\|\mathbf{x}-\mathbf{y}\|^2 \\ (|\langle \mathbf{a},\mathbf{b}\rangle| \leq \|\mathbf{a}\|\|\mathbf{b}\|) \qquad &\leq \phi(\mathbf{y}) + \|\nabla\phi(\mathbf{y})\|\|\mathbf{x}-\mathbf{y}\| + \frac{L}{2}\|\mathbf{x}-\mathbf{y}\|^2. \end{aligned}$$

By Lemma 1.1, $\phi(\mathbf{x})$ is also convex. Moreover, $\nabla\phi(\mathbf{x}_0) = \mathbf{0}$, thus by Theorem 2.4, $\mathbf{x}_0$ is a global minimizer of $\nabla\phi(\mathbf{x})$. So we get

$$\phi(\mathbf{x}_0) = \min_{\mathbf{x}} \phi(\mathbf{x}) \leq \phi(\mathbf{y}) + \|\nabla\phi(\mathbf{y})\|\|\mathbf{x}-\mathbf{y}\| + \frac{L}{2}\|\mathbf{x}-\mathbf{y}\|^2$$

thus

$$\begin{aligned} \phi(\mathbf{x}_0) &\leq \min_{\mathbf{x}} \left[\phi(\mathbf{y}) + \|\nabla\phi(\mathbf{y})\|\|\mathbf{x}-\mathbf{y}\| + \frac{L}{2}\|\mathbf{x}-\mathbf{y}\|^2\right] \\ &\leq \min_{r\geq 0} \left[\phi(\mathbf{y}) + \|\nabla\phi(\mathbf{y})\| r + \frac{L}{2} r^2\right] \\ &= \phi(\mathbf{y}) - \frac{1}{2L}\|\nabla\phi(\mathbf{y})\|^2. \end{aligned}$$

Thus $\phi(\mathbf{x}_0) \leq \phi(\mathbf{y}) - \frac{1}{2L}\|\nabla\phi(\mathbf{y})\|^2$ implies

$$f(\mathbf{x}_0) - \langle \nabla f(\mathbf{x}_0), \mathbf{x}_0\rangle \leq f(\mathbf{y}) - \langle \nabla f(\mathbf{x}_0), \mathbf{y}\rangle - \frac{1}{2L}\|\nabla f(\mathbf{y}) - \nabla f(\mathbf{x}_0)\|^2.$$

Since $\mathbf{x}_0, \mathbf{y}$ are arbitrary, we can also write is as

$$f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{x}\rangle \leq f(\mathbf{y}) - \langle \nabla f(\mathbf{x}), \mathbf{y}\rangle - \frac{1}{2L}\|\nabla f(\mathbf{y}) - \nabla f(\mathbf{x})\|^2,$$

which implies

$$f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{1}{2L} \|\nabla f(\mathbf{y}) - \nabla f(\mathbf{x})\|^2 \leq f(\mathbf{y}).$$

Switching $\mathbf{x}$ and $\mathbf{y}$, we get

$$f(\mathbf{y}) + \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle + \frac{1}{2L} \|\nabla f(\mathbf{y}) - \nabla f(\mathbf{x})\|^2 \leq f(\mathbf{x}),$$

and adding two we get (2.7).

Third, assume (2.7) holds, then $\langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \geq 0$ implies the convexity by Lemma 1.1, and Cauchy-Schwartz inequality gives Lipschitz continuity by

$$\frac{1}{L} \|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|^2 \leq \langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \leq \|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \|\mathbf{x} - \mathbf{y}\|.$$

Finally, we want to show (2.8) is equivalent to (2.5). Assume (2.5) holds, we get (2.8) by adding the following two:

$$0 \leq f(\mathbf{y}) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \leq \frac{L}{2} \|\mathbf{x} - \mathbf{y}\|^2,$$
$$0 \leq f(\mathbf{x}) - f(\mathbf{y}) - \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \leq \frac{L}{2} \|\mathbf{y} - \mathbf{x}\|^2.$$

Assume (2.8) holds, we get (2.5) by Fundamental Theorem of Calculus on $g(t) = f(\mathbf{x} + t(\mathbf{y} - \mathbf{x}))$:

$$f(\mathbf{y}) - f(\mathbf{x}) = \int_0^1 \langle \nabla f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})), \mathbf{y} - \mathbf{x} \rangle dt$$

$$\begin{aligned} \Rightarrow f(\mathbf{y}) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle &= \int_0^1 \langle \nabla f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) - \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle dt \\ &= \int_0^1 \frac{1}{t} \langle \nabla f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) - \nabla f(\mathbf{x}), t(\mathbf{y} - \mathbf{x}) \rangle dt \\ (2.8) \qquad &\leq \int_0^1 Lt \|\mathbf{y} - \mathbf{x}\|^2 dt = \frac{L}{2} \|\mathbf{y} - \mathbf{x}\|^2. \end{aligned}$$

□

Theorem 2.11. *Assume $\nabla f(\mathbf{x})$ is Lipschitz-continuous with Lipschitz constant L and $f(\mathbf{x}) : \mathbb{R}^n \longrightarrow \mathbb{R}$ is strongly convex with $\mu > 0$. Then for any $\mathbf{x}, \mathbf{y}$:*

$$\langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \geq \frac{\mu L}{\mu + L} \|\mathbf{x} - \mathbf{y}\|^2 + \frac{1}{\mu + L} \|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|^2.$$

Remark 2.9. *Plug in $\mu = 0$ and compare it with Theorem 2.8.*

Proof. We prove it by discussing two cases.

First, if $\mu = L$, then we need to show

$$\langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \geq \frac{L}{2}\|\mathbf{x} - \mathbf{y}\|^2 + \frac{1}{2L}\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|^2.$$

Theorem 2.8 gives

$$\frac{1}{2L}\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|^2 \leq \frac{1}{2}\langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle$$

and Lemma 1.1 gives

$$\langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \geq \mu\|\mathbf{x} - \mathbf{y}\|^2 \Rightarrow \frac{1}{2}\langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \geq \frac{L}{2}\|\mathbf{x} - \mathbf{y}\|^2.$$

Thus adding two gives the desired inequality.

Second, if $\mu \neq L$, define $\phi(\mathbf{x}) = f(\mathbf{x}) - \frac{\mu}{2}\|\mathbf{x}\|^2$, then $\nabla \phi(\mathbf{x}) = \nabla f(\mathbf{x}) - \mu \mathbf{x}$. So $\phi(\mathbf{x})$ is a convex function, thus

$$0 \leq \langle \nabla \phi(\mathbf{y}) - \nabla \phi(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle = \langle \nabla f(\mathbf{y}) - \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle - \mu\|\mathbf{y} - \mathbf{x}\|^2 \leq (L - \mu)\|\mathbf{y} - \mathbf{x}\|^2.$$

By (2.8), $\nabla \phi$ is Lipschitz continuous with the Lipschitz constant $L - \mu$.

Thus by using (2.7) on $\phi(\mathbf{x})$, we get

$$\begin{aligned}
\langle \nabla \phi(\mathbf{y}) - \nabla \phi(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle &\geq \frac{1}{L - \mu}\|\nabla \phi(\mathbf{y}) - \nabla \phi(\mathbf{x})\|^2 \\
\langle \nabla f(\mathbf{y}) - \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle - \mu\|\mathbf{x} - \mathbf{y}\|^2 &\geq \frac{1}{L - \mu}\|\nabla f(\mathbf{y}) - \nabla f(\mathbf{x}) - \mu(\mathbf{y} - \mathbf{x})\|^2 \\
\langle \nabla f(\mathbf{y}) - \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle - \mu\|\mathbf{x} - \mathbf{y}\|^2 &\geq \frac{1}{L - \mu}\|\nabla f(\mathbf{y}) - \nabla f(\mathbf{x})\|^2 \\
&+ \frac{\mu^2}{L - \mu}\|\mathbf{y} - \mathbf{x}\|^2 + \frac{-2\mu}{L - \mu}\langle \nabla f(\mathbf{y}) - \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \\
\frac{L + \mu}{L - \mu}\langle \nabla f(\mathbf{y}) - \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle &\geq \frac{1}{L - \mu}\|\nabla f(\mathbf{y}) - \nabla f(\mathbf{x})\|^2 + \frac{L\mu}{L - \mu}\|\mathbf{y} - \mathbf{x}\|^2.
\end{aligned}$$

□

Theorem 2.12 (Global linear rate of gradient descent)**.** *Assume $f(\mathbf{x}) : \mathbb{R}^n \longrightarrow \mathbb{R}$ is strongly convex with $\mu > 0$ and $\nabla f(\mathbf{x})$ is Lipschitz-continuous with Lipschitz constant L. Then $f(\mathbf{x})$ has a unique global minimizer: $f(\mathbf{x}) \geq f(\mathbf{x}_*), \forall \mathbf{x}$. The gradient descent method* (2.1) *with a constant step size $\eta \in (0, \frac{2}{L+\mu}]$ satisfies*

$$\|\mathbf{x}_{k+1} - \mathbf{x}_*\|^2 \leq \left(1 - \frac{2\eta\mu L}{L + \mu}\right)^k \|\mathbf{x}_0 - \mathbf{x}_*\|^2.$$

In particular, if $\eta = \frac{2}{L+\mu}$, then we have

$$\|\mathbf{x}_{k+1} - \mathbf{x}_*\| \leq \left(\frac{\frac{L}{\mu} - 1}{\frac{L}{\mu} + 1}\right)^k \|\mathbf{x}_0 - \mathbf{x}_*\|,$$

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}_*) \leq \frac{L}{2}\left(\frac{\frac{L}{\mu} - 1}{\frac{L}{\mu} + 1}\right)^{2k} \|\mathbf{x}_0 - \mathbf{x}_*\|.$$

Remark 2.10. *For any $\eta \in (0, \frac{2}{L+\mu}]$, the convergence rate for the error $\|\mathbf{x}_k - \mathbf{x}_*\|$ has a linear convergence rate $\mathcal{O}(c^k)$ with $c = \sqrt{1 - \frac{2\eta\mu L}{L+\mu}}$ which is a decreasing function of η. The best rate is achieved at $\eta = \frac{2}{L+\mu}$ with $c = \frac{\frac{L}{\mu}-1}{\frac{L}{\mu}+1}$ which is an increasing function of the condition number $\frac{L}{\mu}$. This implies that the best convergence rate will be worse for a larger condition number.*

Proof. Define $r_k = \|\mathbf{x}_k - \mathbf{x}_*\|$. With $\nabla f(\mathbf{x}_*) = \mathbf{0}$ and Theorem 2.11, we get

$$\begin{aligned}
r_{k+1}^2 &= \|\mathbf{x}_{k+1} - \mathbf{x}_*\|^2 \\
&= \|\mathbf{x}_k - \eta\|\nabla f(\mathbf{x}_k) - \mathbf{x}_*\|^2 \\
&= \|\mathbf{x}_k - \mathbf{x}_*\|^2 + \|\eta\nabla f(\mathbf{x}_k)\|^2 + 2\langle \mathbf{x}_k - \mathbf{x}_*, -\eta\nabla f(\mathbf{x}_k)\rangle \\
&= \|\mathbf{x}_k - \mathbf{x}_*\|^2 + \eta^2\|\nabla f(\mathbf{x}_k)\|^2 - 2\eta\langle \mathbf{x}_k - \mathbf{x}_*, \nabla f(\mathbf{x}_k) - \nabla f(\mathbf{x}_*)\rangle \\
&\leq \|\mathbf{x}_k - \mathbf{x}_*\|^2 + \eta^2\|\nabla f(\mathbf{x}_k)\|^2 - 2\eta\frac{\mu}{\mu + L}\|\mathbf{x}_{k+1} - \mathbf{x}_*\|^2 \\
&- 2\eta\frac{1}{L+\mu}\|\nabla f(\mathbf{x}_k) - \nabla f(\mathbf{x}_*)\|^2 \\
&= \left(1 - 2\eta\frac{\mu}{\mu + L}\right) r_k^2 + (\eta^2 - \frac{2}{L+\mu}\eta))\|\nabla f(\mathbf{x}_k)\|^2.
\end{aligned}$$

Thus for any $\eta \in (0, \frac{2}{L+\mu})$,

$$r_{k+1}^2 \leq \left(1 - 2\eta\frac{\mu}{\mu + L}\right) r_k^2.$$

With descent lemma (Lemma 2.1), we get

$$f(\mathbf{x}_k) - f(\mathbf{x}_*) = \langle \nabla f(\mathbf{x}_*), \mathbf{x} - \mathbf{x}_*\rangle + \frac{L}{2}\|\mathbf{x}_k - \mathbf{x}_*\|^2 = \frac{L}{2}r_k^2 \leq \frac{L}{2}\left(1 - 2\eta\frac{\mu}{\mu + L}\right)^{2k} r_0^2.$$

□

2.2.5 Steepest descent

We can consider a variable step size $\eta_k > 0$ in the gradient descent method

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \eta_k \nabla f(\mathbf{x}_k) \tag{2.9a}$$

where η_k can be taken as the best step size in the following sense

$$\eta_k = \arg\min_{\alpha>0} f(\mathbf{x}_k - \alpha \nabla f(\mathbf{x}_k)). \tag{2.9b}$$

Such an optimal step size is also called *full relaxation.* The method (2.9) is often called *the steepest descent*, which is rarely used in practice unless (2.9b) can be easily computed. Nonetheless, analyzing its convergence rate is a starting point for understanding practical algorithms.

