Notes for Numerical Optimization
Fall 2024

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

Part I

Smooth problems

1

Prerequisites

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.6 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

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

Example 1.2. *Consider $f(\mathbf{x}) = \|\mathbf{x}\|^2 = \sum_i x_i^2$ for $\mathbf{x} = \begin{bmatrix} x_1 & \cdots & x_N\end{bmatrix}^T$, then $\nabla^2 f(\mathbf{x})$ is identity matrix thus $f(\mathbf{x}) = \|\mathbf{x}\|^2$ satisfies the Jensen's inequality:*

$$\|a\mathbf{x}+(1-a)\mathbf{y}\|^2 \leq a\|\mathbf{x}\|^2 + (1-a)\|\mathbf{y}\|^2, \quad \forall a \in (0,1), \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^N.$$

Notice that this is similar to but different from the triangle inequalilty for the norm function:

$$\|a\mathbf{x}+b\mathbf{y}\| \leq a\|\mathbf{x}\| + b\|\mathbf{y}\|, \quad \forall a, b > 0, \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^N.$$

In particular, the function $g(\mathbf{x}) = \|\mathbf{x}\|$ is also convex due to the triangle inequalilty above.

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

It can also be extended to a vector valued function:

Theorem 1.7 (Jensen's inequality in integral form)**.** *Let $\mathbf{g} : \mathbb{R} \to \mathbb{R}^n$ be a single variable vector-valued function, which is integrable on $[a, b]$ in the sense that integral exists for each entry of the vector $\mathbf{g}$. If $\phi : \mathbb{R}^n \longrightarrow \mathbb{R}$ is convex, then*

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

Proof. As in previous theorem, this result can be proven without assuming the differentiability of the convex function. For convenience, assume $\nabla\phi$ exists, then Lemma 1.1 implies

$$\phi(\mathbf{t}) \geq \phi(\mathbf{t}_0) + \langle \nabla\phi(\mathbf{t}_0), \mathbf{t} - \mathbf{t}_0 \rangle. \tag{1.2}$$

Plug in $\mathbf{t}_0 = \frac{1}{b-a}\int_a^b \mathbf{g}(x)dx$ and $\mathbf{t} = \mathbf{g}(x)$ we get

$$\phi[\mathbf{g}(x)] \geq \phi\left(\frac{1}{b-a}\int_a^b \mathbf{g}(x)dx\right) + \langle \nabla\phi(\mathbf{t}_0), \mathbf{g}(x) - \frac{1}{b-a}\int_a^b \mathbf{g}(x)dx \rangle.$$

Integrate both sides for variable x, we get

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

□

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*

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

Proof. If we apply the previous theorem to the norm function, which is convex due to triangle inequality, then we get this inequality. We give a different proof here. 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 1.8 (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 [17, 6]*

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 1.3. *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 1.9. *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 1.4. *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.6) 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 1.8:*

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

thus

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

1.4 Optimality conditions

Definition 1.3. *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 1.10 (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 1.4. $\mathbf{x}^*$ *is called a stationary point or a critical point of the function* $f(\mathbf{x})$ *if* $\nabla f(\mathbf{x}^*) = \mathbf{0}$.

Theorem 1.11 (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 1.10. 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 1.12 (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 1.13. *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 1.6. *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 1.10 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 1.14. *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 1.13, 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 1.15. *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 1.7. *Strict convexity is not enough to ensure the existence of a minimizer. For instance, $f(x) = e^x$ is strictly convex.*

2

The gradient descent method

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

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.

2.1 Stable step sizes

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

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

Remark 2.1. *Notice that there is no assumption on the existence of Hessian. But if assuming $\|\nabla^2 f\| \leq L$, then by Theorem 1.4,*

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

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 2.2. *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

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

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 Convergence for Lipschitz continuous gradient

Theorem 2.1. *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:*

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

and $\lim\limits_{k\to\infty} \|\nabla f(\mathbf{x}_k)\| = 0$ *with*

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

Remark 2.3. *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

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

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

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

□

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

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

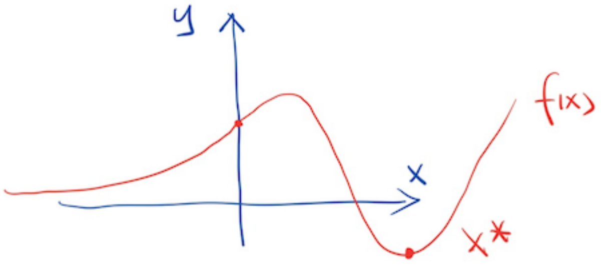
Therefore, $\mathbf{y}_n$ has a limit, which further implies the convergence of $\mathbf{x}_k$. **However, the assumption of convergence of $\sum\limits_{k=0}^{\infty} \|\nabla f(\mathbf{x}_k)\|$ is in general not true**. By the proof of the theorem above, we only have the convergence $\sum\limits_{k=0}^{\infty} \|\nabla f(\mathbf{x}_k)\|^2$, 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 2.1, 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 2.1 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!

2.3 Convergence for convex functions

Theorem 2.2. *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 2.4. *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 1.13, $\mathbf{x}_0$ 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 $\phi(\mathbf{x}_0) \le \phi(\mathbf{y}) - \frac{1}{2L}\|\nabla\phi(\mathbf{y})\|^2$ implies

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

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

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

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

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

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

and adding two we get

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

□

Theorem 2.3. *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: $f(\mathbf{x}) \ge f(\mathbf{x}_*), \forall \mathbf{x}$. 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 2.1, the following holds:*

$$f(\mathbf{x}_k) - f(\mathbf{x}_*) < \frac{1}{k\omega}\|\mathbf{x}_0 - \mathbf{x}_*\|^2, \quad \omega = \eta(\frac{2}{L} - \eta).$$

Remark 2.5. *From the proof, we will see*

1. $\|\mathbf{x}_{k+1} - \mathbf{x}_*\|^2 \le \|\mathbf{x}_k - \mathbf{x}_*\|^2 - \omega\|\nabla f(\mathbf{x}_k)\|^2$ *so* $r_k = \|\mathbf{x}_k - \mathbf{x}_*\|$ *is decreasing but* $\|\mathbf{x}_k - \mathbf{x}_*\| \to 0$ *is wrong because* $\mathbf{x}_*$ *may not be unique.*

2. *The theorem implies* $R_{k+1} = f(\mathbf{x}_{k+1}) - f(\mathbf{x}_*) < \frac{1}{\omega} r_k^2$, *thus* $f(\mathbf{x}_k) - f(\mathbf{x}_*)$ *goes to zero much faster than* $\|\mathbf{x}_k - \mathbf{x}_*\|$ *if* $\mathbf{x}_k \to \mathbf{x}_*$.

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 2.2 in the last inequality.

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

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

thus

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

With Cauchy-Schwartz inequality,

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

which can be written as

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

thus

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

Recall Theorem 2.1 gives

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

thus

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

$$0 \leq R_{k+1} \leq R_k - \omega \|\nabla f(\mathbf{x}_k)\|^2 \leq R_k - \omega \frac{R_k^2}{r_k^2}.$$

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

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

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

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

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

□

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 2.3 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 $x_k \in [-a, a]$.*

2.4 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 1.9) 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 2.4. *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

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

and

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

By Lemma 1.1, $\phi(\mathbf{x})$ is also convex. Moreover, $\nabla\phi(\mathbf{x}_0) = \mathbf{0}$, thus by Theorem 1.13, $\mathbf{x}_0$ is a global minimizer of $\nabla\phi(\mathbf{x})$. So we get

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

thus

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

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

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

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

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

which implies

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

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 2.5. *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 2.6. *Plug in $\mu = 0$ and compare it with Theorem 2.2.*

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 2.2 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 2.6 (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*

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

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

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

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

Remark 2.7. *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 2.5, 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

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

□

2.5 Convergence under the Polyak-Lojasiewicz inequality

The Polyak-Lojasiewicz (PL) inequality or condition is given as

$$\frac{1}{2}\|\nabla f(\mathbf{x})\|^2 \geq \mu(f(\mathbf{x}) - f(\mathbf{x}_*)), \quad \mu > 0, \tag{2.9}$$

where $\mathbf{x}_*$ is one of the global minimizers to $f(\mathbf{x})$. Note that this inequality implies that every critical point is a global minimizer.

Lemma 2.3. *A strongly convex function satisfies the Polyak-Lojasiewicz inequality* (2.9)*.*

Proof. A strongly convex function satisfies

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

By a minimization w.r.t. $\mathbf{y}$, we get $f(\mathbf{x}_*) \geq f(\mathbf{x}) - \frac{1}{2\mu} \|\nabla f(\mathbf{x})\|^2$. □

There are interesting problems such as certain machine learning models, which satisfy the Polyak-Lojasiewicz inequality but are not necessarily strongly convex.

Example 2.5. *Consider $f(\mathbf{x}) = \frac{1}{2}\|A\mathbf{x} - b\|^2$ with $\mathbf{x}, b \in \mathbb{R}^n$ and $A \in \mathbb{R}^{m \times n}$. Then $\nabla^2 f = A^T A$, and $f(\mathbf{x})$ is strongly convex if and only if $\nabla^2 f \geq \mu \mathbb{I}$.*

In general, $A^T A \geq \mu \mathbb{I}$ is not true for $\mu > 0$. For example, the smallest eigenvalue of $A^T A$ is zero if $A \in R^{m \times n}, m < n$. By Theorem 2.7 below, (2.9) still holds even if $A \in R^{m \times n}, m < n$. For simplicity, assume b is in the column space of A, then $f(\mathbf{x}_*) = 0$.

Example 2.6. *The logistic regression cost function*

$$f(\mathbf{x}) = \sum_{i=1}^{n} \log[1 + \exp(\mathbf{b}_i \mathbf{a}_i^T \mathbf{x})]$$

can be written as $f(\mathbf{x}) = g(A\mathbf{x})$, with $g(\mathbf{y})$ being only strictly convex but not strongly convex. For example $g(x) = \log(1 + \exp x)$ has $g''(x) = \frac{e^x}{(e^x+1)^2} \to 0$ as $x \to \infty$. So Theorem 2.7 below does not apply but $f(\mathbf{x})$ satisfies the Polyak-Lojasiewicz inequality on a compact set.

Theorem 2.7. *$f(\mathbf{x}) = g(A\mathbf{x})$ with $A \in \mathbb{R}^{m \times n}$ satisfies the Polyak-Lojasiewicz inequality if $g(\mathbf{x})$ is strongly convex.*

Proof. Strong convexity of $g(\mathbf{x})$ gives

$$g(\mathbf{v}) \geq g(\mathbf{u}) + \langle \nabla g(\mathbf{u}), \mathbf{v} - \mathbf{u} \rangle + \frac{\mu}{2} \|\mathbf{v} - \mathbf{u}\|^2$$

thus

$$\begin{aligned} g(A\mathbf{y}) &\geq g(A\mathbf{x}) + \langle \nabla g(A\mathbf{x}), A\mathbf{y} - A\mathbf{x} \rangle + \frac{\mu}{2} \|A\mathbf{y} - A\mathbf{x}\|^2 \\ &= g(A\mathbf{x}) + \langle A^T \nabla g(A\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{\mu}{2} \|A\mathbf{y} - A\mathbf{x}\|^2 \\ &= g(A\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{\mu}{2} \|A\mathbf{y} - A\mathbf{x}\|^2, \end{aligned}$$

and

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

There might be multiple global minimizers to $f(\mathbf{x})$. Let $\mathbf{x}_p$ be the projection of $\mathbf{x}$ to the set of global minimizers to $f(\mathbf{x})$, then

$$\begin{aligned} f(\mathbf{x}_p) &\geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{x}_p - \mathbf{x} \rangle + \frac{\mu}{2}\|A\mathbf{x}_p - A\mathbf{x}\|^2 \\ f(\mathbf{x}_p) &\geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{x}_p - \mathbf{x} \rangle + \frac{\mu\sigma^2}{2}\|\mathbf{x}_p - \mathbf{x}\|^2, \end{aligned}$$

where σ is the smallest nonzero eigenvalue of A and we used the linear algebra fact that $\|A\mathbf{x}_p - A\mathbf{x}\|^2 \geq \sigma^2\|\mathbf{x}_p - \mathbf{x}\|^2$, which will be proven at the end. So we get

$$\begin{aligned} f(\mathbf{x}_p) &\geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{x}_p - \mathbf{x} \rangle + \frac{\mu\sigma^2}{2}\|\mathbf{x}_p - \mathbf{x}\|^2 \\ &\geq f(\mathbf{x}) + \min_{\mathbf{y}} \left(\langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{\mu\sigma^2}{2}\|\mathbf{y} - \mathbf{x}\|^2 \right) \\ &\geq f(\mathbf{x}) + \min_{\mathbf{y}} \left(\langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{\mu\sigma^2}{2}\|\mathbf{y} - \mathbf{x}\|^2 \right) \\ &= f(\mathbf{x}) - \frac{1}{\mu\sigma^2}\|\nabla f(\mathbf{x})\|^2, \end{aligned}$$

thus (2.9) is satisfied with a parameter $\mu\sigma > 0$.

Finally, we discuss why $\|A\mathbf{x}_p - A\mathbf{x}\|^2 \geq \sigma^2\|\mathbf{x}_p - \mathbf{x}\|^2$. Let the compact SVD (see Appendix A.3) of A be

$$A = \boxed{U}\,\Sigma\,\boxed{V^T}, \quad \Sigma = \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_k \end{bmatrix}, \quad \sigma_1 \geq \cdots \geq \sigma_k = \sigma > 0.$$

In general, we can only have $\frac{\|A\mathbf{y}\|}{\|\mathbf{y}\|} \geq 0, \forall \mathbf{y} \neq \mathbf{0}$, e.g., simply take $\mathbf{x}$ as the right singular vector for zero singular value of A. However, $\mathbf{x}_p - \mathbf{x}$ cannot be the zero singular vector to A if $\mathbf{x}$ is not a minimizer, because otherwise $A(\mathbf{x}_p - \mathbf{x}) = 0 \Rightarrow g(A\mathbf{x}_p) = g(A\mathbf{x})$, which means $\mathbf{x}$ is also a global minimizer.

Now let $\mathbf{v}_1, \cdots, \mathbf{v}_k$ be the right singular vectors of A with nonzero singular values $\sigma_i, i = 1, \cdots, k$. Let $\mathbf{v}_{k+1}, \cdots, \mathbf{v}_n$ be the the right singular vectors of A with zero singular values. Then $\mathbf{v}_1, \cdots, \mathbf{v}_n$ form an orthonormal basis of $\mathbb{R}^n$, and we have $\mathbf{x}_p - \mathbf{x} = \sum\limits_{i=1}^{n} x_i\mathbf{v}_i$. Since $\mathbf{x}_p$ is the projection of $\mathbf{x}$ to the set of global minimizers, i.e., there is no other minimizer that is closer to $\mathbf{x}$, we must have $x_i = 0$ for $i = k+1, \cdots, n$. Assume $x_j \neq 0$ for some $j > k$, let $\mathbf{y} = \mathbf{x}_p - x_j\mathbf{v}_j$, then $A\mathbf{y} = A\mathbf{x}_p \Rightarrow g(A\mathbf{y}) = g(A\mathbf{x}_p)$ implies $\mathbf{y}$ is another minimizer and $\|\mathbf{y} - \mathbf{x}\| < \|\mathbf{x}_p - \mathbf{x}\|$ since

$$\mathbf{y} - \mathbf{x} = \sum_{i=1, j\neq i}^{n} x_i\mathbf{v}_i.$$

Therefore, we must have $\mathbf{x}_p - \mathbf{x} = \sum_{i=1}^{k} x_i \mathbf{v}_i$ thus $\|\mathbf{x}_p - \mathbf{x}\|^2 = x_1^2 + \cdots + x_k^2$. Moreover,

$$A(\mathbf{x}_p - \mathbf{x}) = \boxed{U}\, \Sigma \boxed{V^T}\, \boxed{V} \begin{pmatrix} x_1 \\ \vdots \\ x_k \end{pmatrix} = \boxed{U}\, \Sigma \begin{pmatrix} x_1 \\ \vdots \\ x_k \end{pmatrix} \Rightarrow \|A(\mathbf{x}_p - \mathbf{x})\| \geq \sigma \|\mathbf{x}_p - \mathbf{x}\|.$$

□

Theorem 2.8. *Let $f(\mathbf{x})$ satisfy the Polyak-Lojasiewicz inequality* (2.9) *with $\mu > 0$ and $\nabla f(\mathbf{x})$ is Lipschitz-continuous with Lipschitz constant L. The gradient method with a step-size of* $1/L$, $\mathbf{x}_{k+1} = \mathbf{x}_k - \frac{1}{L}\nabla f(\mathbf{x}_k)$ *has a global linear convergence rate*

$$f(\mathbf{x}_k) - f(\mathbf{x}_*) \leq \left(1 - \frac{\mu}{L}\right)^k (f(x_0) - f(x_*)).$$

Remark 2.8. *Once again, $\mathbf{x}_k \to \mathbf{x}_*$ cannot be true in general since $\mathbf{x}_*$ may not be unique.*

Proof. The Descent Lemma (Lemma 2.1) gives

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

which becomes the following after using $\mathbf{x}_{k+1} = \mathbf{x}_k - \frac{1}{L}\nabla f(\mathbf{x}_k)$ and (2.9) ,

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

$$\Rightarrow f(\mathbf{x}_{k+1}) - f(\mathbf{x}_*) \leq (1 - \frac{\mu}{L})(f(\mathbf{x}_k) - f(\mathbf{x}_*)).$$

□

2.6 Steepest descent

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

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \eta_k \nabla f(\mathbf{x}_k) \tag{2.10a}$$

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

$$\eta_k = \arg\min_{\alpha > 0} f(\mathbf{x}_k - \alpha \nabla f(\mathbf{x}_k)). \tag{2.10b}$$

Such an optimal step size is also called *full relaxation.* The method (2.10) is often called *the steepest descent*, which is rarely used in practice unless (2.10b) can be easily computed. Nonetheless, analyzing its convergence rate is a starting point for understanding practical algorithms.

Theorem 2.9. *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* (2.10) *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 2.9. *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 2.6 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

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}_*) \le f(\mathbf{x}_k) - f(\mathbf{x}_*) - \frac{1}{2L}\|\nabla f(\mathbf{x}_k)\|^2. \tag{2.11}$$

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 (2.11), we get the convergence rate. □

2.7 Quadratic functions

The better convergence rate $\left(\frac{L-\mu}{L+\mu}\right)^2$ can be proven for the steep descent method (2.10) 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

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

Since $\nabla^2 f(\mathbf{x}) \equiv A \geq \mu I$, $f(\mathbf{x})$ 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 (2.10) 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

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

which is the same rate as in Theorem 2.9. In order to get a better rate, we can use the Kantorovich inequality in Theorem A.2:

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

2.8 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 [8] 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 2.3.

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 $\{f(\mathbf{y}_k)\}$.

2.9 Convergence rate of the accelerated gradient method

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 $\mathbf{x} = \mathbf{y}_k$:

$$f(\mathbf{x}_{k+1}) \leq f(\mathbf{y}_k) + \langle \nabla f(\mathbf{y}_k), \mathbf{x}_{k+1} - \mathbf{y}_k \rangle + \frac{L}{2}\|\mathbf{y}_k - \mathbf{x}_{k+1}\|^2. \tag{2.12}$$

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 (2.12), 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:

$$t_k(t_k - 1)R_k - t_k^2 R_{k+1} \geq \frac{L}{2}\|t_k(\mathbf{y}_k - \mathbf{x}_{k+1})\|^2 + L\langle t_k(\mathbf{y}_k - \mathbf{x}_{k+1}), (t_k - 1)\mathbf{x}_k - t_k \mathbf{y}_k - \mathbf{x}_* \rangle. \tag{2.13}$$

Assume we have

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

then

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

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

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

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

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

thus

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

which implies

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

and

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

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 $t_k\geq\frac{k+2}{2}$ by induction.

All the discussions can be summarized as:

Theorem 2.10. *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 2.10. *Obviously the theorem still holds if we plug in $t_k = \frac{k+2}{2}$, 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 $\eta_k = \frac{1}{b_k}\frac{1}{L} \le \frac{1}{(a+\frac{1}{2})}\frac{1}{L}$ with $a > 0$, then

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

we have

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

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

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

Multiply it by t_k:

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

Assume we have

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

then

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

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

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

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 2.11. *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 2.11. *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.*

3

The 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* $\quad \mathbf{p}_k = -\nabla f(\mathbf{x}_k)$.
2. *Newton's method* $\quad \mathbf{p}_k = -[\nabla^2 f(\mathbf{x}_k)]^{-1}\nabla f(\mathbf{x}_k)$.
3. *Quasi Newton's method* $\quad \mathbf{p}_k = -B_k \nabla f(\mathbf{x}_k)$, where $B_k \approx [\nabla^2 f(\mathbf{x}_k)]^{-1}$.
4. *Conjugate Gradient Method* $\quad \mathbf{p}_k = -(\mathbf{x}_k - \mathbf{x}_{k-1} + \beta_k \nabla f(\mathbf{x}_k))$, 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.

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:

$$f(\mathbf{x}_k + \eta_k \mathbf{p}_k) \le f(\mathbf{x}_k) + c_1 \eta_k \langle \nabla f(\mathbf{x}_k),\mathbf{p}_k\rangle, \quad c_1 \in (0,1). \tag{3.1a}$$

The constant c_1 is usually taken as a small number such as 10^{-4}, and (3.1a) is called *Amijo condition.* To avoid unacceptably small step sizes, the *curvature condition* requires

$$\langle \nabla f(\mathbf{x}_k + \eta_k \mathbf{p}_k), \mathbf{p}_k\rangle \ge c_2 \langle \nabla f(\mathbf{x}_k), \mathbf{p}_k\rangle, \quad c_2 \in (c_1, 1). \tag{3.1b}$$

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 (3.1b) simply requires $\phi'(\eta_k) \geq c_2\phi'(0)$, 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 (3.1) with $0 < c_1 < c_2 < 1$ are called the *Wolfe conditions.*

The following are called the *strong Wolfe conditions.*

$$f(\mathbf{x}_k + \eta\mathbf{p}_k) \leq f(\mathbf{x}_k) + c_1\eta\langle \nabla f(\mathbf{x}_k), p_k\rangle, \quad c_1 \in (0, 1). \tag{3.2a}$$

$$|\langle \nabla f(\mathbf{x}_k + \eta_k\mathbf{p}_k), \mathbf{p}_k\rangle| \leq c_2|\langle \nabla f(\mathbf{x}_k), \mathbf{p}_k\rangle|, \quad c_2 \in (c_1, 1). \tag{3.2b}$$

Lemma 3.1. *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* (3.1) *and the strong Wolfe conditions* (3.2)*.*

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 (3.1a) 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 (3.1b). Since ∇f is continuous, there is a small interval containing η_2, in which η satisfies (3.1b). 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 (3.1).

Example 3.1. *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* (3.1a) *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* (3.1)*.*

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 3.1 (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* (3.1)*. Then*

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

Proof. By (3.1b), 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 (3.1a), 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 $a_N = \sum_{k=0}^{N} \cos^2 \theta_k \|\nabla f(\mathbf{x}_k)\|^2$ 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 3.2. *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:):*

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

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

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

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

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

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

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

where $L/\mu = \|B_k\|\|B_k^{-1}\|$ *is the condition number of the Hessian. With Theorem 3.1, 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* (3.1) *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 3.1. *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*

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

3.3 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 $M > 0$:

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

   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.

3.3.1 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 $\mathbf{g}(\mathbf{t}) = \nabla \mathbf{f}(\mathbf{x}_* + \mathbf{t}(\mathbf{x_k} - \mathbf{x}_*))$, we get

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

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 3.2. *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 3.2 (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 3.1. *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

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

□

3.3.2 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 3.3 (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,

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

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 1.8 and Lipschitz continuity of the Hessian,

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

With Lemma 3.2, We also have

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

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

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

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

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

□

Part II

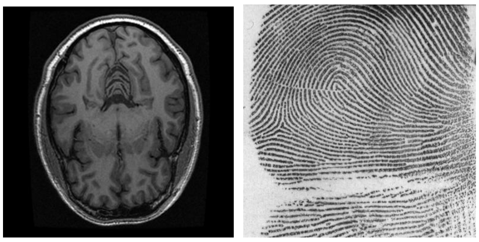
Nonsmooth problems

4

Introduction and extended convex functions

4.1 Motivation and examples

Let $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$ and $\mathbf{x} \in \mathbb{R}^m$. If $m < n$, then $A\mathbf{x} = b$ can have multiple solutions. In many applications such as compressing data, a sparse solution is often needed.



(a) MRI Image (b) Finger Print

Figure 4.1: Many images (or data) have *sparsity*, e.g., it seems unnecessary to store all pixels in an image to store all the information it contains.

Consider the images in Figure 4.1 as an example. The most intuitive Euclidean basis for representing the images is highly redundant, i.e., it is unnecessary to store all pixels in an image to store all the information the image contains. If there is a sparse representation of the data under certain transform, then advantages are gained in compression (e.g., JPEG), interpolation, computation and etc.

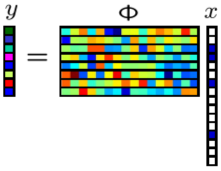
For compressing an image $b \in \mathbb{R}^m$, we consider a *frame* of $\mathbb{R}^m$, which is defined as any spanning set of $\mathbb{R}^m$. For instance, a basis is always a frame, but a frame may not be a basis. In particular, we consider the following setup:

- A redundant frame $\{\phi_i \in \mathbb{R}^m : i = 1, \cdots, n\}$ with $n > m$, consists of many sets of different bases (e.g., from computational harmonic analysis) such as Discrete Fourier Transform, Wavelets, Wavelet Packets, Gabor Transforms, etc. The elements in the frame are also called *atoms.*

- Any vector $y \in \mathbb{R}^m$ can be spanned by the atoms ϕ_i:

$$y = \sum_{i=1}^{m} \phi_i x_i = \Phi \mathbf{x},$$

where $\mathbf{x} = (x_1, \cdots, x_n)^T$ is a coefficient and $\Phi = (\phi_1, \cdots, \phi_n)$.



- The coefficient $\mathbf{x}$ is not unique because the linear system is underdetermined. Ideally, we want to find the coefficient with the smallest $||\mathbf{x}||_0$ (the number of nonzero entries) for compression.

So it motivates the following ℓ^0-minimization

$$\min_{\mathbf{x}} ||\mathbf{x}||_0, \quad \mathbf{x} \quad \text{satisfies} \quad A\mathbf{x} = b, \tag{4.1}$$

which is unfortunately an NP-hard problem due to the nonconvexity of $||\mathbf{x}||_0$.

4.1.1 ℓ^1-norm minimization

The basis pursuit [4] solves the following convex minimization problem:

$$\text{(Basis Pursuit)} \quad \min_{\mathbf{x}} ||\mathbf{x}||_1, \quad \mathbf{x} \quad \text{satisfies} \quad A\mathbf{x} = b. \tag{4.2}$$

The ℓ^1-minimization (4.2) is the convex relaxation of (4.1) and it can be proven to have the same minimizer as the NP-hard ℓ^0-minimization (4.1) for very special problems, e.g., A is a Gaussian random matrix with suitable scaling for n w.r.t. m, see [1]. In applications, the ℓ^1-minimization (4.2) produces sparse minimizers, which are quite useful, even if they are different from the true minimizer of ℓ^0-minimization (4.1), e.g., a good compression does not have to be the best possible compression.

In this chapter, we will use the hard-constrained ℓ^1-minimization (4.2) as an example for discussing optimization algorithms, we will also consider two *easier* problems:

$$\text{(LASSO)} \quad \min_{\mathbf{x}} \mu||\mathbf{x}||_1 + \frac{1}{2}||A\mathbf{x} - b||_2^2, \tag{4.3}$$

and

$$\min_{\mathbf{x}} ||\mathbf{x}||_1 + \frac{1}{2\alpha}||\mathbf{x}||_2^2, \quad \mathbf{x} \quad \text{satisfies} \quad A\mathbf{x} = b. \tag{4.4}$$

The LASSO problem (4.3) was introduced in [16]. It is proven in [18] that the minimizer (4.4) for large enough α minimizes (4.2).

4.1.2 Total variation minimization

The total variation (TV) norm minimization represents a similar yet more chanllenging nonsmooth convex minimization problem. As an example, we will consider the TV norm minimization for image denoising [13, 15].

For simplicity, here we first describe it for a continuum setup. Consider a rectangular domain $\Omega = [0,1] \times [0,1]$, and a function $u(x,y) \in H^1(\Omega)$ (differentiable functions), which represents an image with infinite resolution. Then its total variation norm can be defined as

- isotropic TV-norm: $\|u\|_{TV} = \iint_\Omega \sqrt{|u_x|^2 + |u_y|^2} dxdy$.
- anisotroric TV-norm: $\|u\|_{TV} = \iint_\Omega |u_x| + |u_y| dxdy$

With L^2-norm as $\|u\|_{L^2} = \sqrt{\iint_\Omega |u|^2 dxdy}$, for a given $a(x,y)$, the ROF (Rudin, Osher, and Fatemi, 1992) model [13] is to minimize (over u in a proper function space)

$$\|u\|_{TV} + \frac{1}{2}\lambda\|u - a\|_{L^2}^2,$$

where λ is a fixed parameter.

Notice that the TV-norm contains the absolute value function thus it is similar to the ℓ^1-minimization. For certain noisy images, e.g., Gaussian noise, with suitable λ, the ROF model using isotropic TV-norm can work well, see Figure 8.2 and Figure 8.3.

4.1.3 Constrained minimization

Consider a constrained minimization

$$\min_{\mathbf{x}\in S} f(\mathbf{x}),$$

where C is a convex, e.g., a plane $S = \{\mathbf{x} \in \mathbb{R}^n : A\mathbf{x} = b\}$, or a simplex $S = \{\mathbf{x} \in \mathbb{R}^n : x_i \geq 0, \sum_i x_i \leq 1\}$, or a cone $S = \{\mathbf{x} = (x,y,z) : z \geq \sqrt{x^2 + y^2}\}$, etc.