Theorem 2.13. *For a twice continuously differentiable function $f : \mathbb{R}^n \to \mathbb{R}$, assume $\mu I \le \nabla^2 f(x) \le LI$ where $L > \mu > 0$ are constants (eigenvalues of Hessian have uniform positive bounds), thus f is strongly convex has a unique minimizer $\mathbf{x}_*$. Then the steepest descent method* (2.9) *satisfies*

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}_*) \le \left(1 - \frac{\mu}{L}\right)^k [f(\mathbf{x}_0) - f(\mathbf{x}_*)].$$

Remark 2.11. *The rate $(1 - \frac{\mu}{L})$ is not sharp and in general we have $\left(\frac{L-\mu}{L+\mu}\right)^2 < 1 - \frac{\mu}{L}$, e.g., the provable fastest rate in Theorem 2.12 for a constant step size η is better than the provable rate of steepest descent.*

Proof. For convenience, let $\mathbf{h}_k = \nabla f(\mathbf{x}_k)$. By Multivariate Quadratic Taylor's Theorem (Theorem 1.4), for any $\alpha > 0$, there exists $\theta \in (0, 1)$ and $\mathbf{z}_k = \mathbf{x}_k + \theta(\mathbf{x}_k - \alpha \mathbf{h}_k)$ such that

$$f(\mathbf{x}_k - \alpha \mathbf{h}_k) = f(\mathbf{x}_k) - \alpha \mathbf{h}_k^T \nabla f(\mathbf{x}_k) + \frac{1}{2}\alpha^2 \mathbf{h}_k^T \nabla^2 f(\mathbf{z}_k)\mathbf{h}_k.$$

The assumption $\nabla^2 f(\mathbf{x}) \le LI, \forall \mathbf{x}$ and the Courant-Fischer-Weyl min-max principle (Appendix A.1) implies

$$f(\mathbf{x}_k - \alpha \mathbf{h}_k) \le f(\mathbf{x}_k) - \alpha \mathbf{h}_k^T \nabla f(\mathbf{x}_k) + \frac{1}{2} L\alpha^2 \|\mathbf{h}_k\|^2.$$

The minimum of the left hand side with respect to α is $f(\mathbf{x}_{k+1})$. The right hand side is a quadratic function of α. The inequality above still holds if minimizing both sides with respect to α:

$$f(\mathbf{x}_{k+1}) = \min_{\alpha} f(\mathbf{x}_k - \alpha \mathbf{h}_k) \le f(\mathbf{x}_k) - \alpha \mathbf{h}_k^T \nabla f(\mathbf{x}_k) + \frac{1}{2} L\alpha^2 \|\mathbf{h}_k\|^2,$$

$$f(\mathbf{x}_{k+1}) \le \min_{\alpha} [f(\mathbf{x}_k) - \alpha \mathbf{h}_k^T \nabla f(\mathbf{x}_k) + \frac{1}{2} L\alpha^2 \|\mathbf{h}_k\|^2] = f(\mathbf{x}_k) - \frac{1}{2L} \|\nabla f(\mathbf{x}_k)\|^2,$$

thus

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}_*) \le f(\mathbf{x}_k) - f(\mathbf{x}_*) - \frac{1}{2L}\|\nabla f(\mathbf{x}_k)\|^2. \tag{2.10}$$

Similarly, by Multivariate Quadratic Taylor's Theorem, and lower bound assumption $\mu I \le \nabla^2 f(\mathbf{x})$ with the Courant-Fischer-Weyl min-max principle (Appendix A.1), we get

$$f(\mathbf{x}) \ge f(\mathbf{x}_k) + \nabla f(\mathbf{x}_k)^T(\mathbf{x} - \mathbf{x}_k) + \frac{\mu}{2}\|\mathbf{x} - \mathbf{x}_k\|^2.$$

Minimizing first the right hand side then the left hand side w.r.t. $\mathbf{x}$, we get

$$f(\mathbf{x}) \ge f(\mathbf{x}_k) - \frac{1}{2\mu}\|\nabla f(\mathbf{x}_k)\|^2,$$

$$f(\mathbf{x}_*) \ge f(\mathbf{x}_k) - \frac{1}{2\mu}\|\nabla f(\mathbf{x}_k)\|^2,$$

thus $-\|\nabla f(\mathbf{x}_k)\|^2 \le 2\mu[f(\mathbf{x}_*) - f(\mathbf{x}_k)]$. Plugging it into (2.10), we get the convergence rate. □

2.2.6 Quadratic functions

The better convergence rate $\left(\frac{L-\mu}{L+\mu}\right)^2$ can be proven for the steep descent method (2.9) for a quadratic function

$$f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T A\mathbf{x} - \mathbf{x}^T\mathbf{b},$$

where A is a positive definite matrix with eigenvalues

$$0 < \lambda_1 \le \lambda_2 \cdots \le \lambda_n.$$

Since $\nabla^2 f(\mathbf{x}) \equiv A \ge \mu I$, $f(\mathbf{x})$ is strongly convex thus has a unique minimizer $\mathbf{x}_*$ satisfying $\nabla f(\mathbf{x}_*) = \mathbf{0} \Leftrightarrow A\mathbf{x}_* = \mathbf{b}$. Define

$$E(\mathbf{x}) = \frac{1}{2}(\mathbf{x} - \mathbf{x}_*)^T A(\mathbf{x} - \mathbf{x}_*).$$

Notice that

$$A\mathbf{x}_* = \mathbf{b} \Rightarrow \frac{1}{2}\mathbf{x}_*^T A\mathbf{x}_* = \frac{1}{2}\mathbf{x}_*^T\mathbf{b} \Rightarrow f(\mathbf{x}_*) = -\frac{1}{2}\mathbf{x}_*^T A\mathbf{x}_*,$$

thus

$$E(\mathbf{x}) = f(\mathbf{x}) + \frac{1}{2}\mathbf{x}_*^T A\mathbf{x}_* = f(\mathbf{x}) - f(\mathbf{x}_*).$$

For convenience, let $\mathbf{h}_k = \nabla f(\mathbf{x}_k) = A\mathbf{x}_k - \mathbf{b}$, then

$$f(\mathbf{x}_k - \eta\mathbf{h}_k) = \frac{1}{2}(\mathbf{x}_k - \eta\mathbf{h}_k)^T A(\mathbf{x}_k - \eta\mathbf{h}_k) - (\mathbf{x}_k - \eta\mathbf{h}_k)^T\mathbf{b}.$$

The quadratic function of η above is minimized at $\eta_k = \frac{\mathbf{h}_k^T \mathbf{h}_k}{\mathbf{h}_k^T A \mathbf{h}_k}$. Thus (2.9) becomes

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \frac{\mathbf{h}_k^T \mathbf{h}_k}{\mathbf{h}_k^T A \mathbf{h}_k} \mathbf{h}_k.$$

So

$$\begin{aligned} E(\mathbf{x}_{k+1}) &= \frac{1}{2}(\mathbf{x}_k - \mathbf{x}_* - \eta_k \mathbf{h}_k)^T A (\mathbf{x}_k - \mathbf{x}_* - \eta_k \mathbf{h}_k) \\ &= E(\mathbf{x}_k) - \eta_k \mathbf{h}_k^T A(\mathbf{x}_k - \mathbf{x}_*) + \frac{1}{2}\eta_k^2 \mathbf{h}_k^T A \mathbf{h}_k, \\ \Rightarrow \frac{E(\mathbf{x}_k) - E(\mathbf{x}_{k+1})}{E(\mathbf{x}_k)} &= \frac{\eta_k \mathbf{h}_k^T A(\mathbf{x}_k - \mathbf{x}_*) - \frac{1}{2}\eta_k^2 \mathbf{h}_k^T A \mathbf{h}_k}{\frac{1}{2}(\mathbf{x}_k - \mathbf{x}_*)^T A (\mathbf{x}_k - \mathbf{x}_*)}. \end{aligned}$$

Notice that $A(\mathbf{x}_k - \mathbf{x}_*) = A\mathbf{x}_k - \mathbf{b} = \mathbf{h}_k$ and $\eta_k = \frac{\mathbf{h}_k^T \mathbf{h}_k}{\mathbf{h}_k^T A \mathbf{h}_k}$, we get

$$\frac{E(\mathbf{x}_k) - E(\mathbf{x}_{k+1})}{E(\mathbf{x}_k)} = \frac{2\eta_k \mathbf{h}_k^T \mathbf{h} - \eta_k^2 \mathbf{h}_k^T A \mathbf{h}_k}{\mathbf{h}^T A^{-1} \mathbf{h}} = \frac{\|\mathbf{h}\|^4}{(\mathbf{h}^T A \mathbf{h})(\mathbf{h}^T A^{-1} \mathbf{h})}.$$

We have proved that

$$E(\mathbf{x}_{k+1}) = \left(1 - \frac{\|\mathbf{h}\|^4}{(\mathbf{h}^T A \mathbf{h})(\mathbf{h}^T A^{-1} \mathbf{h})}\right) E(\mathbf{x}_k),$$

or equivalently

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}_*) = \left(1 - \frac{\|\mathbf{h}\|^4}{(\mathbf{h}^T A \mathbf{h})(\mathbf{h}^T A^{-1} \mathbf{h})}\right) [f(\mathbf{x}_k) - f(\mathbf{x}_*)].$$

By the min-max principle (Theorem A.1), we can only get

$$\frac{\mathbf{h}^T A \mathbf{h}}{\|\mathbf{h}\|^2} \leq \lambda_n, \quad \frac{\mathbf{h}^T A^{-1} \mathbf{h}}{\|\mathbf{h}\|^2} \leq \frac{1}{\lambda_1} \Rightarrow 1 - \frac{\|\mathbf{h}\|^4}{(\mathbf{h}^T A \mathbf{h})(\mathbf{h}^T A^{-1} \mathbf{h})} \leq 1 - \frac{\lambda_1}{\lambda_n},$$

which is the same rate as in Theorem 2.13. In order to get a better rate, we can use the Kantorovich inequality in Theorem A.2:

$$1 - \frac{\|\mathbf{h}\|^4}{(\mathbf{h}^T A \mathbf{h})(\mathbf{h}^T A^{-1} \mathbf{h})} \leq 1 - \frac{4\lambda_1 \lambda_n}{(\lambda_1 + \lambda_n)^2} = \frac{(\lambda_n/\lambda_1 - 1)^2}{(\lambda_n/\lambda_1 + 1)^2}.$$

2.3 Line search method

Now we consider a more general method for minimizing $f(\mathbf{x})$:

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \eta_k \mathbf{p}_k,$$

where $\eta_k > 0$ is a step size and $\mathbf{p}_k \in \mathbb{R}^n$ is a search direction. Examples of the search direction include:

1. *Gradient method* $\mathbf{p}_k = -\nabla f(\mathbf{x}_k)$.
2. *Newton's method* $\mathbf{p}_k = -[\nabla^2 f(\mathbf{x}_k)]^{-1}\nabla f(\mathbf{x}_k)$.
3. *Quasi Newton's method* $\mathbf{p}_k = -B_k\nabla f(\mathbf{x}_k)$, where $B_k \approx [\nabla^2 f(\mathbf{x}_k)]^{-1}$.
4. *Conjugate Gradient Method* $\mathbf{p}_k = -(\mathbf{x}_k - \mathbf{x}_{k-1} + \beta_k\nabla f(\mathbf{x}_k))$, where β_k is designed such that $\mathbf{p}_k$ and $\mathbf{x}_k - \mathbf{x}_{k-1}$ are conjugate (orthogonal in some sense).

The search direction $\mathbf{p}_k$ is a descent direction if $\langle \mathbf{p}_k, -\nabla f(\mathbf{x}_k)\rangle > 0$, i.e., $\mathbf{p}_k$ pointing to the negative gradient direction.