The *indicator* function for a set S is defined as

$$\iota_S(\mathbf{x}) = \begin{cases} 0, & \mathbf{x} \in S \\ +\infty, & \mathbf{x} \notin S \end{cases}. \tag{4.5}$$

Then the constrained minimization is equivalent to an unconstrained problem

$$\min_{\mathbf{x}} f(\mathbf{x}) + \iota_S(\mathbf{x}).$$

Such an unconstrained problem is not easier to solve due to the discontinuity of the indicator function, but it is another example of nonsmooth convex minimization.

4.2 More on convex functions

We include some useful properties about convex functions, with some proof. The other omiited proof can be found in classical references such as [12].

4.2.1 Epigraph and continuity

Recall that a function is convex if and only if $f((1-a)\mathbf{x} + a\mathbf{y}) \leq (1-a)f(\mathbf{x}) + af(\mathbf{y})$ for any $a \in [0,1]$.

Definition 4.1. *The epigraph of a function $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ is defined as the set $\{(\mathbf{x}, y) \in \mathbb{R}^n \times \mathbb{R} : y \geq f(\mathbf{x})\}$.*

Definition 4.2. *A set $S \in \mathbb{R}^n$ is convex if $(1-a)\mathbf{x}+\mathbf{y} \in S$ for any $a \in [0,1]$ and any $\mathbf{x}, \mathbf{y} \in S$.*

Theorem 4.1. *For a function $f : \mathbb{R}^n \longrightarrow \mathbb{R}$, it is convex if and only if its epigraph is a convex set.*

Theorem 4.2. *If $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ and $f_i : \mathbb{R}^n \longrightarrow \mathbb{R}$ for $i = 1, \cdots, N$ are convex, then*

1. *$g = f(A\mathbf{x} + b)$ is convex where A is a matrix.*

2. *$g = \sum\limits_{i=1}^{N} f_i(\mathbf{x})$ is convex.*

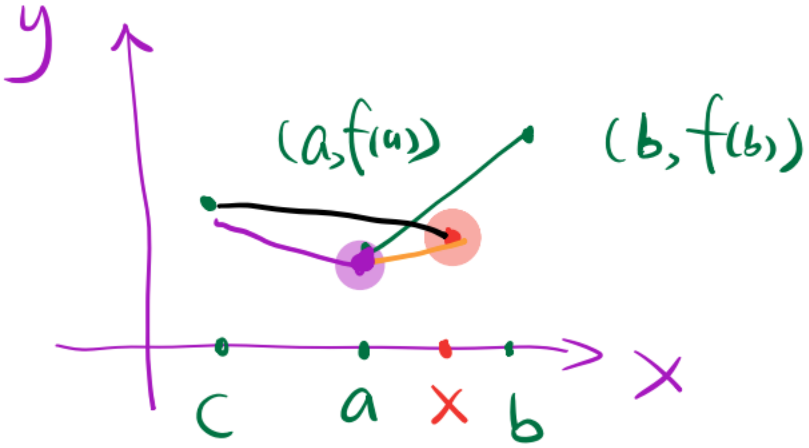
3. *$g = \max_i f_i(\mathbf{x})$ is convex.*

Proof. The first two can be checked by the definition of a convex function. The last one can be verified via the epigraph. □

Theorem 4.3. *If $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ is convex, then f is continuous. Moreover, it is also locally Lipschitz continuous, which means that for any $\mathbf{x}_0$, there is ball centered at $\mathbf{x}_0$ with radius $\delta > 0$, such that for any $\mathbf{x}$ in this ball, $|f(\mathbf{x}) - f(\mathbf{x}_0)| \leq L|\mathbf{x} - \mathbf{x}_0|$ for some constant L.*

Proof. We only consider the proof of continuity for the single variable case. See Corollary 10.1.1 in [12] for the general case.

So for a convex function $f(x)$, we want to show $|f(x) - f(a)| \to 0$ as $x \to a$. We discuss it for two cases: $x > a$ and $x < a$.



First, assume $x > a$, then consider b, c such that $c < a < x < b$. The convexity implies that the point $(x, f(x))$ is below the segment connecting $(a, f(a))$ and $(b, f(b))$ thus the slopes of two segments give

$$\frac{f(x) - f(a)}{x - a} \leq \frac{f(b) - f(a)}{b - a}.$$

Similarly, we get

$$\frac{f(x) - f(a)}{x - a} \geq \frac{f(c) - f(a)}{c - a},$$

thus

$$(x-a)\frac{f(c) - f(a)}{c - a} \leq f(x) - f(a) \leq (x-a)\frac{f(b) - f(a)}{b - a}.$$

By fixing b, c, letting $x \to a$, we get $f(x) \to f(a)$.

The second case of $x < a$ can be similarly shown, thus f is continuous at a. □

4.2.2 Subgradient and subdifferential

Definition 4.3. *For a function* $f : \mathbb{R}^n \longrightarrow \mathbb{R}$*, a vector* $\mathbf{v} \in \mathbb{R}^n$ *is a subgradient (or subderivative) of* $f(\mathbf{x})$ *at* $\mathbf{x}$ *if*

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \mathbf{v}, \mathbf{y} - \mathbf{x} \rangle, \quad \forall \mathbf{y} \in \mathbb{R}^n.$$

The derivative of a smooth $f(x)$ is the slope of the tangent line The definition simply means that the subderivative is the slope of a line which lies below the function graph. For example, for $f(x) = |x|$, at $x = 0$, any line passing $(0, 0)$ with a slope between -1 and 1 lies below the function graph, thus any number in $[-1, 1]$ is a subderivative.

Definition 4.4. *The set of all subderivatives of a function f at $\mathbf{x}$ is called the subdifferential of f at $\mathbf{x}$, denoted as $\partial f(\mathbf{x})$:*

$$\partial f(\mathbf{x}) = \{\mathbf{v} \in \mathbb{R}^n : f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \mathbf{v}, \mathbf{y} - \mathbf{x} \rangle, \quad \forall \mathbf{y} \in \mathbb{R}^n\}.$$

Theorem 4.4. *A function $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ is convex if and only if it has a subderivative at any $\mathbf{x} \in \mathbb{R}^n$.*

Proof. We only prove the only if direction. Let $\mathbf{z} = (1-a)\mathbf{x} + a\mathbf{y}$ with $a \in [0,1]$. Let $\mathbf{v}$ be a subderivative at $\mathbf{x}$, then

$$f(\mathbf{x}) \geq f(\mathbf{z}) + \langle \mathbf{v}, \mathbf{x} - \mathbf{z} \rangle = f(\mathbf{z}) - a\langle \mathbf{v}, \mathbf{y} - \mathbf{x} \rangle,$$

$$f(\mathbf{y}) \geq f(\mathbf{z}) + \langle \mathbf{v}, \mathbf{y} - \mathbf{z} \rangle = f(\mathbf{z}) + (1-a)\langle \mathbf{v}, \mathbf{y} - \mathbf{x} \rangle.$$

Adding them, we get

$$(1-a)f(\mathbf{x}) + af(\mathbf{y}) \geq f(\mathbf{z}) = f((1-a)\mathbf{x} + a\mathbf{y}).$$

□

If a function has subderivatives at $\mathbf{x}$, it is called *subdifferentiable* at $\mathbf{x}$.

Theorem 4.5. *For convex functions $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ and $g : \mathbb{R}^n \longrightarrow \mathbb{R}$,*

1. *f is differentiable implies f is subdifferentiable and $\partial f(\mathbf{x}) = \{\nabla f(\mathbf{x})\}$.*
2. *If the $\partial f(\mathbf{x})$ contains only one element, then f is differentiable at $\mathbf{x}$.*
3. *$\partial[af] = a[\partial f], \quad a \in \mathbb{R}$.*
4. *$\partial[f+g] = \partial f + \partial g$.*

So the subderivative is natural generalization of the derivative. Be aware that the subderivative is NOT the *weak derivative* defined in Sobolev spaces.

Example 4.1. *For $f(x) = |x|$,*

$$\partial f(x) = \begin{cases} \{1\}, & x > 0 \\ \{-1\}, & x < 0 \\ [-1,1], & x = 0 \end{cases}.$$

For $f(\mathbf{x}) = \|\mathbf{x}\|_1 = \sum\limits_{i=1}^{n} x_i$,

$$\forall \mathbf{v} \in \partial f(\mathbf{x}), \quad v_i = \begin{cases} \{1\}, & x_i > 0 \\ \{-1\}, & x_i < 0 \\ [-1,1], & x_i = 0 \end{cases}.$$

4.3 Extended real-valued functions

Notice that the indicator function (4.5) is NOT a function $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ thus the results in the previous subsection may not apply to (4.5). To this end, we define an *extended real-valued function* as a mapping $f : \mathbb{R}^n \longrightarrow \mathbb{R} \cup \{\pm\infty\}$.

4.3.1 Proper, convex, and closed functions

Definition 4.5. *An extended real-valued function $f : \mathbb{R}^n \longrightarrow \mathbb{R} \cup \{\pm\infty\}$ is called convex if its epigraph is a convex set.*

Definition 4.6. *Define the domain of an extended function $f : \mathbb{R}^n \longrightarrow \mathbb{R} \cup \{\pm\infty\}$ as*

$$\mathrm{dom}(f) = \{\mathbf{x} \in \mathbb{R}^n : f(\mathbf{x}) \in \mathbb{R}\}.$$

Definition 4.7. *An extended function $f : \mathbb{R}^n \longrightarrow \mathbb{R} \cup \{\pm\infty\}$ is called proper if it does not attain the value $-\infty$ and there exists at least one $\mathbf{x}$ such that $f(\mathbf{x})$ is a real number, i.e., its domain* $\mathrm{dom}(f)$ *is nonempty.*

Definition 4.8. *An extended function $f : \mathbb{R}^n \longrightarrow \mathbb{R} \cup \{\pm\infty\}$ is called closed if its epigraph is a closed set.*

A closed function is also lower semicontinuous:

Theorem 4.6. *For an extended real-valued function $f : \mathbb{R}^n \longrightarrow \mathbb{R} \cup \{\pm\infty\}$, the following are equivalent:*

1. *f is a closed function.*
2. *The level set $\{\mathbf{x} : f(\mathbf{x}) \leq a\}$ is a closed set for any $a \in$.*
3. *f is* **lower semicontinuous***: for any $\mathbf{x}$, for any sequence $\mathbf{x}_n \to \mathbf{x}$,*

$$f(\mathbf{x}) \leq \liminf_{n\to\infty} f(\mathbf{x}_n).$$

Theorem 4.7. *A proper (extended) function $f : \mathbb{R}^n \longrightarrow \mathbb{R} \cup \{+\infty\}$ is convex if and only if* $\mathrm{dom}(f)$ *is convex and*

$$f(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}) \leq \lambda f(\mathbf{x}) + (1-\lambda)f(\mathbf{y}), \quad \forall \mathbf{x}, \mathbf{y} \in \mathrm{dom}(f), \lambda \in (0,1).$$

A *proper convex function* can also be equivalently defined as:

Definition 4.9. *A proper convex function $f : \mathbb{R}^n \longrightarrow (-\infty, +\infty]$ is a function defined for any $\mathbf{x} \in \mathbb{R}^n$, not identically $+\infty$, satisfying* $\mathrm{dom}(f)$ *is convex and*

$$f(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}) \leq \lambda f(\mathbf{x}) + (1-\lambda)f(\mathbf{y}), \quad \forall \mathbf{x}, \mathbf{y} \in \mathrm{dom}(f), \lambda \in (0,1).$$

For example, an indicator function $\iota_S(\mathbf{x}) = \begin{cases} 0, & \mathbf{x} \in S \\ +\infty, & \mathbf{x} \notin S \end{cases}$ is convex if S is a convex set. And ι_S is a closed function if and only if S is a closed set. So ι_S is lower semicontinuous if and only if S is a convex set.

4.3.2 Existence and boundness of subderivatives

All results in this subsection can be found in [3, Chapter 3].

Theorem 4.8. *For a proper extended function* $f : \mathbb{R}^n \longrightarrow (-\infty, +\infty]$*, for any* $\mathbf{x}$*,* $\partial f(\mathbf{x})$ *is either a closed and convex set or an empty set.*

Theorem 4.9. *Consider a proper extended function* $f : \mathbb{R}^n \longrightarrow (-\infty, +\infty]$*.* **Assume its domain** $\mathrm{dom}(f)$ **is a convex set***, then*

1. *Existence of subderivatives at any* $\mathbf{x} \in \mathrm{dom}(f)$ *implies convexity of* $f(\mathbf{x})$*.*

2. *Convexity of* f *implies that subderivative exists at any* $\mathbf{x}$ *in the interior of* $\mathrm{dom}(f)$*, denoted as* $\mathrm{int}(\mathrm{dom}(f))$ *and* $\partial f(\mathbf{x})$ *is bounded:*

$$\forall \mathbf{v} \in \partial f(\mathbf{x}), \|\mathbf{v}\| \leq C \quad \textit{for some} \quad C.$$

3. *If* $U \subset \mathrm{int}(\mathrm{dom}(f))$ *is a nonempty compact set (bounded and closed set in* $\mathbb{R}^n$*), then convexity of* f *implies* $\bigcup_{\mathbf{x} \in U} \partial f(\mathbf{x})$ *is bounded (all subderivatives in* U *have a uniform bound).*

4. *For boundary points of* $\mathrm{dom}(f)$ *of a convex function, subderivatives exist at the relative interior of* $\mathrm{dom}(f)$ *but they can be unbounded.*

Remark 4.1. *The subdifferential can be extended to functions defined on infinite-dimensional Banach space [11]. See also [2, Chapter 16] for subdifferentials of a lower semicontinuous proper convex function.*

4.3.3 Strong convexity

Definition 4.10. *An extended function* $f : \mathbb{R}^n \longrightarrow \mathbb{R} \cup \{+\infty\}$ *is* μ*-strongly convex if* $\mathrm{dom}(f)$ *is convex and the following holds for any* $\lambda \in (0, 1)$*:*

$$f(\lambda \mathbf{x} + (1-\lambda)\mathbf{y}) \leq \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y}) - \frac{\mu}{2} \lambda (1-\lambda) \|\mathbf{x} - \mathbf{y}\|^2, \quad \forall \mathbf{x}, \mathbf{y} \in \mathrm{dom}(f).$$

Theorem 4.10. *An extended function* $f : \mathbb{R}^n \longrightarrow \mathbb{R} \cup \{+\infty\}$ *is* μ*-strongly convex if and only if* $f(\mathbf{x}) - \frac{\mu}{2}\|\mathbf{x}\|^2$ *is convex.*

Even though $\partial f(\mathbf{x})$ **denotes a set, for simplicity, we often abuse the notation by using it to denote any element in this set.** Lemma 1.1 can be extended as (see [3, Theorem 5.24] for the proof):

Lemma 4.1. *For a proper function* $f(\mathbf{x}) : \mathbb{R}^n \to (-\infty, +\infty]$*,* **assume** $\mathrm{dom}(f)$ **is convex***, then the following are equivalent:*

1. $f(\lambda \mathbf{x} + (1-\lambda)\mathbf{y}) \leq \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y}), \quad \forall \mathbf{x}, \mathbf{y} \in \mathrm{dom}(f), \lambda \in (0, 1).$

2. $f(\mathbf{x}) \geq f(\mathbf{y}) + \langle \partial f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle, \quad \forall \mathbf{x}, \mathbf{y} \in \mathrm{dom}(f).$

3. $\langle \partial f(\mathbf{y}) - \partial f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \geq 0, \quad \forall \mathbf{x}, \mathbf{y} \in \mathrm{dom}(f).$

For **a proper closed and convex** *function, the following are equivalent:*

1. *f is μ-strongly convexity*
2. $f(\mathbf{x}) \geq f(\mathbf{y}) + \langle \partial f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle + \frac{\mu}{2}\|\mathbf{x} - \mathbf{y}\|^2, \quad \forall \mathbf{x}, \mathbf{y} \in \mathrm{dom}(f).$
3. $\langle \partial f(\mathbf{y}) - \partial f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \geq \mu\|\mathbf{x} - \mathbf{y}\|^2, \quad \forall \mathbf{x}, \mathbf{y} \in \mathrm{dom}(f).$

An example is that $\iota_S + \frac{\mu}{2}\|\mathbf{x}\|^2$ is μ-strongly convex for a convex set S.

4.4 Optimality conditions

For a function $f : \mathbb{R}^n \longrightarrow \mathbb{R}$, we have:

Theorem 4.11. *For a convex function $f : \mathbb{R}^n \longrightarrow \mathbb{R}$, $\mathbf{x}_*$ minimizes $f(\mathbf{x})$ if and only if $\mathbf{0} \in \partial f(\mathbf{x}_*)$.*

Proof. $f(\mathbf{x}) \geq f(\mathbf{x}_*) = f(\mathbf{x}_*) + \langle \mathbf{0}, \mathbf{x} - \mathbf{x}_* \rangle \Leftrightarrow \mathbf{0} \in \partial f(\mathbf{x}_*).$ □

For a proper extended function such as $\iota_S(\mathbf{x}) + \frac{\mu}{2}\|\mathbf{x}\|^2$ (4.5), we have

Theorem 4.12. *For a proper convex function $f : \mathbb{R}^n \longrightarrow \mathbb{R} \cup \{+\infty\}$, $\mathbf{x}_*$ minimizes $f(\mathbf{x})$ if and only if $\mathbf{0} \in \partial f(\mathbf{x}_*)$.*

Proof. $f(\mathbf{x}) \geq f(\mathbf{x}_*) = f(\mathbf{x}_*) + \langle \mathbf{0}, \mathbf{x} - \mathbf{x}_* \rangle \Leftrightarrow \mathbf{0} \in \partial f(\mathbf{x}_*), \quad \forall \mathbf{x} \in \mathrm{dom}(f).$ □

Theorem 4.13. *For a* **proper closed** *function $f : \mathbb{R}^n \longrightarrow \mathbb{R} \cup \{+\infty\}$, if it is μ-strongly convex, then it has a unique minimizer $\mathbf{x}_*$.*

For the indicator function ι_S, its subdifferential is given as

$$\partial \iota(\mathbf{x}) = N_S(\mathbf{x}), \quad \mathbf{x} \in S,$$

where N_S is the *normal cone* of S:

Definition 4.11. *For a set $S \subset \mathbb{R}^n$ and $\mathbf{x} \in S$, the normal cone of S at $\mathbf{x}$ is*

$$N_S(\mathbf{x}) = \{\mathbf{y} : \langle \mathbf{y}, \mathbf{z} - \mathbf{x} \rangle \leq 0, \forall \mathbf{z} \in S\},$$

and $N_S(\mathbf{x})$ is an empty set for $\mathbf{x} \notin S$.

For example, for the basis pursuit problem (4.4), it can be written as

$$\min_{\mathbf{x}} ||\mathbf{x}||_1 + \frac{1}{2\alpha}\|\mathbf{x}\|^2 + \iota_{\{\mathbf{x}:A\mathbf{x}=b\}}. \tag{4.6}$$

The subdifferentials of the first two terms have been given in previous sections. For the affine set $S = \{\mathbf{x} : A\mathbf{x} = b\}$ with $A = \mathbb{R}^{m \times n}$, the normal cone is simply $N_S(\mathbf{x}) = \{A^T\mathbf{y} : \forall \mathbf{y} \in \mathbb{R}^m\}$, which is the *row space* of the matrix A.

The function in (4.6) is a proper convex function since the affine set $S = \{\mathbf{x} : A\mathbf{x} = b\}$ is convex. Notice that $S = \{\mathbf{x} : A\mathbf{x} = b\}$ is also a closed set, thus the function is also a proper closed (also lower semicontinuous) convex function, and by Theorem 4.10, it is also $\frac{1}{\alpha}$-strongly convex. Thus it has a unique minimizer.

Problem 4.1. *Derive the subdifferential for the indicator function of a closed unit ball* $S = \{\mathbf{x} \in \mathbb{R}^n : \langle \mathbf{x}, \mathbf{x} \rangle = 1\}$.

5

Subgradient and proximal gradient methods

In this chapter, we introduce some basic algorithms. A suitable application would be the ℓ^1-minimization problem (4.3):

$$\min_{\mathbf{x}} \mu||\mathbf{x}||_1 + \frac{1}{2}||A\mathbf{x} - b||_2^2.$$

5.1 Subgradient method

Now consider minimizing a convex function which is only subdifferentiable but not differentiable. By Theorem 4.11, to find a minimizer of a convex functin $f(\mathbf{x})$, we only need to find a *critical point* by solving an inclusion equation:

$$\mathbf{0} \in \partial f(\mathbf{x}_*),$$

which can be approximated by various iterative schemes.

The simplest method is to use the subgradient method:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \eta_k \mathbf{v}_k, \quad \mathbf{v}_k \in \partial f(\mathbf{x}_k), \tag{5.1}$$

where $\eta_k > 0$ is some step size and $\mathbf{v}_k$ can be chosen as any subderivative in the set $\partial f(\mathbf{x}_k)$.

5.1.1 Convergence of subgradient method

Theorem 5.1. *For the subgradient method*

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \eta_k \frac{\mathbf{v}_k}{\|\mathbf{v}_k\|}, \quad \textit{some} \quad \mathbf{v}_k \in \partial f(\mathbf{x}_k,)$$

assume the boundedness of subderivatives $\|\mathbf{v}_k\| \leq M, \forall k$. *Define*

$$\bar{\mathbf{x}}_k = \operatorname*{argmin}_{1\leq i\leq k} f(\mathbf{x}_k).$$

For a proper convex function $f(\mathbf{x})$ with at least one global minimizer $\mathbf{x}_*$, the following holds:

1. *If $\sum\limits_{k=0}^{\infty} \eta_k = +\infty$ and $\sum\limits_{k=0}^{\infty} \eta_k^2 < +\infty$, then $f(\bar{\mathbf{x}}_k) - f(\mathbf{x}_*) \to 0$ as $k \to \infty$.*
2. *With a step size $\eta_k \equiv \frac{C}{\sqrt{n+1}}$ for $k = 0, 1, \cdots, n$, then*
$$f(\bar{\mathbf{x}}_k) - f(\mathbf{x}_*) = \mathcal{O}(\frac{1}{\sqrt{k}}).$$
3. *With the Polyak's step size $\eta_k = \frac{f(\mathbf{x}_k) - f(\mathbf{x}_*)}{\|\mathbf{v}_k\|}$:*
$$\|\mathbf{x}_{k+1} - \mathbf{x}_*\| \leq \|\mathbf{x}_k - \mathbf{x}_*\|, \quad f(\bar{\mathbf{x}}_k) - f(\mathbf{x}_*) = \mathcal{O}(\frac{1}{\sqrt{k}})$$
.

If further assuming $f(\mathbf{x})$ is μ-strongly convex and **closed (or equivalently lower semicontinuous)***, then with the step size rule $\eta_k = \frac{2}{\mu(k+1)}\|\mathbf{v}_k\|$,*

4. *$f(\bar{\mathbf{x}}_k) - f(\mathbf{x}_*) = \mathcal{O}(\frac{1}{k})$ and $\|\mathbf{x}_k - \mathbf{x}_*\| = \mathcal{O}(\frac{1}{\sqrt{k}})$.*

Proof. **Step I**: By the definition of the subderivative $\mathbf{v}_k$, we have

$$f(\mathbf{x}_k) \leq f(\mathbf{x}_*) + \langle \mathbf{x}_k - \mathbf{x}_*, \mathbf{v}_k \rangle,$$

thus for $k = 0, 1, \cdots, n$,

$$\begin{aligned}
\|\mathbf{x}_{k+1} - \mathbf{x}_*\|^2 &= \left\| \mathbf{x}_k - \eta_k \frac{\mathbf{v}_k}{\|\mathbf{v}_k\|} - \mathbf{x}_* \right\|^2 \\
&= \|\mathbf{x}_k - \mathbf{x}_*\|^2 - \eta_k \frac{2}{\|\mathbf{v}_k\|} \langle \mathbf{x}_k - \mathbf{x}_*, \mathbf{v}_k \rangle + \eta_k^2 \\
&\leq \|\mathbf{x}_k - \mathbf{x}_*\|^2 - \eta_k \frac{2}{\|\mathbf{v}_k\|} (f(\mathbf{x}_k) - f(\mathbf{x}_*)) + \eta_k^2 \\
&\leq \|\mathbf{x}_k - \mathbf{x}_*\|^2 - \eta_k \frac{2}{M} (f(\bar{\mathbf{x}}_k) - f(\mathbf{x}_*)) + \eta_k^2.
\end{aligned}$$

Summing for $k = 0, 1, \cdots, n$, we get

$$\|\mathbf{x}_{n+1} - \mathbf{x}_*\|^2 \leq \|\mathbf{x}_0 - \mathbf{x}_*\|^2 - \frac{2}{M}(f(\bar{\mathbf{x}}_k) - f(\mathbf{x}_*)) \sum_{k=0}^{n} \eta_k + \sum_{k=0}^{n} \eta_k^2$$

$$\Rightarrow f(\bar{\mathbf{x}}_k) - f(\mathbf{x}_*) \leq M \frac{\|\mathbf{x}_0 - \mathbf{x}_*\|^2 + \sum\limits_{k=0}^{n} \eta_k^2}{2 \sum\limits_{k=0}^{n} \eta_k}.$$

Step II: To obtain $f(\bar{\mathbf{x}}_k) - f(\mathbf{x}_*)$ as $k \to \infty$, we need $\sum\limits_{k=0}^{\infty} \eta_k = +\infty$ and $\sum\limits_{k=0}^{\infty} \eta_k^2 = +\infty$, e.g., $\eta_k = \frac{1}{k}$ gives $\sum\limits_{k=0}^{\infty} \frac{1}{k^2} < +\infty$ and $\sum\limits_{k=0}^{\infty} \frac{1}{k} = +\infty$.

Step III: Plugging in $\eta_k \equiv \frac{C}{\sqrt{n+1}}$ for $k = 0, 1, \cdots, n$, we have

$$\begin{aligned} f(\bar{\mathbf{x}}_k) - f(\mathbf{x}_*) &\leq M \frac{\|\mathbf{x}_0 - \mathbf{x}_*\|^2 + \sum\limits_{k=0}^{n} \eta_k^2}{2\sum\limits_{k=0}^{n} \eta_k} = M \frac{\|\mathbf{x}_0 - \mathbf{x}_*\|^2 + \sum\limits_{k=0}^{n} \frac{C^2}{n+1}}{2\sum\limits_{k=0}^{n} \frac{C}{\sqrt{n+1}}} \\ &= M \frac{\|\mathbf{x}_0 - \mathbf{x}_*\|^2 + C^2}{2C\sqrt{n+1}}. \end{aligned}$$

Step IV: For the Polyak's step size $\eta_k = \frac{f(\mathbf{x}_k) - f(\mathbf{x}_*)}{\|\mathbf{v}_k\|}$,

$$\begin{aligned} \|\mathbf{x}_{k+1} - \mathbf{x}_*\|^2 &\leq \|\mathbf{x}_k - \mathbf{x}_*\|^2 - \eta_k \frac{2}{\|\mathbf{v}_k\|}(f(\mathbf{x}_k) - f(\mathbf{x}_*)) + \eta_k^2 \\ &= \|\mathbf{x}_k - \mathbf{x}_*\|^2 - \frac{|f(\mathbf{x}_k) - f(\mathbf{x}_*)|^2}{\|\mathbf{v}_k\|^2} \\ &\leq \|\mathbf{x}_k - \mathbf{x}_*\|^2 - \frac{|f(\mathbf{x}_k) - f(\mathbf{x}_*)|^2}{M^2}. \end{aligned}$$

Summing for $k = 0, 1, \cdots, n$, we get

$$\frac{1}{M^2} \sum_{k=0}^{n} |f(\mathbf{x}_k) - f(\mathbf{x}_*)|^2 \leq \|\mathbf{x}_0 - \mathbf{x}_*\|^2 - \|\mathbf{x}_{n+1} - \mathbf{x}_*\|^2 \leq \|\mathbf{x}_0 - \mathbf{x}_*\|^2,$$

$$\Rightarrow f(\bar{\mathbf{x}}_k) - f(\mathbf{x}_*) \leq \frac{M}{\sqrt{n+1}} \|\mathbf{x}_0 - \mathbf{x}_*\|.$$

Step V: Now assume $f(\mathbf{x})$ is strongly convex, by Lemma 4.1, the strong convexity gives

$$f(\mathbf{x}_*) \geq f(\mathbf{x}_k) + \langle \partial f(\mathbf{x}_k), \mathbf{x}_* - \mathbf{x}_k \rangle + \frac{\mu}{2} \|\mathbf{x}_* - \mathbf{x}_k\|^2,$$

thus

$$-\langle \partial f(\mathbf{x}_k), \mathbf{x}_k - \mathbf{x}_* \rangle \leq f(\mathbf{x}_*) - f(\mathbf{x}_k) - \frac{\mu}{2} \|\mathbf{x}_* - \mathbf{x}_k\|^2$$

and

$$\begin{aligned} \|\mathbf{x}_{k+1} - \mathbf{x}_*\|^2 &= \left\| \mathbf{x}_k - \eta_k \frac{\mathbf{v}_k}{\|\mathbf{v}_k\|} - \mathbf{x}_* \right\|^2 \\ &= \|\mathbf{x}_k - \mathbf{x}_*\|^2 - \eta_k \frac{2}{\|\mathbf{v}_k\|} \langle \mathbf{x}_k - \mathbf{x}_*, \mathbf{v}_k \rangle + \eta_k^2 \\ &= \|\mathbf{x}_k - \mathbf{x}_*\|^2 - \eta_k \frac{2}{\|\mathbf{v}_k\|} \langle \mathbf{v}_k, \mathbf{x}_k - \mathbf{x}_* \rangle + \eta_k^2 \\ &\leq \left(1 - 2\eta_k \frac{1}{\|\mathbf{v}_k\|} \frac{\mu}{2}\right) \|\mathbf{x}_k - \mathbf{x}_*\|^2 - 2\eta_k \frac{1}{\|\mathbf{v}_k\|}][f(\mathbf{x}_k) - f(\mathbf{x}_*)] + \eta_k^2. \end{aligned}$$