2.3.1 The step size

To find a proper step size η_k, it is natural to ask for a sufficient decrease in the cost function:

$$f(\mathbf{x}_k + \eta_k\mathbf{p}_k) \le f(\mathbf{x}_k) + c_1\eta_k\langle\nabla f(\mathbf{x}_k),\mathbf{p}_k\rangle, \quad c_1 \in (0,1). \tag{2.11a}$$

The constant c_1 is usually taken as a small number such as 10^{-4}, and (2.11a) is called *Amijo condition.* To avoid unacceptably small step sizes, the *curvature condition* requires

$$\langle\nabla f(\mathbf{x}_k + \eta_k\mathbf{p}_k), \mathbf{p}_k\rangle \ge c_2\langle\nabla f(\mathbf{x}_k), \mathbf{p}_k\rangle, \quad c_2 \in (c_1, 1). \tag{2.11b}$$

Define $\phi(\eta) = f(\mathbf{x}_k + \eta\mathbf{p}_k)$, then $\phi'(\eta) = \langle\nabla f(\mathbf{x}_k + \eta\mathbf{p}_k), \mathbf{p}_k\rangle$, thus (2.11b) simply requires $\phi'(\eta_k) \ge c_2\phi'(0)$, where $\phi'(0) = \langle\nabla f(\mathbf{x}_k), \mathbf{p}_k\rangle < 0$ for a descent direction $\mathbf{p}_k$. Usually, c_2 is taken as 0.9 for Newton and quasi-Newton methods, and 0.1 in conjugate gradient methods.

The two conditions in (2.11) with $0 < c_1 < c_2 < 1$ are called the *Wolfe conditions.*

The following are called the *strong Wolfe conditions.*

$$f(\mathbf{x}_k + \eta\mathbf{p}_k) \le f(\mathbf{x}_k) + c_1\eta\langle\nabla f(\mathbf{x}_k),p_k\rangle, \quad c_1 \in (0,1). \tag{2.12a}$$

$$|\langle\nabla f(\mathbf{x}_k + \eta_k\mathbf{p}_k), \mathbf{p}_k\rangle| \le c_2|\langle\nabla f(\mathbf{x}_k), \mathbf{p}_k\rangle|, \quad c_2 \in (c_1, 1). \tag{2.12b}$$

Lemma 2.3. *Assume $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ is continuously differentiable and has a lower bound, and $\mathbf{p}_k$ is a descent direction. Then for any $0 < c_1 < c_2 < 1$, there are intervals of η satisfying the Wolfe conditions* (2.11) *and the strong Wolfe conditions* (2.12)*.*

Proof. The line $\ell(\eta) = f(\mathbf{x}_k) + \eta c_1\langle\nabla f(\mathbf{x}_k), \mathbf{p}_k\rangle$ has a negative slope with $0 < c_1 < 1$. So the line must intersect with the graph of $\phi(\eta) = f(\mathbf{x}_k + \eta\mathbf{p}_k)$ at least once for $\eta > 0$, because $0 > \ell'(0) > \phi'(0)$, $\ell(0) = \phi(0)$ and $\phi(\eta)$ is bounded below for all η.

Let $\eta_1 > 0$ be the smallest such intersection point. Then

$$f(\mathbf{x}_k + \eta_1 \mathbf{p}_k) = f(\mathbf{x}_k) + \eta_1 c_1 \langle \nabla f(\mathbf{x}_k), \mathbf{p}_k \rangle,$$

and (2.11a) holds for any $\eta \in (0, \eta_1)$ because $\eta_1 > 0$ is the smallest intersection point.

By the Mean Value Theorem on $\phi(\eta) = f(\mathbf{x}_k + \eta \mathbf{p}_k)$, there is $\eta_2 \in (0, \eta_1)$ such that

$$f(\mathbf{x}_k + \eta_1 \mathbf{p}_k) - f(\mathbf{x}_k) = \langle \nabla f(\mathbf{x}_k + \eta_2 \mathbf{p}_k), \eta_1 \mathbf{p}_k \rangle.$$

By the two equations above, we have

$$\langle \nabla f(\mathbf{x}_k + \eta_2 \mathbf{p}_k), \mathbf{p}_k \rangle = c_1 \langle \nabla f(\mathbf{x}_k), \mathbf{p}_k \rangle > c_2 \langle \nabla f(\mathbf{x}_k), \mathbf{p}_k \rangle.$$

So η_2 satisfies (2.11b). Since ∇f is continuous, there is a small interval containing η_2, in which η satisfies (2.11b). Notice that the left hand side of the inequality above is negative, thus the strong Wolfe conditions also hold at η_2 and in a small interval containing η_2. □

In practice, the search of a proper step size satisfying the Wolfe conditions can be achieved by backtracking, e.g., use $\eta \leftarrow c\eta$ with $c \in (0, 1)$ until the step size satisfies (2.11).

Example 2.5. *For the gradient descent method $\mathbf{p}_k = -\nabla f(\mathbf{x}_k)$ with a fixed step size $\eta < \frac{2}{L}$, where L is the Lipschitz constant for the gradient $\nabla f(\mathbf{x})$, the descent lemma (Lemma 2.1) and sufficient descrease lemma (Lemma 2.2) gives*

$$f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}) \geq \eta(1 - \frac{L}{2}\eta)\|\nabla f(\mathbf{x})\|^2,$$

i.e.,

$$f(\mathbf{x}_k + \eta \mathbf{p}_k) \leq f(\mathbf{x}_k) + \eta(1 - \frac{L}{2}\eta)\langle \nabla f(\mathbf{x}_k), \mathbf{p}_k \rangle.$$

So $\eta < \frac{2}{L}$ satisfies (2.11a) *with $c_1 = 1 - \frac{L}{2}\eta$.*

If we further assume $f(\mathbf{x})$ is strongly convex with $\mu > 0$. Then Lemma 1.1 gives

$$\langle \nabla f(\mathbf{x}_{k+1}) - \nabla f(\mathbf{x}_k), \mathbf{x}_{k+1} - \mathbf{x}_k \rangle \geq \mu \|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2,$$

thus

$$\langle \nabla f(\mathbf{x}_k + \eta \mathbf{p}_k) - \nabla f(\mathbf{x}_k), -\eta \nabla f(\mathbf{x}_k) \rangle \geq \mu \|\eta \nabla f(\mathbf{x}_k)\|^2.$$

So we get

$$\langle \nabla f(\mathbf{x}_k + \eta \mathbf{p}_k), -\nabla f(\mathbf{x}_k) \rangle \geq (\mu\eta - 1)\|\nabla f(\mathbf{x}_k)\|^2,$$

which can be written as

$$\langle \nabla f(\mathbf{x}_k + \eta_k \mathbf{p}_k), \mathbf{p}_k \rangle \geq c_2 \langle \nabla f(\mathbf{x}_k), \mathbf{p}_k \rangle$$

with $c_2 = 1 - \mu\eta$. By requiring $c_1 < c_2 < 1$. So if assuming $L > 2\mu$, which is usually satisfied in practice, then any stable step size $\eta < \frac{2}{L}$ satisfies the Wolfe condition (2.11)*.*

2.3.2 The convergence

We consider the angle θ_k between the negative gradient and the search direction:

$$\cos\theta_k = \frac{\langle -\nabla f(\mathbf{x}_k), \mathbf{p}_k\rangle}{\|\nabla f(\mathbf{x}_k)\|\|\mathbf{p}_k\|}.$$

Theorem 2.14 (Zoutendijk's Theorem)**.** *Assume $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ is continuously differentiable with Lipschitz continuous gradient $\nabla f(\mathbf{x})$, and $f(\mathbf{x})$ is bounded from below. Consider a line search method $\mathbf{x}_{k+1} = \mathbf{x}_k + \eta_k \mathbf{p}_k$, where $\mathbf{p}_k$ is a descent direction and η_k satisfies the Wolfe conditions* (2.11)*. Then*

$$\sum_{k=1}^{\infty} \cos^2\theta_k \|\nabla f(\mathbf{x}_k)\|^2 < +\infty.$$

Proof. By (2.11b), we have

$$\langle \nabla f(\mathbf{x}_{k+1}) - \nabla f(\mathbf{x}_k), \mathbf{p}_k\rangle \geq (c_2 - 1)\langle \nabla f(\mathbf{x}_k), \mathbf{p}_k\rangle.$$

The Lipschitz continuity and Cauchy Schwartz inequality give

$$\langle \nabla f(\mathbf{x}_{k+1}) - \nabla f(\mathbf{x}_k), \mathbf{p}_k\rangle \leq \|\nabla f(\mathbf{x}_{k+1}) - \nabla f(\mathbf{x}_k)\|\|\mathbf{p}_k\| \leq L\|\eta_k \mathbf{p}_k\|\|\mathbf{p}_k\|.$$

Combining the two inequalities, we get

$$\eta_k \geq \frac{c_2 - 1}{L}\frac{\langle \nabla f(\mathbf{x}_k), \mathbf{p}_k\rangle}{\|\mathbf{p}_k\|^2}.$$

Plugging it into (2.11a), we get

$$f(\mathbf{x}_k + \eta_k \mathbf{p}_k) \leq f(\mathbf{x}_k) - c_1 \frac{1 - c_2}{L}\frac{|\langle \nabla f(\mathbf{x}_k), \mathbf{p}_k\rangle|^2}{\|\mathbf{p}_k\|^2},$$

which can be written as

$$f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k) - \omega \cos^2\theta_k \|\nabla f(\mathbf{x}_k)\|^2, \quad \omega = c_1 \frac{1 - c_2}{L}.$$

Summing it up, since $f(\mathbf{x}) \geq C$, we get

$$\sum_{k=0}^{N} \cos^2\theta_k \|\nabla f(\mathbf{x}_k)\|^2 \leq \frac{1}{\omega}[f(\mathbf{x}_0) - f(\mathbf{x}_{N+1})] \leq \frac{1}{\omega}[f(\mathbf{x}_0) - C].$$

So $a_N = \sum\limits_{k=0}^{N} \cos^2\theta_k \|\nabla f(\mathbf{x}_k)\|^2$ is a bounded and increasing sequence, thus the infinite series converges. □

The convergence of the series in Zoutendijk's Theorem gives $\cos^2\theta_k \|\nabla f(\mathbf{x}_k)\| \to 0$. Thus if $\cos^2\theta_k \geq \delta > 0, \forall k$, then $\|\nabla f(\mathbf{x}_k)\| \to 0$.

Example 2.6. *Consider Newton's method with* $\mathbf{p}_k = -[\nabla^2 f(\mathbf{x}_k)]^{-1}\nabla f(\mathbf{x}_k)$. *Assume the Hessian has some uniform positive bounds for eigenvalues (i.e., the Hessian is* **positive definite** *with a uniformly bounded condition number:):*

$$\mu I \le \nabla^2 f(\mathbf{x}) \le LI, \quad L \ge \mu > 0, \forall \mathbf{x},$$

then we have (eigenvalues of A are reciprocals of eigenvalues of A^{-1})

$$\frac{1}{L} I \le [\nabla^2 f(\mathbf{x})]^{-1} \le \frac{1}{\mu} I, \quad L \ge \mu > 0, \forall \mathbf{x}.$$

For convenience, let $B_k = [\nabla^2 f(\mathbf{x})]^{-1}$ and $\mathbf{h}_k = \nabla f(\mathbf{x}_k)$. Since B_k is positive definite, its eigenvalues are also singular values. By the definition of spectral norm, we get

$$\|\mathbf{p}_k\| = \|B_k \nabla f(\mathbf{x}_k)\| \le \|B_k\| \|\nabla f(\mathbf{x}_k)\| \le \frac{1}{\mu}\|\nabla f(\mathbf{x}_k)\| = \frac{1}{\mu}\|\mathbf{h}_k\|.$$

By the Courant-Fischer-Weyl min-max principle (Appendix A.1), we have

$$\cos\theta_k = \frac{\langle -\nabla f(\mathbf{x}_k), \mathbf{p}_k\rangle}{\|\nabla f(\mathbf{x}_k)\|\|\mathbf{p}_k\|} = \frac{\mathbf{h}_k^T B_k \mathbf{h}_k}{\|\mathbf{h}_k\|\|\mathbf{p}_k\|} \ge \mu \frac{\mathbf{h}_k^T B_k \mathbf{h}_k}{\|\mathbf{h}_k\|\|\mathbf{h}_k\|} \ge \frac{\mu}{L} = \frac{1}{L/\mu},$$

where $L/\mu = \|B_k\|\|B_k^{-1}\|$ is the condition number of the Hessian. With Theorem 2.14, we get $\|\nabla f(\mathbf{x}_k)\| \to 0$. Recall that a strongly convex function has a unique critical point which is the global minimizer. So the Newton's method with a step size satisfying the Wolfe conditions (2.11) *converges to the unique minimizer $\mathbf{x}_*$ for a strongly convex function $f(\mathbf{x})$ if $\|\nabla^2 f(\mathbf{x})\|$ has a uniform upper bound, see the problem below.*

Problem 2.1. *Recall that $\|\nabla f(\mathbf{x}_k)\| \to 0$ may not even imply $\mathbf{x}_k$ converges to a critical point, see Example 2.2. Prove that $\|\nabla f(\mathbf{x}_k)\| \to 0$ implies $\mathbf{x}_k$ converges to the global minimizer under the assumption*

$$\mu I \le \nabla^2 f(\mathbf{x}) \le LI, \quad L \ge \mu > 0, \forall \mathbf{x}.$$

2.4 Local convergence rate

So far we have only discussed the global convergence, e.g., the convergence for arbitrary initial guess $\mathbf{x}_0$ in an iterative method. If the initial guess is very close to a minimizer, we can discuss the *local convergence.*

We will make the following assumptions:

1. The Hessian exists and is Lipschitiz continuous with parameter $M > 0$:

$$\|\nabla^2 f(\mathbf{x}) - \nabla^2 f(\mathbf{y})\| \le M\|\mathbf{x} - \mathbf{y}\|, \quad \forall \mathbf{x}, \mathbf{y},$$

where the left hand side is the matrix spectral norm.