We get

$$\begin{aligned} f(\mathbf{x}_k) - f(\mathbf{x}_*) &\leq \frac{\|\mathbf{v}_k\|}{2\eta_k}\|\mathbf{x}_k - \mathbf{x}_*\|^2 + \left(\frac{\|\mathbf{v}_k\|}{2\eta_k} - \frac{2}{\mu}\right)\|\mathbf{x}_{k+1} - \mathbf{x}_*\|^2 + \frac{\|\mathbf{v}_k\|}{2}\eta_k \\ &= \frac{\mu(k-1)}{4}\|\mathbf{x}_k - \mathbf{x}_*\|^2 - \frac{\mu(k+1)}{4}\|\mathbf{x}_{k+1} - \mathbf{x}_*\|^2 + \frac{1}{\mu(k+1)}\|\mathbf{v}_k\|^2, \end{aligned}$$

and

$$k[f(\mathbf{x}_k) - f(\mathbf{x}_*)] \leq \frac{\mu k(k-1)}{4}\|\mathbf{x}_k - \mathbf{x}_*\|^2 - \frac{\mu k(k+1)}{4}\|\mathbf{x}_{k+1} - \mathbf{x}_*\|^2 + \frac{k}{\mu(k+1)}\|\mathbf{v}_k\|^2.$$

Summing it for $k = 0, 1, \cdots, n$, we get

$$\sum_{k=0}^{n} k[f(\mathbf{x}_k) - f(\mathbf{x}_*)] \leq -\frac{\mu n(n+1)}{4}\|\mathbf{x}_{n+1} - \mathbf{x}_*\|^2 + \frac{M^2}{\mu}\sum_{k=0}^{n}\frac{k}{k+1} \leq \frac{M^2}{\mu}\sum_{k=0}^{n}\frac{k}{k+1}$$

$$\Rightarrow \frac{n(n+1)}{2}[f(\bar{\mathbf{x}}_k) - f(\mathbf{x}_*)] = \sum_{k=0}^{n} k[f(\bar{\mathbf{x}}_k) - f(\mathbf{x}_*)] \leq \frac{M^2}{\mu}\sum_{k=0}^{n}\frac{k}{k+1} \leq \frac{M^2}{\mu}n,$$

which gives $f(\bar{\mathbf{x}}_k) - f(\mathbf{x}_*) \leq \frac{2M^2}{\mu}\frac{1}{n+1}$.

Finally we also have $\frac{\mu}{2}\|\bar{\mathbf{x}}_k - \mathbf{x}_*\|^2 \leq f(\bar{\mathbf{x}}_k) - f(\mathbf{x}_*)$, thus $\|\bar{\mathbf{x}}_k - \mathbf{x}_*\| \leq \frac{2M}{\mu}\frac{1}{\sqrt{n+1}}$. □

5.1.2 Polyak's step size and Fejér monotonicity

In Theorem 5.1, without the strong convexity, the Polyay's step size ensures $\|\mathbf{x}_{k+1} - \mathbf{x}_*\| \leq \|\mathbf{x}_k - \mathbf{x}_*\|$ for one global minimizer $\mathbf{x}_*$, which is called *Fejér monotonicity*:

Definition 5.1. *A sequence $\{\mathbf{x}_k\}_{k=0}^{\infty}$ is called Fejér monotone w.r.t. a set S if there is one $\mathbf{y} \in S$ such that $\|\mathbf{x}_{k+1} - \mathbf{y}\| \leq \|\mathbf{x}_k - \mathbf{y}\|$ for any y.*

Let X_* be the set of all global minimizers of the convex function in Theorem 5.1, then $\{\mathbf{x}_k\}_{k=0}^{\infty}$ is Fejér monotone w.r.t. X_*, which does not imply $\mathbf{x}_k \to \mathbf{x}_*$ since $\mathbf{x}_* \in X_*$ may not be unique. On the other hand, we have (see [3, Theorem 8.17]):

Theorem 5.2. *For subgradient method with the Polyay's step size for a proper convex function in Theorem 5.1, $\mathbf{x}_k$ converges to one global minimizer of $f(\mathbf{x})$.*

5.2 Proximal point method

For solving $\mathbf{0} \in \partial f(\mathbf{x}_*)$, instead of using $\partial f(\mathbf{x}_k)$, if using the subderivative at $\mathbf{x}_{k+1}$, then the method is called *proximal point method*:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \eta_k \mathbf{v}_{k+1}, \quad \mathbf{v}_{k+1} \in \partial f(\mathbf{x}_{k+1}). \tag{5.2}$$

For readers who are familiar with ordinary differential equations (ODEs), consider approximating $u'(t) = F(u)$, then the explicit *forward Euler method* is

$$u_{k+1} = u_k + \Delta t F(u_k),$$

and implicit *backward Euler method* is

$$u_{k+1} = u_k + \Delta t F(u_{k+1}).$$

Notice that subgradient method (5.1) is similar to the forward Euler method and the proximal point method (5.2) is similar to the backward Euler method, if we set $F = -\nabla f$.

For simplicity, consider a constant step size η in (5.2). For implementation, we rewrite it as

$$\mathbf{x}_{k+1} + \eta \partial f(\mathbf{x}_{k+1}) = \mathbf{x}_k \Leftrightarrow \mathbf{x}_{k+1} = (I + \eta \partial f)^{-1}(\mathbf{x}_k).$$

By Theorem 4.12, $(I + \eta \partial f)^{-1}$ is equivalent to the following proximal operator:

Definition 5.2. *The proximal operator of a convex function $f(\mathbf{x})$ with a parameter $\gamma > 0$ is defined as the following function:*

$$\operatorname{Prox}_f^\gamma(\mathbf{x}) = \operatorname*{argmin}_{\mathbf{u}} f(\mathbf{u}) + \frac{1}{2\gamma}\|\mathbf{u} - \mathbf{x}\|^2.$$

The proximal operator $\operatorname{Prox}_f^\gamma(\mathbf{x})$ is a well defined for the following:

1. For a convex function $f : \mathbb{R}^n \longrightarrow \mathbb{R}$, $f(\mathbf{u}) + \frac{1}{2\gamma}\|\mathbf{u} - \mathbf{x}\|^2$ is a strongly convex function, so it has a unique minimizer.

2. For a **proper closed** convex function $f : \mathbb{R}^n \longrightarrow \mathbb{R} \cup \{+\infty\}$, $f(\mathbf{u}) + \frac{1}{2\gamma}\|\mathbf{u} - \mathbf{x}\|^2$ is a **proper closed** strongly convex function, so it has a unique minimizer (Theorem 4.13).

In general the function $\operatorname{Prox}_f^\gamma(\mathbf{x})$ does not have an explicit formula. But for special functions, explicit formulae are available:

1. For $f(x) = |x|$, $\operatorname{Prox}_f^\gamma(x) = \begin{cases} x - \gamma, & x > 1, \\ x + \gamma, & x < -1, \\ 0, & x \in [-1, 1]. \end{cases}$

2. For $f(\mathbf{x}) = \|\mathbf{x}\|_1$, $\operatorname{Prox}_f^\gamma(\mathbf{x}) = \mathbf{v}, \quad \mathbf{v}_i = \begin{cases} x_i - \gamma, & x_i > 1, \\ x_i + \gamma, & x_i < -1, \\ 0, & x_i \in [-1, 1]. \end{cases}$

3. For an indicator function (4.5) of a convex **closed** set S, $\operatorname{Prox}_f^\gamma(\mathbf{x})$ is the Euclidean projection of $\mathbf{x}$ to S.

Problem 5.1. *Derive the explicit proximal formulae in the examples above.*

5.2.1 The Moreau-Yosida regularization

By Theorem 4.13, we have

Definition 5.3. *The Moreau-Yosida regularization (a.k.a. Moreau envelope) of a function $f(\mathbf{x})$ is*

$$f_\eta(\mathbf{x}) = \min_{\mathbf{u}}[f(\mathbf{u}) + \frac{1}{2\eta}\|\mathbf{u}-\mathbf{x}\|^2],$$

which is well defined if $f : \mathbb{R}^n \longrightarrow \mathbb{R} \cup \{+\infty\}$ is a closed (or equivalently lower semicontinuous) proper convex function.

Recall that the proximal operator is given as

$$\text{Prox}_f^\eta(\mathbf{x}) = \operatorname*{argmin}_{\mathbf{u}}[f(\mathbf{u}) + \frac{1}{2\eta}\|\mathbf{u}-\mathbf{x}\|^2],$$

which is an operator thus different from the function $f_\eta(\mathbf{x})$.

Theorem 5.3. *For a closed (or equivalently lower semicontinuous) proper convex function $f(\mathbf{x})$, its Moreau-Yosida regularization (a.k.a. Moreau envelope) satisfies*

1. *$f_\eta(\mathbf{x})$ is convex and differentiable.*

2. *$\nabla f_\eta(\mathbf{x}) = \frac{1}{\eta}[\mathbf{x} - \text{Prox}_f^\eta(\mathbf{x})]$, $\quad \text{Prox}_f^\eta(\mathbf{x}) = \mathbf{x} - \eta\nabla f_\eta(\mathbf{x})$.*

3. *$\nabla f_\eta(\mathbf{x})$ is Lipschitz-continuous with $L = \frac{1}{\eta}$.*

Remark 5.1. *It can be proven that f is strongly convex if and only f_η is strongly convex, see [10].*

Problem 5.2. *Prove that $f_\eta(\mathbf{x})$ is convex.*

Proof. See [3, Theorem 6.60] for the differentiability. We give a brief proof of 2 and 3. Let $\mathbf{u} = \text{Prox}_f^\eta(\mathbf{x})$ and $\mathbf{v} = \text{Prox}_f^\eta(\mathbf{y})$. Since $\mathbf{u}$ is the minimizer to $g(\mathbf{z}) = f(\mathbf{z}) + \frac{1}{2\eta}\|\mathbf{z}-\mathbf{x}\|^2$, we have

$$g(\mathbf{v}) \geq g(\mathbf{u}) + \langle 0, \mathbf{v}-\mathbf{u}\rangle + \frac{1}{2\eta}\|\mathbf{v}-\mathbf{u}\|^2$$

$$\Rightarrow f(\mathbf{v}) + \frac{1}{2\eta}\|\mathbf{v}-\mathbf{x}\|^2 \geq f(\mathbf{u}) + \frac{1}{2\eta}\|\mathbf{u}-\mathbf{x}\|^2 + \frac{1}{2\eta}\|\mathbf{v}-\mathbf{u}\|^2$$

thus

$$\begin{aligned}
f_\eta(\mathbf{y}) =& f(\mathbf{v}) + \frac{1}{2\eta}\|\mathbf{v}-\mathbf{y}\|^2 \\
=& f(\mathbf{v}) + \frac{1}{2\eta}\|\mathbf{v}-\mathbf{x}\|^2 + \frac{\langle \mathbf{v}-\mathbf{x}, \mathbf{x}-\mathbf{y}\rangle}{\eta} + \frac{1}{2\eta}\|\mathbf{x}-\mathbf{y}\|^2 \\
\geq& f(\mathbf{u}) + \frac{1}{2\eta}\|\mathbf{u}-\mathbf{x}\|^2 + \frac{1}{2\eta}\|\mathbf{v}-\mathbf{u}\|^2 + \frac{\langle \mathbf{v}-\mathbf{x}, \mathbf{x}-\mathbf{y}\rangle}{\eta} + \frac{1}{2\eta}\|\mathbf{x}-\mathbf{y}\|^2 \\
=& f(\mathbf{u}) + \frac{1}{2\eta}\|\mathbf{u}-\mathbf{x}\|^2 + \langle\frac{\mathbf{x}-\mathbf{u}}{\eta}, \mathbf{y}-\mathbf{x}\rangle + \frac{1}{2\eta}\|\mathbf{x}-\mathbf{y}\|^2 \\
&+ \langle\frac{\mathbf{u}-\mathbf{v}}{\eta}, \mathbf{y}-\mathbf{x}\rangle + \frac{1}{2\eta}\|\mathbf{u}-\mathbf{v}\|^2 \\
=& f(\mathbf{u}) + \frac{1}{2\eta}\|\mathbf{u}-\mathbf{x}\|^2 + \langle\frac{\mathbf{x}-\mathbf{u}}{\eta}, \mathbf{y}-\mathbf{x}\rangle + \frac{\eta}{2}\left\|\frac{\mathbf{x}-\mathbf{y}}{\eta} - \frac{\mathbf{u}-\mathbf{v}}{\eta}\right\|^2
\end{aligned}$$

Step I: by throwing away the last quadratic term, we obtain

$$f_\eta(\mathbf{y}) \geq f_\eta(\mathbf{x}) + \langle\frac{\mathbf{x}-\mathbf{u}}{\eta}, \mathbf{y}-\mathbf{x}\rangle$$

thus $\frac{\mathbf{x}-\mathbf{u}}{\eta} = \nabla f_\eta(\mathbf{x})$ due to the differentiability of f_η.

Step II: the inequality above gives

$$f_\eta(\mathbf{y}) \geq f_\eta(\mathbf{x}) + \langle\frac{\mathbf{x}-\mathbf{u}}{\eta}, \mathbf{y}-\mathbf{x}\rangle + \frac{\eta}{2}\left\|\frac{\mathbf{x}-\mathbf{y}}{\eta} - \frac{\mathbf{u}-\mathbf{v}}{\eta}\right\|^2.$$

By switching $\mathbf{x}$ and $\mathbf{y}$, we also have

$$f_\eta(\mathbf{x}) \geq f_\eta(\mathbf{y}) + \langle\frac{\mathbf{y}-\mathbf{v}}{\eta}, \mathbf{x}-\mathbf{y}\rangle + \frac{\eta}{2}\left\|\frac{\mathbf{x}-\mathbf{y}}{\eta} - \frac{\mathbf{u}-\mathbf{v}}{\eta}\right\|^2.$$

Adding them, we get

$$\langle\frac{\mathbf{x}-\mathbf{u}}{\eta} - \frac{\mathbf{y}-\mathbf{v}}{\eta}, \mathbf{x}-\mathbf{y}\rangle \geq \eta\left\|\frac{\mathbf{x}-\mathbf{u}}{\eta} - \frac{\mathbf{y}-\mathbf{v}}{\eta}\right\|^2,$$

$$\langle\nabla f_\eta(\mathbf{x}) - \nabla f_\eta(\mathbf{y}), \mathbf{x}-\mathbf{y}\rangle \geq \eta\|\nabla f_\eta(\mathbf{x}) - \nabla f_\eta(\mathbf{y})\|^2.$$

With Cauchy-Schwartz inequality $\langle\nabla f_\eta(\mathbf{x}) - \nabla f_\eta(\mathbf{y}), \mathbf{x}-\mathbf{y}\rangle \leq \|\nabla f_\eta(\mathbf{x}) - \nabla f_\eta(\mathbf{y})\|\|\mathbf{x}-\mathbf{y}\|$, we get the Lipschitz continuity. □

5.2.2 The first convergence proof of the proximal point method

Assume $\mathbf{x}_*$ minimized a closed convex function $f(\mathbf{x})$, the Moreau-Yosida regularized function $f_\eta(\mathbf{x})$ has the same minimizer $\mathbf{x}_*$:

$$\min_{\mathbf{x}} f_\eta(\mathbf{x}) = \min_{\mathbf{x},\mathbf{u}}[f(\mathbf{u}) + \frac{1}{2\eta}\|\mathbf{x}-\mathbf{u}\|^2] \geq \min_{\mathbf{x},\mathbf{u}} f(\mathbf{u}) = f(\mathbf{x}_*),$$

where the equal sign attains if $\mathbf{u} = \mathbf{x} = \mathbf{x}_*$.