2. There exists a **local minimum** $\mathbf{x}_*$, and the Hessian $\nabla^2 f(\mathbf{x}^*)$ is positive definite:
$$\mu I \leq \nabla^2 f(\mathbf{x}^*) \leq LI, \quad L \geq \mu > 0.$$
Notice that this does not imply the function is strongly convex.

2.4.1 Gradient descent

Consider the gradient descent method

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \eta \nabla f(\mathbf{x}_k).$$

By Fundamental Theorem of Calculus on the single variable vector-valued function $\mathbf{g(t)} = \nabla \mathbf{f}(\mathbf{x}_* + \mathbf{t}(\mathbf{x_k} - \mathbf{x}_*))$, we get

$$\nabla f(\mathbf{x}_k) = \nabla f(\mathbf{x}_*) - \nabla f(\mathbf{x}_*) = \int_0^1 \nabla^2 f(\mathbf{x}_* + t(\mathbf{x}_k - \mathbf{x}_*))(\mathbf{x}_k - \mathbf{x}_*) dt = G(\mathbf{x}_k - \mathbf{x}_*),$$

where
$$G_k = \int_0^1 \nabla^2 f(\mathbf{x}_* + t(\mathbf{x}_k - \mathbf{x}_*)) dt.$$

Then

$$\mathbf{x}_{k+1} - \mathbf{x}_* = \mathbf{x}_k - \mathbf{x}_* - \eta G_k(\mathbf{x}_k - \mathbf{x}_*) = (I - \eta G_k)(\mathbf{x}_k - \mathbf{x}_*)$$
$$\Rightarrow \|\mathbf{x}_{k+1} - \mathbf{x}_*\| \leq \|I - \eta G_k\| \|\mathbf{x}_k - \mathbf{x}_*\|.$$

Lemma 2.4. *If $\nabla^2 f(\mathbf{x})$ is Lipschitiz continuous with parameter $M > 0$ and $\|\mathbf{x} - \mathbf{y}\| = r$, then*

$$\nabla^2 f(\mathbf{x}) - MrI \leq \nabla^2 f(\mathbf{y}) \leq \nabla^2 f(\mathbf{x}) + MrI.$$

Proof. Let $H = \nabla^2 f(\mathbf{y}) - \nabla^2 f(\mathbf{x})$. Since H is real symmetric, its singular values are absolute values of its eigenvalues, Lipschitiz continuity gives $\|H\| \leq M\|\mathbf{x} - \mathbf{y}\| = Mr \Rightarrow |\lambda_i(H)| \leq Mr$, where $\lambda_i(H)$ denotes the eigenvalue. So $\lambda_i(H) - Mr \leq 0$ and $Mr - \lambda_i(H) \geq 0$. □

Theorem 2.15 (Local linear rate of gradient descent)**.** *Let $f(\mathbf{x})$ satisfy the assumptions in this section. Let $\mathbf{x}_0$ be close enough to a strict local minimizer $\mathbf{x}_*$:*
$$r_0 = \|\mathbf{x}_0 - \mathbf{x}_*\| < \bar{r} = \frac{2\mu}{M}.$$
Then the gradient descent method with a fixed step size $0 < \eta < \frac{2}{L+\mu}$ satisfies

$$\|\mathbf{x}_{k+1} - \mathbf{x}_*\| \leq c_k \|\mathbf{x}_k - \mathbf{x}_*\|,$$

where

$$c_k = \max\{|1 - \eta(\mu - \frac{1}{2}M\|\mathbf{x}_k - \mathbf{x}_*\|)|, |1 - \eta(L + \frac{1}{2}M\|\mathbf{x}_k - \mathbf{x}_*\|)|\} < 1.$$

In particular, if $\eta = \frac{2}{L+\mu}$,

$$\|\mathbf{x}_{k+1} - \mathbf{x}_*\| \leq \frac{\bar{r} r_0}{\bar{r} - r_0} \left(1 - \frac{2\mu}{L + 3\mu}\right)^k \|\mathbf{x}_0 - \mathbf{x}_*\|.$$

Remark 2.12. *The numbers* μ *and* L *in this local convergence rate theorem are eigenvalues bounds of the Hessian at only* $\mathbf{x}_*$, *rather than uniform bounds for the Hessian at all* $\mathbf{x}$.

Proof. Let $r_k = \|\mathbf{x}_k - \mathbf{x}_*\|$, by the lemma above, we have

$$\nabla^2 f(\mathbf{x}_*) - tMr_k I \leq \nabla^2 f(\mathbf{x}_* + t(\mathbf{x}_k - \mathbf{x}_*)) \leq \nabla^2 f(\mathbf{x}_*) + tMr_k I$$

thus

$$(\mu - tMr_k)I \leq \nabla^2 f(\mathbf{x}_* + t(\mathbf{x}_k - \mathbf{x}_*)) \leq (L + tMr_k)I.$$

Notice that the inequalities still hold after integration. For instance,

$$(\mu - tMr_k)I \leq \nabla^2 f(\mathbf{x}_* + t(\mathbf{x}_k - \mathbf{x}_*)) \Leftrightarrow \nabla^2 f(\mathbf{x}_* + t(\mathbf{x}_k - \mathbf{x}_*)) - (\mu - tMr_k)I \geq 0,$$

and

$$\int_0^1 [\nabla^2 f(\mathbf{x}_* + t(\mathbf{x}_k - \mathbf{x}_*)) - (\mu - tMr_k)I] dt \geq 0$$

because

$$\forall \mathbf{z}, \quad \mathbf{z}^T [\nabla^2 f(\mathbf{x}_* + t(\mathbf{x}_k - \mathbf{x}_*)) - (\mu - tMr_k)I] \mathbf{z} \geq 0$$

$$\Rightarrow \mathbf{z}^T \int_0^1 [\nabla^2 f(\mathbf{x}_* + t(\mathbf{x}_k - \mathbf{x}_*)) - (\mu - tMr_k)I] dt \mathbf{z} \geq 0.$$

So after integration we get

$$(\mu - \frac{1}{2}Mr_k)I \leq G_k \leq (L + \frac{1}{2}Mr_k)I,$$

$$[1 - \eta(L + \frac{1}{2}Mr_k)]I \leq I - \eta G_k \leq [1 - \eta(\mu - \frac{1}{2}Mr_k)]I.$$

So

$$\|I - \eta G_k\| \leq \max\{|a_k(\eta)|, |b_k(\eta)|\}$$

where

$$a_k(\eta) = 1 - \eta(\mu - \frac{1}{2}Mr_k), \quad b_k(\eta) = 1 - \eta(L + \frac{1}{2}Mr_k).$$

Notice that $a_k(0) = 1$ and $a_k'(\eta) = -(\mu - \frac{1}{2}Mr_k) < 0$, if assuming $r_k < \frac{2\mu}{M}$. And $b_k(0) = 1$ and $b_k'(\eta) = -(L + \frac{1}{2}Mr_k) < 0$. For small enough η, $\|I - \eta G_k\| < 1$, which can ensure $r_{k+1} < r_k$ since $\|\mathbf{x}_{k+1} - \mathbf{x}_*\| \leq \|I - \eta G_k\| \|\mathbf{x}_k - \mathbf{x}_*\|$.

In particular, under the assumption $r_k < \bar{r}$, it is straightforward to check that

$$\eta < \frac{2}{\mu} \Rightarrow |a_k(\eta)| < 1,$$

$$\eta \leq \frac{2}{L+\mu} \Rightarrow |b_k(\eta)| < 1.$$

Now set $\eta = \frac{2}{L+\mu}$, then $b_k(\eta) < 0$ and $a_k(\eta) > 0$. In this case, with $\eta = \frac{2}{L+\mu}$ it is straightforward to check that

$$|a_k(\eta)| = |b_k(\eta)| = \frac{L-\mu}{L+\mu} + \eta\frac{1}{2}Mr_k.$$

Therefore, $r_{k+1} \leq \|I - \eta G_k\| r_k$ gives

$$r_{k+1} \leq \frac{L-\mu}{L+\mu} r_k + \frac{M}{L+\mu} r_k^2.$$

Let $a_k = \frac{M}{L+\mu} r_k$ and $q = \frac{2\mu}{L+\mu} < 1$, then it is equivalent to

$$a_{k+1} \leq (1-q)a_k + a_k^2 = a_k[1+(a_k-q)] = a_k\frac{1-(a_k-q)^2}{1-(a_k-q)} \leq a_k\frac{1}{1-(a_k-q)} = \frac{a_k}{1+q-a_k}.$$

$$\Rightarrow \frac{1}{a_{k+1}} \geq \frac{1+q}{a+k} - 1 \Rightarrow \frac{q}{a_{k+1}} - 1 \geq \frac{q(1+q)}{a_k} - q - 1 = (1+q)(\frac{q}{a_k} - 1).$$

So we get

$$\frac{q}{a_{k+1}} - 1 \geq (1+q)^k(\frac{q}{a_0} - 1) = (1+q)^k(\frac{\bar{r}}{r_0} - 1),$$

thus

$$a_k \leq \frac{qr_0}{r_0 + (1+q)^k(\bar{r} - r_0)} \leq \frac{qr_0}{\bar{r} - r_0}\frac{1}{(1+q)^k}.$$

□

2.4.2 Newton's method

Newton's method is the most well-known method to approximately solve a nonlinear equation $F(\mathbf{x}) = \mathbf{0}$ where $F : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ is a smooth function:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \nabla F(\mathbf{x}_k)^{-1} F(\mathbf{x}_k),$$

where ∇F is the Jacobian matrix.

The Babylonian method for finding square roots, especially the root of 2, has been known since the ancient Babylon period around the 17th century BC. It is preciously Newton's method applying to the function $F(x) = x^2 - 2$:

$$x_{k+1} = x_k - F(x_k)/F'(x_k) = x_k - (x_k^2 - 2)/(2x_k) = x_k/2 + 1/x_k.$$

If $x_0 = 1$, then $x_3 = 1.41421568627$ and $|x_3 - \sqrt{2}| = 2.12E - 6$.

When applying the Newton's method to $\nabla f(\mathbf{x}) = 0$ for finding minimizers of $f(\mathbf{x})$, we obtain the Newton's method for finding critical points:

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

Another way to derive the simplest Newton's method is to consider a quadratic function:

$$\phi(\mathbf{x}) = f(\mathbf{x}_k) + (\mathbf{x} - \mathbf{x}_k)^T \nabla f(\mathbf{x}_k) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_k)^T \nabla^2 f(\mathbf{x}_k)(\mathbf{x} - \mathbf{x}_k).$$

Assume the Hessian is positive definite, define $\mathbf{x}_{k+1}$ as the minimizer of $\phi(\mathbf{x})$. Then

$$\mathbf{0} = \nabla \phi(\mathbf{x}_{k+1}) = \nabla f(\mathbf{x}_k) + \nabla^2 f(\mathbf{x}_k)(\mathbf{x}_{k+1} - \mathbf{x}_k)$$

gives the Newton's method.

Theorem 2.16 (Local quadratic rate of Newton's method)**.** *Let $f(\mathbf{x})$ satisfy the assumptions in this section. Let $\mathbf{x}_0$ be close enough to a strict local minimizer $\mathbf{x}_*$:*

$$r_0 = \|\mathbf{x}_0 - \mathbf{x}_*\| < \bar{r} = \frac{2\mu}{3M}.$$

Then $r_k = \|\mathbf{x}_k - \mathbf{x}_*\| < \bar{r}$, and Newton's method converges quadratically,

$$\|\mathbf{x}_{k+1} - \mathbf{x}_*\| \le \frac{M\|\mathbf{x}_k - \mathbf{x}_*\|^2}{2(\mu - M\|\mathbf{x}_k - \mathbf{x}_*\|)} \le \frac{3M}{2\mu}\|\mathbf{x}_k - \mathbf{x}_*\|^2.$$

Proof.

$$\begin{aligned}
\mathbf{x}_{k+1} - \mathbf{x}_* &= \mathbf{x}_k - \mathbf{x}_* - [\nabla^2 f(\mathbf{x}_k)]^{-1}[\nabla f(\mathbf{x}_k) - \nabla f(\mathbf{x}_*)] \\
&= \mathbf{x}_k - \mathbf{x}_* - [\nabla^2 f(\mathbf{x}_k)]^{-1} \int_0^1 \nabla f^2(\mathbf{x}_* + t(\mathbf{x}_k - \mathbf{x}_*))(\mathbf{x}_k - \mathbf{x}_*) dt \\
&= [\nabla^2 f(\mathbf{x}_k)]^{-1} G_k (\mathbf{x}_k - \mathbf{x}_*)
\end{aligned}$$

where

$$G_k = \int_0^1 [\nabla^2 f(\mathbf{x}_k) - \nabla f^2(\mathbf{x}_* + t(\mathbf{x}_k - \mathbf{x}_*)] dt.$$

By Theorem 1.7 and Lipschitz continuity of the Hessian,

$$\begin{aligned}
\|G_k\| &\le \int_0^1 \|\nabla^2 f(\mathbf{x}_k) - \nabla f^2(\mathbf{x}_* + t(\mathbf{x}_k - \mathbf{x}_*)\| dt \\
&\le \int_0^1 M(1-t)\|\mathbf{x}_k - \mathbf{x}_*\| dt \\
&= \frac{1}{2} r_k M.
\end{aligned}$$

With Lemma 2.4, We also have

$$\nabla f^2(\mathbf{x}_k) \geq \nabla f^2(\mathbf{x}_*) - Mr_k I \geq (\mu - Mr_k)I.$$

So if $r_k < \frac{\mu}{M}$, $\nabla f^2(\mathbf{x}_k) > 0$ and

$$\|[\nabla f^2(\mathbf{x}_k)]^{-1}\| \leq (\mu - Mr_k)^{-1}.$$

Thus if $r_k < \frac{2\mu}{3M}$, we get

$$r_{k+1} \leq \|[\nabla f^2(\mathbf{x}_k)]^{-1}\|\|G_k(\mathbf{x}_k - \mathbf{x}_*)\| \leq \|[\nabla f^2(\mathbf{x}_k)]^{-1}\|\|G_k\| r_k \leq \frac{Mr_k^2}{2(\mu - Mr_k)} \leq r_k.$$

□

2.5 Accelerated gradient method

The accelerated gradient descent method is a very popular class of first order methods for large scale minimization problems. The original accelerated gradient method [3] proposed by Nesterov in 1983 takes the following form:

$$\boxed{\begin{cases} \mathbf{x}_{k+1} &= \mathbf{y}_k - \eta_k \nabla f(\mathbf{y}_k) \\ t_{k+1} &= \frac{1}{2}\left(1 + \sqrt{4t_k^2 + 1}\right) \\ \mathbf{y}_{k+1} &= \mathbf{x}_{k+1} + \frac{t_k - 1}{t_{k+1}}(\mathbf{x}_{k+1} - \mathbf{x}_k) \end{cases} \qquad \mathbf{x}_0 = \mathbf{y}_0, t_0 = 1.}$$

For convenience, we can take $\eta_k = \frac{1}{L}$ where L is Lipschitz constant of the gradient $\nabla f(\mathbf{x})$, and use a slightly different $t_{k+1} = \frac{k+2}{2}$, then we have a slightly different version of Nesterov's accelerated gradient method:

$$\boxed{\begin{cases} \mathbf{x}_{k+1} &= \mathbf{y}_k - \frac{1}{L}\nabla f(\mathbf{y}_k) \\ \mathbf{y}_{k+1} &= \mathbf{x}_{k+1} + \frac{k-1}{k+2}(\mathbf{x}_{k+1} - \mathbf{x}_k) \end{cases} \qquad \mathbf{x}_0 = \mathbf{y}_0.}$$

This method requires only one evaluation of the gradient per iteration, yet a global $\mathcal{O}(\frac{1}{k^2})$ convergence rate can be proven for a convex function $f(\mathbf{x})$ with a Lipschtitz continuous gradient. Recall that the gradient descent method has a global $\mathcal{O}(\frac{1}{k})$ convergence rate for the same function as proven in Theorem 2.9.

However, the provable rate $O(\frac{1}{k})$ or $O(\frac{1}{k^2})$ usually represents the worst case scenario of all iterates in an iterative algorithm. The worst case may or may not happen in practice. Thus the accelerated gradient method is not necessarily faster than the gradient descent method for a given convex functions $f(\mathbf{x})$ with Lipschtitz continuous gradient, even though it is indeed better in many applications.

Recall that we get the stable step size $\eta \in (0, \frac{2}{L}]$ for the gradient descent method by requiring cost function to decrease in each iteration $f(\mathbf{x}_{k+1}) < f(\mathbf{x}_k)$. But in the accelerated gradient method, there is no monotonicity guarantee on the sequences $\{f(\mathbf{x}_k)\}$ and $\{f(\mathbf{y}_k)\}$.

2.5.1 Convergence rate

To prove the convergence rate $\mathcal{O}(\frac{1}{k^2})$ and also to see how the sequence t_k and step sizes η_k should be chosen, we consider the following method for a convex function $f(\mathbf{x})$ with Lipschitz continuous gradient (with Lipschitz constant L):

$$\boxed{\begin{cases} \mathbf{x}_{k+1} &= \mathbf{y}_k - \eta_k \nabla f(\mathbf{y}_k) \\ \mathbf{y}_{k+1} &= \mathbf{x}_{k+1} + \frac{t_k - 1}{t_{k+1}}(\mathbf{x}_{k+1} - \mathbf{x}_k) \end{cases} \qquad \mathbf{x}_0 = \mathbf{y}_0.}$$

Apply the descent lemme (Lemma 2.1) to $\mathbf{y} = \mathbf{x}_{k+1}$ and $\mathbf{x} = \mathbf{y}_k$:

$$f(\mathbf{x}_{k+1}) \leq f(\mathbf{y}_k) + \langle \nabla f(\mathbf{y}_k), \mathbf{x}_{k+1} - \mathbf{y}_k \rangle + \frac{L}{2}\|\mathbf{y}_k - \mathbf{x}_{k+1}\|^2. \tag{2.13}$$

The convexity implies

$$f(\mathbf{x}) \geq f(\mathbf{y}) + \langle \mathbf{x} - \mathbf{y}, \nabla f(\mathbf{y}) \rangle,$$

thus

$$f(\mathbf{x}_k) \geq f(\mathbf{y}_k) + \langle \mathbf{x}_k - \mathbf{y}_k, \nabla f(\mathbf{y}_k) \rangle.$$

Subtracting two inequalities, we get

$$\begin{aligned} f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}) &\geq -\frac{L}{2}\|\mathbf{y}_k - \mathbf{x}_{k+1}\|^2 + \langle \mathbf{x}_k - \mathbf{x}_{k+1}, \nabla f(\mathbf{y}_k) \rangle \\ &= -\frac{L}{2}\|\mathbf{y}_k - \mathbf{x}_{k+1}\|^2 + \langle \mathbf{x}_k - \mathbf{x}_{k+1}, \frac{1}{\eta_k}(\mathbf{y}_k - \mathbf{x}_{k+1}) \rangle \\ &= -\frac{L}{2}\|\mathbf{y}_k - \mathbf{x}_{k+1}\|^2 + \langle \mathbf{x}_k - \mathbf{y}_k + \mathbf{y}_k - \mathbf{x}_{k+1}, \frac{1}{\eta_k}(\mathbf{y}_k - \mathbf{x}_{k+1}) \rangle \\ &= (\frac{1}{\eta_k} - \frac{L}{2})\|\mathbf{y}_k - \mathbf{x}_{k+1}\|^2 + \frac{1}{\eta_k}\langle \mathbf{y}_k - \mathbf{x}_{k+1}, \mathbf{x}_k - \mathbf{y}_k \rangle. \end{aligned}$$

Thus

$$\eta_k[f(\mathbf{x}_k) - f(\mathbf{x}_{k+1})] \geq (1 - \eta_k \frac{L}{2})\|\mathbf{y}_k - \mathbf{x}_{k+1}\|^2 + \langle \mathbf{y}_k - \mathbf{x}_{k+1}, \mathbf{x}_k - \mathbf{y}_k \rangle.$$

Similarly, convexity implies

$$f(\mathbf{x}_*) \geq f(\mathbf{y}_k) + \langle \mathbf{x}_* - \mathbf{y}_k, \nabla f(\mathbf{y}_k) \rangle.$$

Subtract it with (2.13), we get

$$\begin{aligned}
f(\mathbf{x}_*) - f(\mathbf{x}_{k+1}) &\geq -\frac{L}{2}\|\mathbf{y}_k - \mathbf{x}_{k+1}\|^2 + \langle \mathbf{x}_* - \mathbf{x}_{k+1}, \nabla f(\mathbf{y}_k)\rangle \\
&= -\frac{L}{2}\|\mathbf{y}_k - \mathbf{x}_{k+1}\|^2 + \langle \mathbf{x}_* - \mathbf{x}_{k+1}, \frac{1}{\eta_k}(\mathbf{y}_k - \mathbf{x}_{k+1})\rangle \\
&= -\frac{L}{2}\|\mathbf{y}_k - \mathbf{x}_{k+1}\|^2 + \langle \mathbf{x}_* - \mathbf{y}_k + \mathbf{y}_k - \mathbf{x}_{k+1}, \frac{1}{\eta_k}(\mathbf{y}_k - \mathbf{x}_{k+1})\rangle \\
&= (\frac{1}{\eta_k} - \frac{L}{2})\|\mathbf{y}_k - \mathbf{x}_{k+1}\|^2 + \frac{1}{\eta_k}\langle \mathbf{y}_k - \mathbf{x}_{k+1}, \mathbf{x}_* - \mathbf{y}_k\rangle.
\end{aligned}$$

Now assume $\eta_k = \frac{1}{L}$, then we have

$$\begin{aligned}
f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}) &\geq \frac{L}{2}\|\mathbf{y}_k - \mathbf{x}_{k+1}\|^2 + L\langle \mathbf{y}_k - \mathbf{x}_{k+1}, \mathbf{x}_k - \mathbf{y}_k\rangle, \\
f(\mathbf{x}_*) - f(\mathbf{x}_{k+1}) &\geq \frac{L}{2}\|\mathbf{y}_k - \mathbf{x}_{k+1}\|^2 + L\langle \mathbf{y}_k - \mathbf{x}_{k+1}, \mathbf{x}_* - \mathbf{y}_k\rangle.
\end{aligned}$$

Next, let $R_k = f(\mathbf{x}_k) - f(\mathbf{x}_*)$ where $\mathbf{x}_*$ is a global minimizer. Then multiplying the first inequality by $t_k - 1$ and add it the second one, we get

$$(t_k - 1)R_k - t_k R_{k+1} \geq \frac{L}{2}t_k\|\mathbf{y}_k - \mathbf{x}_{k+1}\|^2 + L\langle \mathbf{y}_k - \mathbf{x}_{k+1}, (t_k - 1)\mathbf{x}_k - t_k\mathbf{y}_k - \mathbf{x}_*\rangle.$$

Multiply it by t_k:

$$t_k(t_k - 1)R_k - t_k^2 R_{k+1} \geq \frac{L}{2}\|t_k(\mathbf{y}_k - \mathbf{x}_{k+1})\|^2 + L\langle t_k(\mathbf{y}_k - \mathbf{x}_{k+1}), (t_k - 1)\mathbf{x}_k - t_k\mathbf{y}_k - \mathbf{x}_*\rangle. \tag{2.14}$$

Assume we have

$$t_{k+1}^2 - t_{k+1} \leq t_k^2,$$

then

$$t_{k-1}^2 R_k - t_k^2 R_{k+1} \geq \frac{L}{2}\|t_k(\mathbf{y}_k - \mathbf{x}_{k+1})\|^2 + L\langle t_k(\mathbf{y}_k - \mathbf{x}_{k+1}), (t_k - 1)\mathbf{x}_k - t_k\mathbf{y}_k - \mathbf{x}_*\rangle. \tag{2.15}$$