Due to the $\mathrm{Prox}_f^\eta(\mathbf{x}) = \mathbf{x} - \eta\nabla f_\eta(\mathbf{x})$, the proximal point method for minimizing $f(\mathbf{x})$ is simply the gradient descent method minimizing $f_\eta(\mathbf{x})$:

$$\mathbf{x}_{k+1} = \mathrm{Prox}_f^\eta(\mathbf{x}_k) = \mathbf{x}_k - \eta\nabla f_\eta(\mathbf{x}_k).$$

Recall that ∇f_η is Lipschitz continuous with Lipschitz constant $L = \frac{1}{\eta}$, thus the $\mathcal{O}(\frac{1}{k})$ convergence rate theorem for gradient descent (Theorem 2.3) immediately applies to the proximal point method, which is however in the form of $f_\eta(\mathbf{x}_k) - f_\eta(\mathbf{x}_*) = \mathcal{O}(\frac{1}{k})$.

If assuming strong convexity of $f(\mathbf{x})$, the linear convergence rate theorem (Theorem 2.6) for gradient descent gives a convergence rate for small enough step size.

In Section 5.4, we will prove a better result

$$\|\mathbf{x}_k - \mathbf{x}_*\|^2 \leq \left(\frac{1}{1+2\eta\mu}\right)^k \|\mathbf{x}_k - \mathbf{x}_*\|^2, \quad \forall \eta > 0.$$

5.3 The proximal gradient method

Now we consider the composite optimization problem

$$\min_{\mathbf{x}} f(\mathbf{x}) + g(\mathbf{x}),$$

where f is a closed (or equivalently lower semicontinous) proper convex function, and g is a differentiable convex function with Lipschitz continuous gradient.

Now we consider the following proximal gradient method by using the proximal operator of f and the gradient of g:

$$\mathbf{x}_{k+1} = (I + \eta\partial f)^{-1}[\mathbf{x}_k - \eta\nabla g(\mathbf{x}_k)] = \mathrm{Prox}_f^\eta[\mathbf{x}_k - \eta\nabla g(\mathbf{x}_k)].$$

Notice that the discussion in this section applies to both gradient descent method (if $f(\mathbf{x}) = 0$) and the proximal point method (if $g(\mathbf{x}) = 0$).

5.3.1 Forward-backward splitting

The proximal operator is the simple forward-backward splitting. For instance, consider solving the (or inclusion) equation for two vector valued (or set-valued) operators A and B:

$$\mathbf{0} = A(\mathbf{x}) + B(\mathbf{x}) \quad (\text{ or } \mathbf{0} \in A(\mathbf{x}) + B(\mathbf{x})),$$

The forward-backward splitting is given as

$$\mathbf{x}_{k+1} = (\text{ or } \in) \quad \mathbf{x}_k + \tau[A(\mathbf{x}_k) + B(\mathbf{x}_{k+1})].$$

Examples:

1. Plug in $\tau = -\eta$, $A(\mathbf{x}) = \partial g(\mathbf{x}) = \{\nabla g(\mathbf{x})\}$ and $B(\mathbf{x}) = \partial f(\mathbf{x})$, then we obtain the proximal gradient method.

2. In particular, if $f(\mathbf{x})$ is an indicator function of a closed convex set S, then its proximal operator $(I+\eta\partial f)^{-1}(\mathbf{x}) = \mathrm{Prox}_f^{\eta}(\mathbf{x})$ is the Euclidean projection to S, denoted as $P_S(\mathbf{x})$. In this case, the forward-backward splitting or the proximal gradient method is also called the **projected gradient method**:
$$\mathbf{x}_{k+1} = P_S[\mathbf{x}_k - \eta\nabla g(\mathbf{x}_k)],$$
which is an intuitive method for $\min_{\mathbf{x}\in S} g(\mathbf{x})$.

3. Consider solving the ODE $\mathbf{x}'(t) = A(\mathbf{x}) + B(\mathbf{x})$, then the implicit-explicit (IMEX) scheme is precisely the forward-backward splitting:
$$\mathbf{x}_{k+1} = \mathbf{x}_k + \Delta t[A(\mathbf{x}_k) + B(\mathbf{x}_{k+1})].$$
For example, for the convection diffusion equation $u_t = u_x + u_x x$, the IMEX method is a popular choice for the time discretization:
$$u^{n+1} = u^n + \Delta t[u_x^n + u_{xx}^n].$$

Notice that the gradient descent method for minimizing a function $g(\mathbf{x})$ is the same as the forward Euler method $\mathbf{x}_{k+1} = \mathbf{x}_k - \Delta t\nabla g(\mathbf{x}_k)$ for approximating ODE $\mathbf{x}'(t) = -\nabla g(\mathbf{x})$. Similarly, the proximal gradient method for composite optimization is the same as the IMEX method for ODE. Though such a connection between optimization and ODE solvers is often used to construct new methods for either optimizaiton or ODE solvers, there is a significant difference between the gradient descent method and the forward Euler method: a ODE solver is used to approximate and capture time dynamics, yet the gradient descent is to find only the minimizer, which is the equilibrium solution to the ODE $\mathbf{x}'(t) = -\nabla g(\mathbf{x})$.

5.3.2 Properties of the proximal operator

Theorem 5.4. *For a proper closed (or equivalently lower semicontinuous) convex function $f : \mathbb{R}^n \longrightarrow \mathbb{R}$, then following are equivalent:*

1. $\mathbf{u} = \mathrm{Prox}_f^{\gamma}(\mathbf{x}) = (I + \gamma\partial f)^{-1}(\mathbf{x})$.
2. $\mathbf{x} - \mathbf{u} \in \gamma\partial f(\mathbf{u})$.
3. $\frac{1}{\gamma}\langle \mathbf{x} - \mathbf{u}, \mathbf{y} - \mathbf{u}\rangle \le f(\mathbf{y}) - f(\mathbf{u}), \forall \mathbf{y}$ *or equivalently*
$$f(\mathbf{y}) \ge f(\mathbf{u}) + \langle \frac{1}{\gamma}(\mathbf{x} - \mathbf{u}), \mathbf{y} - \mathbf{u}\rangle, \quad \forall \mathbf{y}.$$

Proof. The equivalence between 1 and 2 is implied by Theorem 4.12. For 3, it simply means that $\frac{1}{\gamma}(\mathbf{x} - \mathbf{u})$ is the slope of a subtangent line. □

5.3.3 Convergence under convexity

Theorem 5.5 (Sufficient Decrease Lemma). *Assume $f : \mathbb{R}^n \longrightarrow \mathbb{R} \cup \{+\infty\}$ is a closed (or equivalently lower semicontinous) proper convex function, and $g : \mathbb{R}^n \longrightarrow \mathbb{R}$ is a differentiable convex function with ∇g being Lipschitz continuous with Lipschitz constant L. Let $F = f + g$, and $\bar{\mathbf{x}} = \mathrm{Prox}_f^\eta[\mathbf{x} - \eta\nabla g(\mathbf{x})]$, then*

$$F(\mathbf{x}) - F(\bar{\mathbf{x}}) \geq \left(\frac{1}{\eta} - \frac{L}{2}\right)\|\bar{\mathbf{x}} - \mathbf{x}\|^2.$$

Remark 5.2. *By setting $g(\mathbf{x}) \equiv 0$, we get the unconditional stability of the proximal point method $\mathbf{x}_{k+1} = (I + \eta\partial f)^{-1}(\mathbf{x}_k)$ for minimizing $f(\mathbf{x})$:*

$$f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}) \geq \frac{1}{\eta}\|\bar{\mathbf{x}}_k - \mathbf{x}_{k+1}\|^2, \quad \forall \eta > 0.$$

Proof. By Theorem 5.4, we have

$$\bar{\mathbf{x}} = \mathrm{Prox}_f^\eta[\mathbf{x} - \eta\nabla g(\mathbf{x})] \Rightarrow \frac{1}{\eta}\langle \mathbf{x} - \eta\nabla g(\mathbf{x}) - \bar{\mathbf{x}}, \mathbf{x} - \bar{\mathbf{x}}\rangle \leq f(\mathbf{x}) - f(\bar{\mathbf{x}}),$$

$$\Rightarrow \langle \nabla g(\mathbf{x}), \mathbf{x} - \bar{\mathbf{x}}\rangle \leq -\frac{1}{\eta}\|\mathbf{x} - \bar{\mathbf{x}}\|^2 + f(\mathbf{x}) - f(\bar{\mathbf{x}})$$

$$\Rightarrow f(\bar{\mathbf{x}}) \leq f(\mathbf{x}) - \langle \nabla g(\mathbf{x}), \mathbf{x} - \bar{\mathbf{x}}\rangle - \frac{1}{\eta}\|\mathbf{x} - \bar{\mathbf{x}}\|^2.$$

We get the desired result after combining it with the Descent Lemma (Lemma 2.1) on g:

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

□

Theorem 5.6 (Prox-Grad Inequality). *Assume $f : \mathbb{R}^n \longrightarrow \mathbb{R} \cup \{+\infty\}$ is a closed (or equivalently lower semicontinous) proper convex function, and $g : \mathbb{R}^n \longrightarrow \mathbb{R}$ is a differentiable convex function with ∇g being Lipschitz continuous with Lipschitz constant L. Let $F = f + g$, and $\bar{\mathbf{y}} = \mathrm{Prox}_f^\eta[\mathbf{y} - \eta\nabla g(\mathbf{y})]$,* **with step size** *$\eta \leq \frac{1}{L}$, then*

$$F(\mathbf{x}) - F(\bar{\mathbf{y}}) \geq g(\mathbf{x}) - g(\mathbf{y}) - \langle \nabla g(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle + \frac{1}{2\eta}\|\mathbf{x} - \bar{\mathbf{y}}\|^2 - \frac{1}{2\eta}\|\mathbf{x} - \mathbf{y}\|^2.$$

In particular, with $\eta = \frac{1}{L}$, we have

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

Remark 5.3. *Let $g(\mathbf{x}) \equiv 0$, then we get the following inequality for the proximal operator:*

$$f(\mathbf{x}) - f(\mathrm{Prox}_f^\eta(\mathbf{y})) \geq \frac{1}{2\eta}\|\mathbf{x} - \mathrm{Prox}_f^\eta(\mathbf{y})\|^2 - \frac{1}{2\eta}\|\mathbf{x} - \mathbf{y}\|^2, \forall \eta > 0.$$

Remark 5.4. *By setting* $\mathbf{y} = \mathbf{x} = \mathbf{x}_k$*, we get the following property of the proximal gradient method:*

$$F(\mathbf{x}_k) - F(\mathbf{x}_{k+1}) \geq \frac{1}{2\eta}\|\mathbf{x}_k - \mathbf{x}_{k+1}\|^2, \quad \forall \eta \leq \frac{1}{L}.$$

Proof. Define $\phi(\mathbf{u}) = g(\mathbf{y}) + \langle \nabla g(\mathbf{y}), \mathbf{u} - \mathbf{y}\rangle + f(\mathbf{u}) + \frac{1}{2\eta}\|\mathbf{u} - \mathbf{y}\|^2$, then it is a proper closed $\frac{1}{\eta}$-strongly convex function. By Theorem 4.12 and Theorem 4.13, ϕ has a unique minimizer $\mathbf{u}_*$ and it satisfies

$$\begin{aligned} \mathbf{0} &\in \nabla g(\mathbf{y}) + \partial f(\mathbf{u}_*) + \frac{1}{\eta}(\mathbf{u}_* - \mathbf{y}) \\ \Leftrightarrow \mathbf{u}_* &= (I + \eta \partial f)^{-1}(\mathbf{y} - \eta \nabla g(\mathbf{y})), \end{aligned}$$

which means $\bar{\mathbf{y}} = \mathbf{u}_*$ is the mimizer to $\phi(\mathbf{u})$.

By Lemma 4.1, we have

$$\phi(\mathbf{x}) \geq \phi(\mathbf{y}) + \langle \partial\phi(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle + \frac{1}{2\eta}\|\mathbf{x} - \mathbf{y}\|^2, \quad \forall \mathbf{x}, \mathbf{y} \in \mathrm{dom}(\phi).$$

Plugging in $\mathbf{y} = \bar{\mathbf{y}}$ (recall $\bar{\mathbf{y}}$ is the minimizer of ϕ) and $\partial\phi(\bar{\mathbf{y}}) \ni \mathbf{0}$, we have

$$\phi(\mathbf{x}) - \phi(\bar{\mathbf{y}}) \geq \frac{1}{2\eta}\|\mathbf{x} - \bar{\mathbf{y}}\|^2.$$

By the Descent Lemma (Lemma 2.1), we have

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

$$\phi(\bar{\mathbf{y}}) = g(\mathbf{y}) + \langle \nabla g(\mathbf{y}), \bar{\mathbf{y}} - \mathbf{y}\rangle + \frac{1}{2\eta}\|\bar{\mathbf{y}} - \mathbf{y}\|^2 + f(\bar{\mathbf{y}}) \geq g(\bar{\mathbf{y}}) + f(\bar{\mathbf{y}}) = F(\bar{\mathbf{y}}).$$

Combine the two inequalities above, we have

$$\phi(\mathbf{x}) - F(\bar{\mathbf{y}}) \geq \frac{1}{2\eta}\|\mathbf{x} - \bar{\mathbf{y}}\|^2,$$

which is the desired result. □

Theorem 5.7. *(Fejér monotonicity and* $\mathcal{O}(\frac{1}{k})$ *convergence rate.) Assume* $f : \mathbb{R}^n \longrightarrow \mathbb{R} \cup \{+\infty\}$ *is a closed (or equivalently lower semicontinous) proper convex function, and* $g : \mathbb{R}^n \longrightarrow \mathbb{R}$ *is a differentiable convex function with* ∇g *being Lipschitz continuous with Lipschitz constant* L*. Assume the existence of global minimizers of* $F(\mathbf{x}) = f(\mathbf{x}) + g(\mathbf{x})$*. With step size* $\eta \leq \frac{1}{L}$*, the iterates of the proximal gradient method satisfies*

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

where $\mathbf{x}_*$ *is any global minimizer to* $f(\mathbf{x}) + g(\mathbf{x})$*. With step size* $\eta = \frac{1}{L}$*, the iterates of the proximal gradient method satisfies*

$$F(\mathbf{x}_{k+1}) - F(\mathbf{x}_*) \leq \frac{L}{2}\|\mathbf{x}_0 - \mathbf{x}_*\|\frac{1}{k}, \quad \forall k.$$

Proof. **Step I**: By the Prox-Grad inequality above, we have

$$F(\mathbf{x}) - F(\bar{\mathbf{y}}) \geq g(\mathbf{x}) - g(\mathbf{y}) - \langle \nabla g(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle + \frac{1}{2\eta}\|\mathbf{x} - \bar{\mathbf{y}}\|^2 - \frac{1}{2\eta}\|\mathbf{x} - \mathbf{y}\|^2.$$

Convexity gives $g(\mathbf{x}) - g(\mathbf{y}) - \langle \nabla g(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \geq 0$, thus

$$F(\mathbf{x}) - F(\bar{\mathbf{y}}) \geq \frac{1}{2\eta}\|\mathbf{x} - \bar{\mathbf{y}}\|^2 - \frac{1}{2\eta}\|\mathbf{x} - \mathbf{y}\|^2.$$

Plug in $\mathbf{x} = \mathbf{x}_*$, $\mathbf{y} = \mathbf{x}_k \Rightarrow \bar{\mathbf{y}} = \mathbf{x}_{k+1}$, then

$$0 \geq 2\eta[F(\mathbf{x}_*) - F(\mathbf{x}_{k+1})] \geq \|\mathbf{x}_{k+1} - \mathbf{x}_*\| - \|\mathbf{x}_k - \mathbf{x}_*\|.$$

Step II: Sum up the inequality above, we have

$$\frac{2}{L}\sum_{k=0}^{n-1}[F(\mathbf{x}_*) - F(\mathbf{x}_{k+1})] \geq \|\mathbf{x}_n - \mathbf{x}_*\| - \|\mathbf{x}_0 - \mathbf{x}_*\| \geq -\|\mathbf{x}_0 - \mathbf{x}_*\|.$$

By Theorem 5.5 (Sufficient Decrease Lemma), we have

$$F(\mathbf{x}) - F(\bar{\mathbf{x}}) \geq \frac{L}{2}\|\mathbf{x} - \bar{\mathbf{x}}\| \Rightarrow F(\mathbf{x}_k) - F(\mathbf{x}_{k+1}) \geq \frac{L}{2}\|\mathbf{x}_k - \mathbf{x}_{k+1}\|,$$

thus

$$F(\mathbf{x}_k) \geq F(\mathbf{x}_{k+1}) \Rightarrow \frac{2}{L}n[F(\mathbf{x}_n) - F(\mathbf{x}_*)] \geq \frac{2}{L}\sum_{k=0}^{n-1}[F(\mathbf{x}_{k+1}) - F(\mathbf{x}_*)] \geq \|\mathbf{x}_0 - \mathbf{x}_*\|.$$

□

The Fejér monotonicity and $F(\mathbf{x}_k) \to F(\mathbf{x}_*)$ does imply convergence to some global minimizer, which is not necessarily $\mathbf{x}_*$. See [3] for general statements. For simplicity, we give a proof for a well defined function $F = f + g$ on the whole space $\mathbb{R}^n$:

Theorem 5.8. *Assume $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ is a convex function, and $g : \mathbb{R}^n \longrightarrow \mathbb{R}$ is a differentiable convex function with ∇g being Lipschitz continuous with Lipschitz constant L. Assume the existence of global minimizers of $F(\mathbf{x}) = f(\mathbf{x}) + g(\mathbf{x})$. With step size $\eta \leq \frac{1}{L}$, the iterates $\mathbf{x}_k$ of the proximal gradient method converges to one of the global minimizers as $k \to \infty$.*

Remark 5.5. *By setting $g(\mathbf{x}) \equiv 0$, we get the convergence of the proximal point method for minimizing $f(\mathbf{x})$.*

Proof. The Fejér monotonicity in the previous theorem implies the boundedness of the sequence $\{\mathbf{x}_k\}_{k=0}^{\infty} \subset \mathbb{R}^n$, thus there is a convergent subsequence $\mathbf{x}_{k_j} \to \mathbf{y}_*$ as $j \to \infty$.

By Theorem 4.3, the convexity implies the continuity of F, thus $F(\mathbf{x}_{k_j}) \to F(\mathbf{y}_*)$.

The convergence rate theorem above also gives $F(\mathbf{x}_{k_j}) \to F(\mathbf{x}_*)$ for any global minimizer $\mathbf{x}_*$, thus $F(\mathbf{y}_*) = F(\mathbf{x}_*)$. So $\mathbf{y}_*$ is one of the global minimizers.

So the Fejér monotonicity also applies to $\mathbf{y}_*$:

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

which forces $\mathbf{x}_k \to \mathbf{y}_*$. □

5.3.4 Convergence under strong convexity

Theorem 5.9. *(Linear convergence rate.) Assume $f : \mathbb{R}^n \longrightarrow \mathbb{R} \cup \{+\infty\}$ is a closed (or equivalently lower semicontinous) proper convex function, and $g : \mathbb{R}^n \longrightarrow \mathbb{R}$ is a μ-strongly convex function with ∇g being Lipschitz continuous with Lipschitz constant L. Then $F(\mathbf{x}) = f(\mathbf{x}) + g(\mathbf{x})$ is a proper lower semicontinous μ-strongly convex function, thus it has a unique global minimizer $\mathbf{x}_*$. With step size $\eta = \frac{1}{L}$, the iterates of the proximal gradient method satisfies*

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

Proof. The Prox-Grad inequality gives

$$F(\mathbf{x}_*)-F(\mathbf{x}_{k+1}) \geq g(\mathbf{x}_*)-g(\mathbf{x}_k)-\langle\nabla g(\mathbf{x}_k), \mathbf{x}_*-\mathbf{x}_k\rangle+\frac{L}{2}\|\mathbf{x}_*-\mathbf{x}_{k+1}\|^2-\frac{L}{2}\|\mathbf{x}_*-\mathbf{x}_k\|^2.$$

Strong convexity gives

$$g(\mathbf{x}_*) - g(\mathbf{x}_k) - \langle\nabla g(\mathbf{x}_k), \mathbf{x}_* - \mathbf{x}_k\rangle + \frac{\mu}{2}\|\mathbf{x}_* - \mathbf{x}_k\|^2$$

thus

$$0 \geq F(\mathbf{x}_*) - F(\mathbf{x}_{k+1}) \geq \frac{L}{2}\|\mathbf{x}_* - \mathbf{x}_{k+1}\|^2 - (\frac{L}{2} - \frac{\mu}{2})\|\mathbf{x}_* - \mathbf{x}_k\|^2.$$

So

$$\|\mathbf{x}_{k+1} - \mathbf{x}_*\|^2 \leq (1 - \frac{\mu}{L})\|\mathbf{x}_k - \mathbf{x}_*\|^2$$

and

$$F(\mathbf{x}_{k+1})-F(\mathbf{x}_*) \leq (\frac{L}{2}-\frac{\mu}{2})\|\mathbf{x}_*-\mathbf{x}_k\|^2-\frac{L}{2}\|\mathbf{x}_*-\mathbf{x}_{k+1}\|^2 < (\frac{L}{2}-\frac{\mu}{2})\|\mathbf{x}_*-\mathbf{x}_k\|^2.$$

□

5.4 The second convergence proof of proximal point method

We now reconsider the proximal point method for minimizing $f(\mathbf{x})$:

$$\mathbf{x}_{k+1} = (I + \eta\partial f)^{-1}(\mathbf{x}_k) = \mathrm{Prox}_f^{\eta}(\mathbf{x}_k).$$

By letting $g(\mathbf{x}) \equiv 0$ in theorems for the proximal gradient method, we can get the following theorem:

Theorem 5.10. *(Fejér monotonicity and $\mathcal{O}(\frac{1}{k})$ convergence rate.) Assume $f : \mathbb{R}^n \longrightarrow \mathbb{R} \cup \{+\infty\}$ is a closed (or equivalently lower semicontinous) proper convex function. Assume the existence of global minimizers of $f(\mathbf{x})$.* **For any positive step size** *$\eta > 0$, the iterates of the proximal point method satisfies*

$$\|\mathbf{x}_{k+1} - \mathbf{x}_*\| \leq \|\mathbf{x}_k - \mathbf{x}_*\|, \quad \forall k,$$
$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}_*) \leq \frac{1}{2\eta}\|\mathbf{x}_0 - \mathbf{x}_*\|\frac{1}{k}, \quad \forall k,$$

where $\mathbf{x}_*$ is any global minimizer to $f(\mathbf{x})$.

Theorem 5.11. *(Linear convergence rate under strong convexity.) Assume $f : \mathbb{R}^n \longrightarrow \mathbb{R} \cup \{+\infty\}$ is a closed (or equivalently lower semicontinous) proper μ-strongly convex function, then it has a unique global minimizers $\mathbf{x}_*$.* **For any positive step size** *$\eta > 0$, the iterates of the proximal point method satisfies*

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

Remark 5.6. *The proximal point method is unconditionally stable, i.e., $f(\mathbf{x}_k)$ is under control for any step size. With strong convexity, faster rate is achieved for larger step size η.*

Proof. For the proximal point method, there is some subderivative $\mathbf{v}_{k+1} \in \partial f(\mathbf{x}_{k+1})$ such that

$$\begin{aligned}\mathbf{x}_k &= \mathbf{x}_{k+1} + \eta\mathbf{v}_{k+1}\\ \mathbf{x}_k - \mathbf{x}_* &= \mathbf{x}_{k+1} - \mathbf{x}_* + \eta\mathbf{v}_{k+1}\\ \|\mathbf{x}_k - \mathbf{x}_*\|^2 &= \|\mathbf{x}_{k+1} - \mathbf{x}_* + \eta\mathbf{v}_{k+1}\|^2\end{aligned}$$

By Lemma 4.1, we have

$$\begin{aligned}\|\mathbf{x}_k - \mathbf{x}_*\|^2 &= \|\mathbf{x}_{k+1} - \mathbf{x}_*\|^2 + 2\langle \mathbf{x}_{k+1} - \mathbf{x}_*, \eta\mathbf{v}_{k+1}\rangle + \|\eta\mathbf{v}_{k+1}\|^2\\ &\geq \|\mathbf{x}_{k+1} - \mathbf{x}_*\|^2 + 2\eta\langle \mathbf{x}_{k+1} - \mathbf{x}_*, \mathbf{v}_{k+1}\rangle\\ &= \|\mathbf{x}_{k+1} - \mathbf{x}_*\|^2 + 2\eta\langle \mathbf{x}_{k+1} - \mathbf{x}_*, \partial f(\mathbf{x}_{k+1}) - \partial f(\mathbf{x}_{k+1})\rangle\\ &\geq \|\mathbf{x}_{k+1} - \mathbf{x}_*\|^2 + 2\eta\mu\|\mathbf{x}_{k+1} - \mathbf{x}_*\|^2\end{aligned}$$

thus we obtain

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

Let $g(\mathbf{x}) \equiv 0$ in the Prox-Grad Inequality, then we get the following inequality for the proximal operator:

$$f(\mathbf{x}_*) - f(\mathbf{x}_{k+1}) \geq \frac{1}{2\eta}\|\mathbf{x}_* - \mathbf{x}_{k+1}\|^2 - \frac{1}{2\eta}\|\mathbf{x}_* - \mathbf{x}_k\|^2, \forall \eta > 0.$$

Thus

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

□

5.5 The fast proximal gradient method

The proximal gradient method can be accelerated by combining with Nesterov's acceleration method in Section 2.8. The vanilla version of the fast proximal gradient method can be written as

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

The sequence t_k should satisfy $t_k^2 - t_k \leq t_{k-1}^2$, for which we can simply use $t_k = \frac{k+1}{2}$:

$$\begin{cases} \mathbf{x}_{k+1} &= (I+\eta\partial f)^{-1}[\mathbf{y}_k - \eta_k \nabla g(\mathbf{y}_k)] \\ \mathbf{y}_{k+1} &= \mathbf{x}_{k+1} + \frac{k-1}{k+2}(\mathbf{x}_{k+1}-\mathbf{x}_k) \end{cases} \qquad \mathbf{x}_0 = \mathbf{y}_0.$$

5.5.1 Convergence rate under convexity

Theorem 5.12 ($\mathcal{O}(\frac{1}{k^2})$ convergence rate)**.** *Assume $f : \mathbb{R}^n \longrightarrow \mathbb{R} \cup \{+\infty\}$ is a closed (or equivalently lower semicontinous) proper convex function, and $g : \mathbb{R}^n \longrightarrow \mathbb{R}$ is a differentiable convex function with ∇g being Lipschitz continuous with Lipschitz constant L. Assume the existence of global minimizers of $F(\mathbf{x}) = f(\mathbf{x}) + g(\mathbf{x})$. With step size $\eta = \frac{1}{L}$, the iterates of* *the fast proximal gradient method with* $t_k^2 - t_k \leq t_{k-1}^2$ *(e.g.,* $t_k = \frac{k+1}{2}$ *or* $t_{k+1} = \frac{1}{2}\left(1+\sqrt{4t_k^2+1}\right)$*) satisfies*

$$F(\mathbf{x}_{k+1}) - F(\mathbf{x}_*) \leq 2L\|\mathbf{x}_0 - \mathbf{x}_*\|\frac{1}{k^2}, \quad \forall k,$$

where $\mathbf{x}_*$ *is any global minimizer to* $f(\mathbf{x}) + g(\mathbf{x})$*.*

Remark 5.7. *By letting* $f(\mathbf{x}) \equiv 0$*, this is the same result as Theorem 2.10.*

Proof. The Prox-Grad inequality

$$\begin{aligned}F(\mathbf{x}) - F(\bar{\mathbf{y}}) &\geq g(\mathbf{x}) - g(\mathbf{y}) - \langle \nabla g(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle + \frac{1}{2\eta}\|\mathbf{x} - \bar{\mathbf{y}}\|^2 - \frac{1}{2\eta}\|\mathbf{x} - \mathbf{y}\|^2, \\ &\geq \frac{L}{2}\|\mathbf{x} - \bar{\mathbf{y}}\|^2 - \frac{L}{2}\|\mathbf{x} - \mathbf{y}\|^2.\end{aligned}$$

Plugging $\mathbf{x} = \frac{1}{t_k}\mathbf{x}_* + (1 - \frac{1}{t_k})\mathbf{x}_k$ and $\mathbf{y} = \mathbf{y}_k$ (then $\bar{\mathbf{y}} = \mathbf{x}_{k+1}$) into we get

$$F\left(\frac{1}{t_k}\mathbf{x}_* + (1 - \frac{1}{t_k})\mathbf{x}_k\right) - F(\mathbf{x}_{k+1}) \geq \frac{L}{2}\|\frac{1}{t_k}\mathbf{x}_* + (1 - \frac{1}{t_k})\mathbf{x}_k - \mathbf{x}_{k+1}\|^2 - \frac{L}{2}\|\frac{1}{t_k}\mathbf{x}_* + (1 - \frac{1}{t_k})\mathbf{x}_k - \mathbf{y}_k\|^2$$

$$= \frac{L}{2t_k^2}\|\mathbf{x}_* + (t_k - 1)\mathbf{x}_k - t_k\mathbf{x}_{k+1}\|^2 - \frac{L}{2t_k^2}\|\mathbf{x}_* + (t_k - 1)\mathbf{x}_k - t_k\mathbf{y}_k\|^2.$$