For the right hand side dot product, let

$$\mathbf{a} = t_k\mathbf{y}_k, \quad \mathbf{b} = t_k\mathbf{x}_{k+1}, \quad \mathbf{c} = (t_k - 1)\mathbf{x}_k + \mathbf{x}_*,$$

then the right hand side can be written as

$$\frac{L}{2}\left(\|\mathbf{a} - \mathbf{b}\|^2 + 2\langle \mathbf{c} - \mathbf{a}, \mathbf{a} - \mathbf{b}\rangle\right) = \frac{L}{2}\left(\|\mathbf{b} - \mathbf{c}\|^2 - \|\mathbf{a} - \mathbf{c}\|^2.\right)$$

It can be written as

$$t_{k-1}^2 R_k - t_k^2 R_{k+1} \geq \frac{L}{2}\left(\|t_k\mathbf{x}_{k+1} - [(t_k - 1)\mathbf{x}_k + \mathbf{x}_*]\|^2 - \|t_k\mathbf{y}_k - [(t_k - 1)\mathbf{x}_k + \mathbf{x}_*]\|^2\right).$$

Let $\mathbf{u}_{k+1} = t_k\mathbf{x}_{k+1} - [(t_k - 1)\mathbf{x}_k + \mathbf{x}_*]$, then with

$$\mathbf{y}_{k+1} = \mathbf{x}_{k+1} + \frac{t_k - 1}{t_{k+1}}(\mathbf{x}_{k+1} - \mathbf{x}_k) \Rightarrow t_{k+1}\mathbf{x}_{k+1} + (t_k - 1)\mathbf{x}_k = t_{k+1}\mathbf{y}_{k+1},$$

we get

$$t_k\mathbf{y}_k - [(t_k - 1)\mathbf{x}_k + \mathbf{x}_*] = t_{k-1}\mathbf{x}_k - [(t_{k-1} - 1)\mathbf{x}_{k-1} + \mathbf{x}_*] = \mathbf{u}_k.$$

So

$$t_{k-1}^2 R_k - t_k^2 R_{k+1} \geq \frac{L}{2}(\|\mathbf{u}_{k+1}\|^2 - \|\mathbf{u}_k\|^2)$$

thus

$$t_k^2 R_{k+1} + \frac{L}{2}\|\mathbf{u}_{k+1}\|^2 \leq t_{k-1}^2 R_k + \frac{L}{2}\|\mathbf{u}_k\|^2,$$

which implies

$$t_k^2 R_{k+1} \leq t_k^2 R_{k+1} + \frac{L}{2}\|\mathbf{u}_{k+1}\|^2 \leq t_0^2 R_1 + \frac{L}{2}\|\mathbf{u}_1\|^2,$$

and

$$R_{k+1} \leq \frac{1}{t_k^2}[t_0^2 R_1 + \frac{L}{2}\|\mathbf{u}_1\|^2].$$

So in order to obtain $\mathcal{O}(\frac{1}{k^2})$, we should use t_k satisfying $t_k = \mathcal{O}(k)$. For instance, assume $t_k^2 - t_k = t_{k-1}^2$ with $t_0 = 1$, then we can easily show $t_k \geq \frac{k+1}{2}$ by induction.

All the discussions can be summarized as:

Theorem 2.17. *Assume the function $f(\mathbf{x}) : \mathbb{R}^n \longrightarrow \mathbb{R}$ is convex with a global minimizer $\mathbf{x}_*$. Assume $\nabla f(\mathbf{x})$ is Lipschitz continuous with constant L. Assume $t_k^2 - t_k = t_{k-1}^2$ with $t_0 = 1$. Then the following accelerated gradient method*

$$\boxed{\begin{cases} \mathbf{x}_{k+1} &= \mathbf{y}_k - \frac{1}{L}\nabla f(\mathbf{y}_k) \\ \mathbf{y}_{k+1} &= \mathbf{x}_{k+1} + \frac{t_k-1}{t_{k+1}}(\mathbf{x}_{k+1} - \mathbf{x}_k) \end{cases} \qquad \mathbf{x}_0 = \mathbf{y}_0,}$$

satisfies

$$f(\mathbf{x}_k) - f(\mathbf{x}_*) \leq \frac{4}{k^2}\left(f(\mathbf{x}_1) - f(\mathbf{x}_*) + \frac{L}{2}\|\mathbf{x}_1 - \mathbf{x}_*\|^2\right).$$

Remark 2.13. *Obviously the theorem still holds if we plug in $t_k = \frac{k+1}{2}$, then the algorithm is simplied to*

$$\boxed{\begin{cases} \mathbf{x}_{k+1} &= \mathbf{y}_k - \frac{1}{L}\nabla f(\mathbf{y}_k) \\ \mathbf{y}_{k+1} &= \mathbf{x}_{k+1} + \frac{k-1}{k+2}(\mathbf{x}_{k+1} - \mathbf{x}_k) \end{cases} \qquad \mathbf{x}_0 = \mathbf{y}_0.}$$

To consider a variable step size, now assume $\eta_k = \frac{1}{b_k}\frac{1}{L} \leq \frac{1}{(a+\frac{1}{2})}\frac{1}{L}$ with $a > 0$, then

$$\eta_k - \frac{L}{2} \geq aL, \quad \frac{1}{\eta_k} = b_k L, \quad b_k \geq a + \frac{1}{2}$$

we have

$$\begin{aligned} f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}) &\geq aL\|\mathbf{y}_k - \mathbf{x}_{k+1}\|^2 + b_k L\langle \mathbf{y}_k - \mathbf{x}_{k+1}, \mathbf{x}_k - \mathbf{y}_k\rangle, \\ f(\mathbf{x}_*) - f(\mathbf{x}_{k+1}) &\geq aL\|\mathbf{y}_k - \mathbf{x}_{k+1}\|^2 + b_k L\langle \mathbf{y}_k - \mathbf{x}_{k+1}, \mathbf{x}_* - \mathbf{y}_k\rangle. \end{aligned}$$

Multiplying the first one by $(t_k - 1)$ and add it to the second one, we get

$$(t_k-1)R_k - t_k R_{k+1} \geq aLt_k\|\mathbf{y}_k - \mathbf{x}_{k+1}\|^2 + b_k L\langle \mathbf{y}_k - \mathbf{x}_{k+1}, (t_k-1)\mathbf{x}_k - t_k\mathbf{y}_k - \mathbf{x}_*\rangle.$$

Multiply it by t_k:

$$t_k(t_k-1)R_k - t_k^2 R_{k+1} \geq aL\|t_k(\mathbf{y}_k - \mathbf{x}_{k+1})\|^2 + b_k L\langle t_k(\mathbf{y}_k - \mathbf{x}_{k+1}), (t_k-1)\mathbf{x}_k - t_k\mathbf{y}_k - \mathbf{x}_*\rangle. \tag{2.16}$$

Assume we have

$$t_{k+1}^2 - t_{k+1} \leq t_k^2,$$

then

$$t_{k-1}^2 R_k - t_k^2 R_{k+1} \geq aL\|t_k(\mathbf{y}_k - \mathbf{x}_{k+1})\|^2 + b_k L\langle t_k(\mathbf{y}_k - \mathbf{x}_{k+1}), (t_k-1)\mathbf{x}_k - t_k\mathbf{y}_k - \mathbf{x}_*\rangle. \tag{2.17}$$

For the right hand side dot product, let

$$\mathbf{a} = t_k\mathbf{y}_k, \quad \mathbf{b} = t_k\mathbf{x}_{k+1}, \quad \mathbf{c} = (t_k - 1)\mathbf{x}_k + \mathbf{x}_*.$$

Assume $b_k \leq 2a$, which implies $a \geq \frac{1}{2}$, then the right hand side can be written as

$$\begin{aligned} t_{k-1}^2 R_k - t_k^2 R_{k+1} &\geq \frac{b_k L}{2}\left(\frac{2a}{b_k}\|\mathbf{a} - \mathbf{b}\|^2 + 2\langle \mathbf{c} - \mathbf{a}, \mathbf{a} - \mathbf{b}\rangle\right) \\ &\geq \frac{b_k L}{2}\left(\|\mathbf{a} - \mathbf{b}\|^2 + 2\langle \mathbf{c} - \mathbf{a}, \mathbf{a} - \mathbf{b}\rangle\right) \\ &= \frac{b_k L}{2}\left(\|\mathbf{b} - \mathbf{c}\|^2 - \|\mathbf{a} - \mathbf{c}\|^2\right) \\ &\geq \frac{(a+\frac{1}{2})L}{2}\left(\|\mathbf{b} - \mathbf{c}\|^2 - \|\mathbf{a} - \mathbf{c}\|^2\right) \end{aligned}$$

It can be written as

$$t_{k-1}^2 R_k - t_k^2 R_{k+1} \geq \frac{2a+1}{4}L(\|\mathbf{u}_{k+1}\|^2 - \|\mathbf{u}_k\|^2)$$

thus

$$t_k^2 R_{k+1} + \frac{2a+1}{4}L\|\mathbf{u}_{k+1}\|^2 \leq t_{k-1}^2 R_k + \frac{2a+1}{4}L\|\mathbf{u}_k\|^2,$$

which implies

$$t_k^2 R_{k+1} \leq t_k^2 R_{k+1} + \frac{2a+1}{4} L \|\mathbf{u}_{k+1}\|^2 \leq t_0^2 R_1 + \frac{2a+1}{4} L \|\mathbf{u}_1\|^2,$$

and

$$R_{k+1} \leq \frac{1}{t_k^2} [t_0^2 R_1 + \frac{2a+1}{4} L \|\mathbf{u}_1\|^2].$$

Theorem 2.18. *Assume the function $f(\mathbf{x}) : \mathbb{R}^n \longrightarrow \mathbb{R}$ is convex with a global minimizer $\mathbf{x}_*$. Assume $\nabla f(\mathbf{x})$ is Lipschitz continuous with constant L. Assume $t_k^2 - t_k = t_{k-1}^2$ with $t_0 = 1$. Consider the following accelerated gradient method*

$$\boxed{\begin{cases} \mathbf{x}_{k+1} & = \mathbf{y}_k - \eta_k \nabla f(\mathbf{y}_k) \\ \mathbf{y}_{k+1} & = \mathbf{x}_{k+1} + \frac{t_k - 1}{t_{k+1}} (\mathbf{x}_{k+1} - \mathbf{x}_k) \end{cases} \qquad \mathbf{x}_0 = \mathbf{y}_0.}$$

If

$$\frac{1}{2a} \frac{1}{L} \leq \eta_k \leq \frac{1}{a + \frac{1}{2}} \frac{1}{L}, \quad a \geq \frac{1}{2}, \quad \forall k,$$

then

$$f(\mathbf{x}_k) - f(\mathbf{x}_*) \leq \frac{4}{k^2} \left(f(\mathbf{x}_1) - f(\mathbf{x}_*) + \frac{2a+1}{4} L \|\mathbf{x}_1 - \mathbf{x}_*\|^2 \right).$$

Remark 2.14. *Notice that we only have $\eta_k \leq \frac{1}{L}$. Even though it may converge with a slightly larger η_k in practice, the accelerated gradient method might blow up for a step size like $\eta = \frac{2}{L}$, which is however a stable one for the gradient descent method.*

Appendices

Appendix A

Linear algebra

A.1 Eigenvalues and Courant-Fischer-Weyl min-max principle

Notations and quick facts:

- A^T denote the transpose. A^* denote the conjugate transpose of A.
- A matrix $A \in \mathbb{C}^{n\times n}$ is called Hermitian if $A^* = A$. Any Hermitian matrix A has real eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ with a complete set of orthonormal eigenvectors.
- Any real symmetric matrix has real eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ with a complete set of **real** orthonormal eigenvectors.

For a Hermitian matrix A, Rayleigh-Ritz quotient is defined as

$$R_A(x) = \frac{x^*Ax}{x^*x}, \quad x \in \mathbb{C}^n.$$

Let $\{v_j \in \mathbb{C}^n : j = 1, \cdots, n\}$ be orthonormal eigenvectors of A then they form a basis. Thus any vector x can be expressed as $x = \sum\limits_{j=1}^{n} a_j v_j$. Let V be a matrix with columns as v_j and a be a column vector with entries a_j. Then $x = Va$ and $x^*x = a^*V^*Va = a^*a = \sum\limits_{j=1}^{n} |a_j|^2$. Let Λ be a diagonal matrix with diagonal entries λ_j. We have $Av_j = \lambda_j v_j$ thus $Ax = \sum\limits_{j=1}^{n} a_j Av_j = \sum\limits_{j=1}^{n} a_j \lambda_j v_j = V\Lambda a$. Thus $x^*Ax = a^*V^*V\Lambda a = a^*\Lambda a = \sum\limits_{j=1}^{n} \lambda_j |a_j|^2$. So we get

$$\lambda_n \sum_{j=1}^{n} |a_j|^2 \leq \sum_{j=1}^{n} \lambda_j |a_j|^2 \leq \lambda_1 \sum_{j=1}^{n} |a_j|^2,$$

which is the min-max principle.

Theorem A.1 (Courant-Fischer-Weyl min-max principle)**.** *Let λ_1 and λ_n be the largest and the smallest eigenvalues of a Hermitian matrix A, then for any vector $x \in \mathbb{C}^n$,*

$$\lambda_n \leq \frac{x^* A x}{x^* x} \leq \lambda_1.$$

Next, we consider a positive definite matrix A, i.e., the eigenvalues are positive:

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n > 0.$$

Then A is invertible and A^{-1} has the same eigenvectors v_i with eigenvalues λ_i^{-1}.

Theorem A.2 (Kantorovich inequality)**.** *Let $A \in \mathbb{C}^{n\times n}$ be a positive definite matrix, then*

$$\frac{\|x\|^4}{(x^*Ax)(x^*A^{-1}x)} \geq \frac{4\lambda_1\lambda_n}{(\lambda_1+\lambda_n)^2}, \quad \forall x \in \mathbb{C}^n.$$

Proof. With similar discussions as before, we get

$$\frac{\|x\|^4}{(x^*Ax)(x^*A^{-1}x)} = \frac{\left[\sum\limits_{j=1}^{n} |a_j|^2\right]^2}{\left[\sum\limits_{j=1}^{n} \lambda_j |a_j|^2\right]\left[\sum\limits_{j=1}^{n} |a_j|^2/\lambda_j\right]} = \frac{1}{\sum\limits_{j=1}^{n} \lambda_j b_j} \frac{1}{\sum\limits_{j=1}^{n} b_j/\lambda_j},$$

where $b_j = \frac{|a_j|^2}{\sum\limits_{j=1}^{n} |a_j|^2}$. We can rewrite it as

$$\frac{\|x\|^4}{(x^*Ax)(x^*A^{-1}x)} = \frac{\phi(b)}{\psi(b)},$$

where $\phi(b) = \frac{1}{\sum\limits_{j=1}^{n} \lambda_j b_j}$ and $\psi(b) = \sum\limits_{j=1}^{n} b_j/\lambda_j$.

Consider the convex function $g(\lambda) = \frac{1}{\lambda}$, then $\phi(b) = g(\lambda_*)$ with a specific point $\lambda_* = \sum\limits_{j=1}^{n} \lambda_j b_j$.

Consider a line segment connecting $(\lambda_1, \frac{1}{\lambda_1})$ and $(\lambda_n, \frac{1}{\lambda_n})$ in the same plane where the graph of $g(\lambda)$ lies. Then this line segment intersects with the vertical line $\lambda = \lambda_*$ at some point $(\lambda_*, \frac{c}{\lambda_1} + \frac{d}{\lambda_n})$ where $c + d = 1$ and $c, d > 0$.

Notice that all the b_j form a set of convex combination coefficients, thus the value of $\psi(b)$ can be regarded as a convex combination of points $(\lambda_j, \frac{1}{\lambda_j})$ for all j, which is a point in the same plane. In particular, this point is on the vertical line $\lambda = \lambda_*$, and lower than the intersection point $(\lambda_*, \frac{c}{\lambda_1} + \frac{d}{\lambda_n})$, and higher than $(\lambda_*, \frac{1}{\lambda_*})$ due to the convexity of the function $g(\lambda) = \frac{1}{\lambda}$.

So we have

$$\frac{\phi(b)}{\psi(b)} \geq \frac{1/\lambda_*}{\frac{c}{\lambda_1} + \frac{d}{\lambda_n}}.$$

Notice that λ^* can also be written as $\lambda^* = c\lambda_1 + d\lambda_n$. Since $c = 1 - d$ and $d = 1 - c$, we get

$$\frac{c}{\lambda_1} + \frac{d}{\lambda_n} = \frac{c\lambda_n + d\lambda_1}{\lambda_1\lambda_n} = \frac{(1-d)\lambda_n + (1-c)\lambda_1}{\lambda_1\lambda_n} = \frac{\lambda_1 + \lambda_n - \lambda_*}{\lambda_1\lambda_n}.$$

Thus

$$\frac{\phi(b)}{\psi(b)} \geq \frac{1/\lambda_*}{\frac{c}{\lambda_1} + \frac{d}{\lambda_n}} = \frac{1/\lambda_*}{\frac{\lambda_1+\lambda_n-\lambda_*}{\lambda_1\lambda_n}} \geq \min_{\lambda\in(\lambda_n,\lambda_1)} \frac{1/\lambda}{\frac{\lambda_1+\lambda_n-\lambda}{\lambda_1\lambda_n}}.$$

The minimum value is achieved at $\lambda = (\lambda_1 + \lambda_n)/2$. Plug it in, the proof is concluded. □

A.2 Singular values

For a matrix $A \in \mathbb{C}^{m\times n}$, let A^* denote the conjugate transpose of A. Then A^*A and AA^* are both positive semi-definite (or definite) Hermitian matrices thus have real non-negative eigenvalues, denoted as $\lambda_i(A^*A)$ and $\lambda_i(AA^*)$ ordering by magnitudes.

The matrix A has $l = \min\{m, n\}$ singular values, defined as

$$\sigma_i(A) = \sqrt{\lambda_i(A^*A)} = \sqrt{\lambda_i(AA^*)}.$$

The singular values are defined for any matrix A and are always real non-negative. Eigenvalues are defined for square matrices and are not necessarily real.

A.3 Singular value decomposition

Theorem A.3. *Let $l \leq \min\{m, n\}$. Any matrix $A \in \mathbb{C}^{m\times n}$ of rank k has a decomposition $A = U\Sigma V^*$ (***singular value decomposition (SVD)***) where U of size $m \times l$ and V of size $n \times l$ have orthonormal columns and Σ of size $l \times l$ is diagonal matrix with singular values of A. It also has a compact decomposition $A = U_1\Sigma_1V_1$ (***compact SVD***) where where U of size $m \times k$ and V of size $n\times k$ have orthonormal columns and Σ_1 of size $k\times k$ is diagonal matrix with nonzero singular values of A.*

Proof. Assume $n \leq m$, we consider the matrix A^*A (if $n > m$, similar procedure for AA^*). The matrix A^*A is positive semi-definite Hermitian thus has non-negative real eigenvalues with a complete set of orthonormal eigenvectors. And A^*A has the same rank as A (why? good excercise to figure it out), thus A^*A has k nonzero eigenvalues. Let D be a $k \times k$ diagonal matrix with all nonzero eigenvalues of A^*A as diagonal entries, and V be a $n \times n$ matrix with orthonormal eigenvectors as columns. Then

$$V^*A^*AV = \begin{bmatrix} D & 0 \\ 0 & 0 \end{bmatrix}.$$

Let $V = [V_1\, V_2]$ corresponding to nonzero and zero eigenvalues, then

$$\begin{bmatrix} V_1^* \\ V_2^* \end{bmatrix} A^*A \begin{bmatrix} V_1 & V_2 \end{bmatrix} = \begin{bmatrix} D & 0 \\ 0 & 0 \end{bmatrix}.$$

By multiplying matrices in the left hand side above, we get

$$V_1^*A^*AV_1 = D, \quad V_2^*A^*AV_2 = 0.$$

Recall $V = [V_1\, V_2]$ has orthonormal columns thus $VV^* = I$, which implies $V_1V_1^* + V_2V_2^* = I$.

Next, since V_2 consists of eigenvectors to zero eigenvalue of A^*A, we get $A^*AV_2 = 0$ thus $V_2^*A^*AV_2 = 0$. So we must have $AV_2 = 0$ because it contradicts with $V_2^*A^*AV_2 = 0$ otherwise.

Let $U_1 = AV_1D^{-\frac{1}{2}}$ where $D^{\frac{1}{2}}$ is defined as taking square root for diagonal entries of D. Then

$$U_1D^{\frac{1}{2}}V_1^* = AV_1V_1^* = A(I - V_2V_2^*) = A - (AV_2)V_2^* = A.$$

The decomposition $A = U_1D^{\frac{1}{2}}V_1^*$ is exactly the compact SVD. Pick any U_2 of size $n \times (n-k)$ such that $U = [U_1\ U_2]$ is a unitary matrix and define Σ of size $n \times n$ as

$$\Sigma = \begin{bmatrix} D^{\frac{1}{2}} & 0 \\ 0 & 0 \end{bmatrix},$$

then $A = U\Sigma V$ is the full SVD. □

From the proof above, we get the following facts:

- The columns of V (right-singular vectors) are eigenvectors of A^*A.
- The columns of U (left-singular vectors) are eigenvectors of $AA*$.
- A real matrix A has real singular vectors.
- Let u_i and v_i be i-th columns of U and V corresponding i-th singular value $\sigma_i(A)$, then

$$Av_i = \sigma_i u_i, \quad A^*u_i = \sigma_i v_i.$$

- The rank of A is also the number of nonzero singular values of A.
- The compact SVD of A looks like this:

  $$A = \boxed{U_1}\Sigma_1\boxed{\;V_1^*\;}$$

  with

  $$\Sigma_1 = \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_k \end{bmatrix}.$$

  It is a convention to order σ_i in decreasing order: $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_k$.
- For a Hermitian (or real symmetric) positive semi-definite (PSD) matrix A and its SVD $A = U\Sigma V^*$ we must have $U = V$, thus its SVD $A = U\Sigma U^*$ is also its eigenvalue decomposition. Therefore, singular values are also eigenvalues for PSD matrices.

A.4 Vector norms

For $x = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix}^T$:

- *2-norm*: $\|x\| = \sqrt{\sum\limits_{j=1}^{n} |x|_j^2}$.
- *1-norm*: $\|x\|_1 = \sum\limits_{j=1}^{n} |x|_j$.
- ∞*-norm*: $\|x\|_\infty = \max_j |x|_j$.

A.5 Matrix norms

For a rank k matrix $A = (a_{ij})$ of size $m\times n$, assume its SVD is $A = U\Sigma V$ with nonzero singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_k$. Let $\boldsymbol{\sigma} = \begin{bmatrix} \sigma_1 & \sigma_2 & \cdots & \sigma_k \end{bmatrix}^T$. There are many norms of matrices. The following are a few important ones:

- *Spectral norm*: $\|A\|$ is defined as $\|A\| = \max\limits_{x\in\mathbb{C}^n} \frac{\|Ax\|}{\|x\|}$ ($x \in \mathbb{R}^n$ for real matrices) and $\|A\|$ is equal to the largest singular value of A. By Courant-Fischer-Weyl min-max principle Theorem A.1,

  $$\frac{\|Ax\|}{\|x\|} = \sqrt{\frac{\|Ax\|^2}{\|x\|^2}} = \sqrt{\frac{x^*A^*Ax}{x^*x}} \leq \sqrt{\lambda_1(A^*A)}.$$

  By taking $x = v_1$, the eigenvector of A^*A corresponding to $\lambda_1(A^*A)$, we get $\|A\| = \sqrt{\lambda_1(A^*A)} = \sigma_1$.

- *Frobenius norm*: $\|A\|_F = \sqrt{tr(A^*A)} = \sqrt{\sum\limits_{i=1}^{m}\sum\limits_{j=1}^{n}|a_{ij}|^2}$. We have $\|A\|_F = \|\boldsymbol{\sigma}\|$ because

  $$\|A\|_F = \sqrt{tr(V^*\Sigma U^*U\Sigma V)} = \sqrt{tr(V^*\Sigma^2 V)} = \sqrt{tr(VV^*\Sigma^2)} = \sqrt{\sum_j \sigma_j^2},$$

  where we have used the property of trace function $tr(ABC) = tr(CAB)$ for three matrices A, B, C of proper sizes.

- *Nuclear norm*: $\|A\|_* = \sigma_1 + \sigma_2 + \cdots \sigma_k$. Then the nuclear norm of A is simply $\|\boldsymbol{\sigma}\|_1$.

- *Matrix 1-norm*: $\|A\|_1 = \max\limits_{x\in\mathbb{C}^n} \frac{\|Ax\|_1}{\|x\|_1}$ ($x \in \mathbb{R}^n$ for real matrices). Since Ax is a linear combination of columns of A, therefore $\|Ax\|_1$ for $\|x\|_1 = 1$ is less than or equal to a convex combination of 1-norm of columns of A thus $\|A\|_1 = \max\limits_j \sum\limits_{i=1}^{m} |a_{ij}|$.

- *Matrix ∞-norm*: $\|A\|_\infty = \max\limits_{x\in\mathbb{C}^n} \frac{\|Ax\|_\infty}{\|x\|_\infty}$ ($x \in \mathbb{R}^n$ for real matrices). It is easy to show $\|A\|_\infty = \max\limits_i \sum\limits_{j=1}^{n} |a_{ij}|$.

Useful facts:

- For a matrix norm $|||A|||$ induced by vector norms such as spectral norm, $1-norm$ and ∞-norm, by definition we have

  $$|||Ax||| \leq |||A||| \cdot |||x|||.$$

  Since $|||ABx||| \leq |||A||| \cdot |||Bx||| \leq |||A||| \cdot |||B||| \cdot |||x|||$, we also have

  $$|||AB||| \leq |||A||| \cdot |||B|||.$$

- For a matrix norm $|||A|||$ defined through singular values such as spectral norm, Frobenius norm and nuclear norm, it is invariant after unitary transformation: let T and S be unitary matrices, then $|||A||| = |||TAS|||$. Notice that $TAS = (TU)\Sigma(V^*S)$ is the SVD of TAS, so TAS has the same singular values as A.

A.6 Normal matrices

A matrix A is normal if $A^*A = AA^*$. The following are equivalent:

- $A^*A = AA^*$.

- $\sigma_i(A) = |\lambda_i(A)|$.
- A is diagonalizable by unitary matrix: $A = U\Lambda U^*$ where Λ is diagonal. (Obviously, $A = U\Lambda U^*$ is also its eigenvalue decomposition. In other words, A has a complete set of orthonormal eigenvectors (but eigenvalues could be negative, could be complex). If Λ has negative or complex diagonal entries, then $A = U\Lambda U^*$ is not SVD and its SVD has the form $A = U|\Lambda|V^*$ where $|\Lambda|$ is a diagonal matrix with diagonal entries $|\lambda_i|$.)

The equivalency can be easily established by SVD. All Hermitian matrices including PSD matrices are normal. Here is one non-Hermitian normal matrix example: a matrix A is skew-Hermitian if $A^* = -A$. Skew-Hermitian matrices are normal and always have purely imaginary eigenvalues.

Appendix B

Discrete Laplacian

B.1 Finite difference approximations

For a smooth function $u(x)$, define the following finite difference operators approximating $u'(x)$ at the point $\bar{x}$:

- Forward Difference: $D_+u(\bar{x}) = \frac{u(\bar{x}+h)-u(\bar{x})}{h}$.
- Backward Difference: $D_-u(\bar{x}) = \frac{u(\bar{x})-u(\bar{x}-h)}{h}$.
- Centered Difference: $D_0u(\bar{x}) = \frac{u(\bar{x}+h)-u(\bar{x}-h)}{2h}$.

By Taylor expansion, the truncation errors of these operators are

$$D_\pm u(\bar{x}) = u'(\bar{x}) + \mathcal{O}(h), \quad D_0u(\bar{x}) = u'(\bar{x}) + \mathcal{O}(h^2).$$

Define $\hat{D}_0u(\bar{x}) = \frac{u(\bar{x}+h/2)-u(\bar{x}-h/2)}{h}$, then a classial second order finite difference approximation to $u''(x)$ at $\bar{x}$ is given by (denoted by D^2):

$$D^2u(\bar{x}) = D_+D_-u(\bar{x}) = \hat{D}_0\hat{D}_0u(\bar{x}) = \frac{u(\bar{x}+h) - 2u(\bar{x}) + u(\bar{x}-h)}{h^2} = u''(\bar{x}) + \mathcal{O}(h^2).$$

The Poisson's equations are

- 1D: $u''(x) = f(x)$
- 2D: $\Delta u(x,y) = u_{xx} + u_{yy} = f(x,y)$.
- 3D: $\Delta u(x,y,z) = f(x,y,z)$.

B.2 1D BVP: Dirichlet b.c.

Consider solving the 1D Poisson's equation with homogeneous Dirichlet boundary conditions:

$$\begin{cases} -u''(x) = f(x), & x \in (0,1), \\ u(0) = 0, \ u(1) = 0. \end{cases} \tag{B.1}$$

Discretize the domain $[0, 1]$ by a uniform grid with spacing $h = \frac{1}{n+1}$ and n interior nodes: $x_j = jh$, $j = 1, 2, \cdots, n$. See Figure B.1. Let $u(x)$ denote the true solution and $f_j = f(x_j)$. For convenience, define two ghost points $x_0 = 0$ and $x_{n+1} = 1$. Let u_j be the value of the numerical solution at x_j. Since two end values are given as $u(0) = 0, u(1) = 0$, only the interior point values $u_j (j = 1, \cdots, n)$ are unknowns. After approximating $\frac{d^2}{dx^2}$ by D^2, we get a finite difference scheme

$$-D^2 u_j = \frac{-u_{j-1} + 2u_j - u_{j+1}}{h^2} = f_j, \quad j = 1, 2, \cdots, n \tag{B.2}$$

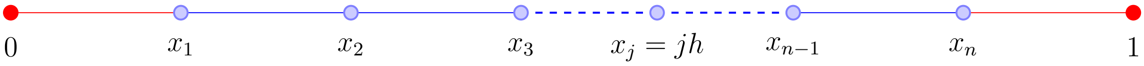


Figure B.1: An illustration of the discretized domain.

Define

$$U_h = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}, \quad F = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_n \end{bmatrix}, \quad K = \frac{1}{h^2} \begin{pmatrix} 2 & -1 & & & & \\ -1 & 2 & -1 & & & \\ & -1 & 2 & -1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & -1 & 2 & -1 \\ & & & & -1 & 2 \end{pmatrix}.$$

With the boundary values $u_0 = 0$ and $u_{n+1} = 0$ from the boundary condition, we can rewrite the finite difference scheme in the matrix vector form:

$$KU_h = F.$$

B.2.1 Eigenvalues of K

In general it is difficult to find exact eigenvalues of a large matrix. For the K matrix, if U is an eigenvector, then $KU = \lambda U$ approximates the eigenfunction problem:

$$-u'' = \lambda u, \qquad u(0) = u(1) = 0. \tag{B.3}$$

This is standard knowledge in an ordinary differential equation course to find such eigenfunctions as $\sin(m\pi x)$ with eigenvalues $\lambda_m = m^2\pi^2$ for $m = 1, 2, \cdots$. So we expect that the eigenvectors of K would look like $\sin(m\pi x)$ for small h. With the following trigonometric formulas,

$$\sin(m\pi x_{j+1}) = \sin(m\pi(x_j + h)) = \sin(m\pi x_j)\cos(m\pi h) + \cos(m\pi x_j)\sin(m\pi h),$$

$$\sin(m\pi x_{j-1}) = \sin(m\pi(x_j - h)) = \sin(m\pi x_j)\cos(m\pi h) - \cos(m\pi x_j)\sin(m\pi h),$$

thus,

$$-\sin(m\pi x_{j-1}) + 2\sin(m\pi x_j) - \sin(m\pi x_{j+1}) = (2 - 2\cos(m\pi h))\sin(m\pi x_j).$$

Notice the facts that $\sin(m\pi x_0) = 0$ and $\sin(m\pi x_{n+1}) = 0$, we also have

$$2\sin(m\pi x_1) - \sin(m\pi x_2) = (2 - 2\cos(m\pi h))\sin(m\pi x_1),$$

$$-\sin(m\pi x_{n-1}) + 2\sin(m\pi x_n) = (2 - 2\cos(m\pi h))\sin(m\pi x_n).$$

Let $\mathbf{x} = [x_1, x_2, \cdots, x_n]^T$, then the eigenvectors of K are $\mathbf{v}_m = \sin(m\pi\mathbf{x})$:

$$K\sin(m\pi\mathbf{x}) = \frac{1}{h^2}(2 - 2\cos(m\pi h))\sin(m\pi\mathbf{x}), \quad m = 1, 2, \cdots, n,$$

with eigenvalues

$$\lambda_m = \frac{1}{h^2}[2 - 2\cos(m\pi h)] = 4\frac{1}{h^2}\sin^2(m\frac{\pi}{2}h).$$

Since all eigenvalues are positive, K is a positive definite matrix, thus singular values are also eigenvalues. We have

$$\|K\| = \sigma_1 = \max_m 4\sin^2(m\frac{\pi}{2}h) = 4\frac{1}{h^2}\sin^2(\frac{\pi}{2}\frac{n}{n+1}) \le 4\frac{1}{h^2},$$

and

$$\min_m 4\sin^2(m\frac{\pi}{2}h) = 4\frac{1}{h^2}\sin^2(\frac{\pi}{2}\frac{1}{n+1})$$

Thus we have

$$4\frac{1}{h^2}\sin^2(\frac{\pi}{2}h)I \le K < \frac{4}{h^2}I$$

for any n where $h = \frac{1}{n+1}$.

Define the eigenvector matrix as $S = [\sin(\pi\mathbf{x}) \quad \sin(2\pi\mathbf{x}) \quad \cdots \quad \sin(n\pi\mathbf{x})]$ and consider the diagonal matrix Λ with diagonal entries $\frac{2-2\cos(m\pi h)}{h^2}, m = 1, \cdots, n$. Then $K = S\Lambda S^{-1}$, and $K^{-1} = S\Lambda^{-1}S^{-1}$. Therefore we get

$$\frac{1}{4}h^2 I \le K^{-1} < \frac{h^2}{4\sin^2(\frac{\pi}{2}h)}I.$$

We can check that $4\frac{1}{h^2}\sin^2(\frac{\pi}{2}h)$ is a decreasing function of h, and $4\frac{1}{h^2}\sin^2(\frac{\pi}{2}h) \to \pi^2$ as $h \to 0$ L'Hospital's rule.

Thus we also have

$$\frac{1}{4}h^2 I \le K^{-1} < \frac{1}{\pi^2}I,$$

and $\|K^{-1}\| \le \frac{1}{\pi^2}$.

Appendix C

Basic Theorems in Analysis

The following results are standard in many real analysis books, e.g. [2].

C.1 Completeness of Real Numbers

Theorem C.1 (Completeness Theorem for Sequences)**.** *If a sequence of real numbers* $\{a_n\} \subset \mathbb{R}$ *is monotone and bounded, then it converges.*

Theorem C.2 (Completeness Theorem for Sets)**.** *If a set of real numbers* $S \subset \mathbb{R}$ *is bounded, then its supremum and infimum exist.*

C.2 Compactness

Definition C.1. *A subset* S *in* $\mathbb{R}^n$ *is called compact if any sequence* $\{a_n\} \subseteq S$ *has a convergent subsequence* $\{a_{n_i}\}$ *with limit point in* S*.*

Theorem C.3 (Heine–Borel)**.** *A subset* S *in* $\mathbb{R}^n$ *is compact if and only if it is closed and bounded.*

Theorem C.4 (Bolzano–Weierstrass)**.** *Any bounded sequence in* $\mathbb{R}^n$ *has a convergent subsequence.*

Using Theorems above and proof by contradiction, we can show

Theorem C.5. *A continuous function* $f(\mathbf{x})$ *attains its maximum and minimum on a compact set in* $\mathbb{R}^n$*.*

C.3 Cauchy Sequence

Definition C.2. *A sequence* $\{\mathbf{x}_k\} \subset \mathbb{R}^n$ *is Cauchy if*

$$\forall \varepsilon > 0, \exists N, \forall m, n \geq N, \|\mathbf{x}_m - \mathbf{x}_n\| < \varepsilon.$$

Theorem C.6. *A sequence* $\{\mathbf{x}_k\} \subset \mathbb{R}^n$ *converges if and only if it is a Cauchy sequence.*

C.4 Infinite Series

Theorem C.7. *If* $\sum\limits_{n=0}^{\infty} a_n$ *converges, then* $\lim\limits_{n\to\infty} a_n = 0$.

Theorem C.8. *For a decreasing function* $f(x)$, $\sum\limits_{n=1}^{\infty} f(n)$ *converges if and only if* $\int_N^\infty f(x)dx$ *is finite for some* $N > 0$.

The theorem above implies $\sum\limits_{n=1}^{\infty} \frac{1}{n^2}$ converges and the *Harmonic Sum* $\sum\limits_{n=1}^{\infty} \frac{1}{n} = +\infty$.

References

Bibliography

[1] Douglas R Farenick and Fei Zhou. Jensen's inequality relative to matrix-valued measures. *Journal of mathematical analysis and applications*, 327(2):919–929, 2007.

[2] Arthur Mattuck. *Introduction to analysis.* Prentice Hall, 1999.

[3] Yurii Evgen'evich Nesterov. A method of solving a convex programming problem with convergence rate $o(\frac{1}{k^2})$. In *Doklady Akademii Nauk*, volume 269, pages 543–547. Russian Academy of Sciences, 1983.

[4] Ting On To and Kai Wing Yip. A generalized jensen's inequality. *Pacific Journal of Mathematics*, 58(1):255–259, 1975.