Convexity of F gives

$$F\left(\frac{1}{t_k}\mathbf{x}_* + (1 - \frac{1}{t_k})\mathbf{x}_k\right) \leq \frac{1}{t_k}F\left(\mathbf{x}_*\right) + (1 - \frac{1}{t_k})F\left(\mathbf{x}_k\right).$$

Let $R_k = F(\mathbf{x}_k) - F(\mathbf{x}_*)$, then

$$F\left(\frac{1}{t_k}\mathbf{x}_* + (1 - \frac{1}{t_k})\mathbf{x}_k\right) - F(\mathbf{x}_{k+1}) \leq (1 - \frac{1}{t_k})[F\left(\mathbf{x}_k\right) - F(\mathbf{x}_*)] - [F\left(\mathbf{x}_{k+1}\right) - F(\mathbf{x}_*)].$$

$$= (1 - \frac{1}{t_k})R_k - R_{k+1}.$$

Combining two inequalities, we get

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

If we define $\mathbf{u}_k = \mathbf{x}_* + (t_{k-1} - 1)\mathbf{x}_{k-1} - t_{k-1}\mathbf{x}_k$, then

$$\|\mathbf{x}_* + (t_k - 1)\mathbf{x}_k - t_k\mathbf{x}_{k+1}\|^2 = \|\mathbf{u}_{k+1}\|^2$$

Next, by plugging in

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

we get

$$\begin{aligned}\|\mathbf{x}_* + (t_k - 1)\mathbf{x}_k - t_k\mathbf{y}_k\|^2 &= \|\mathbf{x}_* + (t_k - 1)\mathbf{x}_k - t_k\mathbf{x}_k - (t_{k-1} - 1)(\mathbf{x}_k - \mathbf{x}_{k-1})\|^2 \\ &= \|\mathbf{x}_* + (t_{k-1} - 1)\mathbf{x}_{k-1} - t_{k-1}\mathbf{x}_k\|^2 = \|\mathbf{u}_k\|^2.\end{aligned}$$

So we get

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

So with $t_0 = 1$, we have

$$t_k^2 R_{k+1} + \frac{L}{2}\|\mathbf{u}_{k+1}\|^2 \leq t_0^2 R_1 + \frac{L}{2}\|\mathbf{u}_1\|^2 = F(\mathbf{x}_1) - F(\mathbf{x}_*) + \frac{L}{2}\|\mathbf{x}_* - \mathbf{x}_1\|^2.$$

Plugging $\mathbf{x} = \mathbf{x}_*$ and $\mathbf{y} = \mathbf{y}_0$ (then $\bar{\mathbf{y}} = \mathbf{x}_1$) into the Prox-Grad inequality

$$\begin{aligned}F(\mathbf{x}) - F(\bar{\mathbf{y}}) &\geq g(\mathbf{x}) - g(\mathbf{y}) - \langle \nabla g(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle + \frac{1}{2\eta}\|\mathbf{x} - \bar{\mathbf{y}}\|^2 - \frac{1}{2\eta}\|\mathbf{x} - \mathbf{y}\|^2, \\ &\geq \frac{L}{2}\|\mathbf{x} - \bar{\mathbf{y}}\|^2 - \frac{L}{2}\|\mathbf{x} - \mathbf{y}\|^2,\end{aligned}$$

so with $\mathbf{x}_0 = \mathbf{y}_0$, we get

$$\begin{aligned}F(\mathbf{x}_*) - F(\mathbf{x}_1) &\geq \frac{L}{2}\|\mathbf{x}_* - \mathbf{x}_1\|^2 - \frac{L}{2}\|\mathbf{x}_* - \mathbf{y}_0\|^2 \\ &\Rightarrow F(\mathbf{x}_1) - F(\mathbf{x}_*) + \frac{L}{2}\|\mathbf{x}_* - \mathbf{x}_1\|^2 \leq \frac{L}{2}\|\mathbf{x}_* - \mathbf{x}_0\|^2.\end{aligned}$$

Thus

$$R_{k+1} \leq \frac{L}{2}\|\mathbf{x}_0 - \mathbf{x}_*\|^2 \frac{1}{t_k^2}.$$

Finally, the $\frac{1}{k^2}$ rate is obtained from the rate $t_k = \mathcal{O}(\frac{1}{k})$. For instance, we can take $t_k = \frac{k+1}{2}$. We can also take

$$t_{k+1} = \frac{1}{2}\left(1 + \sqrt{4t_k^2 + 1}\right) \Rightarrow t_k \geq \frac{k+2}{2},$$

which can be verified by induction.

□

5.5.2 Convergence rate under strong convexity

If $g(\mathbf{x})$ is also μ-strongly convex, then $\sigma = \frac{L}{\mu}$ denotes its condition number. See [3, Theorem 10.42] for the convergence rate of a modified fast proximal gradient method:

Theorem 5.13. *Assume $f : \mathbb{R}^n \longrightarrow \mathbb{R} \cup \{+\infty\}$ is a closed (or equivalently lower semicontinous) proper convex function, and $g : \mathbb{R}^n \longrightarrow \mathbb{R}$ is a μ-strongly convex function with ∇g being Lipschitz continuous with Lipschitz constant L. Then $F(\mathbf{x}) = f(\mathbf{x}) + g(\mathbf{x})$ is a proper lower semicontinous μ-strongly convex function, thus it has a unique global minimizer $\mathbf{x}_*$. With step size $\eta = \frac{1}{L}$, the iterates of the following modified fast proximal gradient method*

$$\begin{cases} \mathbf{x}_{k+1} & = (I + \eta \partial f)^{-1}[\mathbf{y}_k - \eta_k \nabla g(\mathbf{y}_k)] \\ \mathbf{y}_{k+1} & = \mathbf{x}_{k+1} + \frac{\sqrt{\sigma}-1}{\sqrt{\sigma}+1}(\mathbf{x}_{k+1} - \mathbf{x}_k) \end{cases} \qquad \mathbf{x}_0 = \mathbf{y}_0.$$

satisfies

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

5.5.3 Restarted fast proximal gradient method

For $f(\mathbf{x}) = \|\mathbf{x}\|_1$, the proximal gradient method is often called ISTA (iterative shrinkage/thresholding algorithm), and the fast proximal gradient method is called FISTA.

The performace of the fast proximal gradient method can be significantly improved if using a restarted version (reset $t_N = 1, \mathbf{y}_N = \mathbf{x}_N$ every N iterations):

$$\begin{cases} \mathbf{x}_{k+1} & = (I + \eta \partial f)^{-1}[\mathbf{y}_k - \eta_k \nabla g(\mathbf{y}_k)] \\ t_{k+1} & = \frac{1}{2}\left(1 + \sqrt{4t_k^2 + 1}\right) \\ \mathbf{y}_{k+1} & = \mathbf{x}_{k+1} + \frac{t_k - 1}{t_{k+1}}(\mathbf{x}_{k+1} - \mathbf{x}_k) \end{cases}, \quad \text{if } \frac{k}{N} \text{ is an integer, set } \mathbf{y}_k = \mathbf{x}_k, t_k = 1.$$

5.6 Comparison and examples

So we can summarize all the **global convergence rate** results so far in this chapter as follows (see Section 6.2 for the linear rate of the proximal point method):

Assumptions	Convexity	Strong Convexity	Step Size
Subgradient Method	$\mathcal{O}(\frac{1}{\sqrt{k}})$	$\mathcal{O}(\frac{1}{k})$	$\eta_n \equiv \frac{1}{\sqrt{k}}$ or Polyak step size
Proximal Point Method	$\mathcal{O}(\frac{1}{k})$	$\mathcal{O}(\left(\frac{1}{1+\eta\mu}\right)^{2k})$	$\forall \eta > 0$
Proximal Gradient Method	$\mathcal{O}(\frac{1}{k})$	$\mathcal{O}((1-\frac{\mu}{L})^k)$	$\eta \leq \frac{1}{L}$
Fast Proximal Gradient	$\mathcal{O}(\frac{1}{k^2})$	$\mathcal{O}(\left(1-\sqrt{\frac{\mu}{L}}\right)^k)$	$\eta \leq \frac{1}{L}$

Table 5.1: The convergence rate for $F(\mathbf{x}_k) - F(\mathbf{x}_*)$ of different algorithms minimizing $F(\mathbf{x}) = f(\mathbf{x}) + g(\mathbf{x})$, where $f(\mathbf{x})$ and $g(\mathbf{x})$ are convex, μ is the strong convexity parameter of g, and L is the Lipschitz constant of ∇g.

Assumptions	Convexity	Strong Convexity
Gradient Descent with $\eta < \frac{2}{L}$	$\mathcal{O}(\frac{1}{k})$	
Fast Gradient Descent, $\eta = \frac{1}{L}$	$\mathcal{O}(\frac{1}{k^2})$	
Gradient Descent with $\eta \leq \frac{2}{L+\mu}$		$\mathcal{O}(\left(1-\frac{2\eta\mu L}{L+\mu}\right)^{2k})$
Gradient Descent with $\eta = \frac{2}{L+\mu}$		$\mathcal{O}(\left(\frac{L/\mu-1}{L/\mu+1}\right)^{2k})$
Fast Gradient Descent, $\eta = \frac{1}{L}$		$\mathcal{O}(\left(1-\sqrt{\frac{\mu}{L}}\right)^k)$

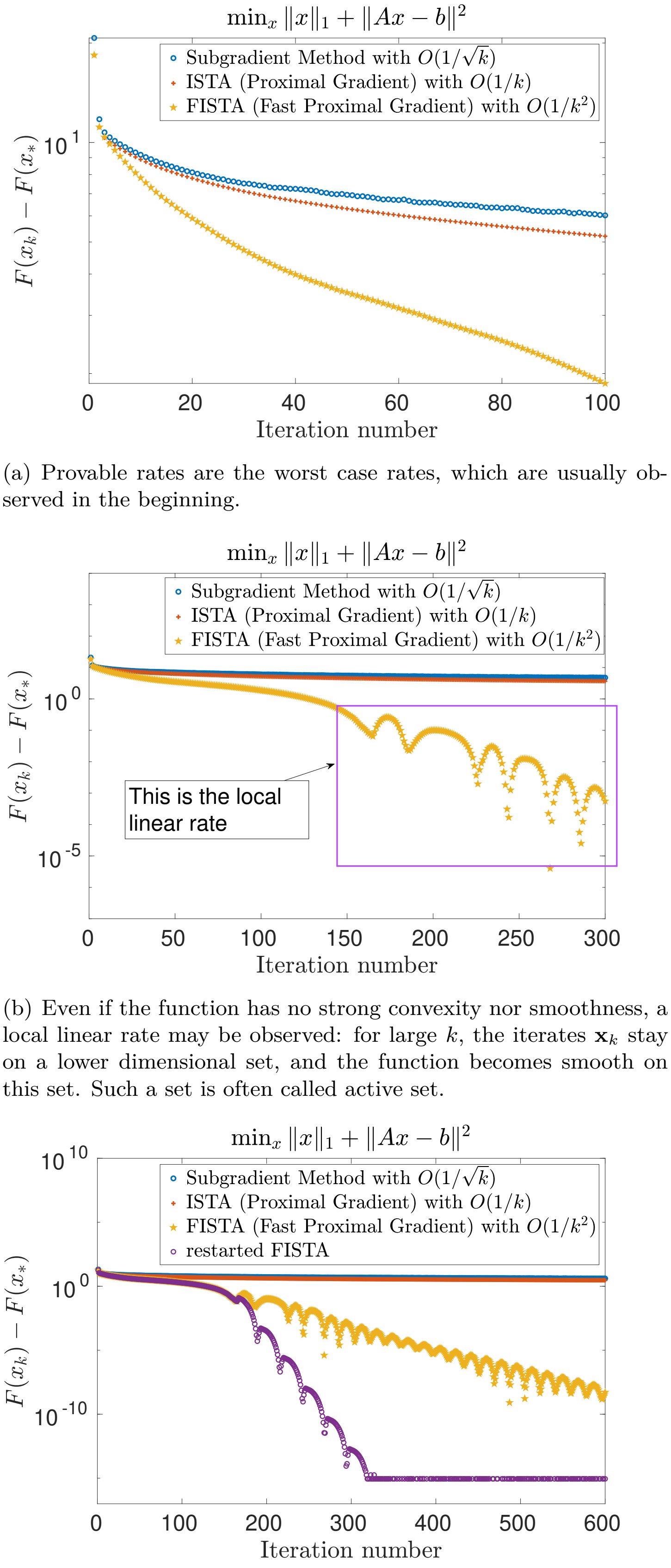
Table 5.2: The global convergence rate for $f(\mathbf{x}_k)-f(\mathbf{x}_*)$ of minimizing $f(\mathbf{x})$, where $f(\mathbf{x})$ is convex, μ is the strong convexity parameter of f, and L is the Lipschitz constant of ∇f. Remark: $1-\sqrt{\frac{\mu}{L}} < \left(\frac{L/\mu-1}{L/\mu+1}\right)^2$ if $\mu/L \leq 0.085$.

As a comparison, we also recall the global and local convergence rates proven for the smooth problems

Finally we take look at how slow/fast these methods can be for minimizing $\|\mathbf{x}\|_1 + \|A\mathbf{x} - b\|^2$. See Figure 5.1. We remark that monotone decay for $F(\mathbf{x}_k) - F(\mathbf{x}_*)$ can be proven for the subgradient method and the proximal gradient method, but for the fast proximal gradient method. We can observe that indeed $F(\mathbf{x}_k) - F(\mathbf{x}_*)$ is not monotone for the fast proximal gradient method, but obviously the fast proximal gradient method is much better.

Assumptions	Rate
Gradient Descent with $\|\mathbf{x}_0 - \mathbf{x}_*\| \leq \frac{2\mu}{M}$ and $\eta = \frac{2}{L+\mu}$	$\frac{\|\mathbf{x}_{k+1}-\mathbf{x}_*\|}{\|\mathbf{x}_k-\mathbf{x}_*\|} \leq 1 - \frac{2\mu}{L+3\mu}$
Newton's method with $\|\mathbf{x}_0 - \mathbf{x}_*\| \leq \frac{2\mu}{3M}$	$\frac{\|\mathbf{x}_{k+1}-\mathbf{x}_*\|}{\|\mathbf{x}_k-\mathbf{x}_*\|^2} \leq \frac{3M}{2\mu}$

Table 5.3: The **local** convergence rate for minimizing $f(\mathbf{x})$, where $f(\mathbf{x})$ is **not necessarily convex**, $\mathbf{x}_*$ is a local minimizer, μ is the strong convexity parameter of f, and M is the Lipschitz constant of the Hessian $\nabla^2 f$.



(a) Provable rates are the worst case rates, which are usually observed in the beginning.

(b) Even if the function has no strong convexity nor smoothness, a local linear rate may be observed: for large k, the iterates $\mathbf{x}_k$ stay on a lower dimensional set, and the function becomes smooth on this set. Such a set is often called active set.

(c) For ℓ^1 problem, restarted FISTA can perform extremely well.

Figure 5.1: A LASSO problem (4.3) with $A \in \mathbb{R}^{40\times 1000}$. The cost function $F(\mathbf{x}) = \|\mathbf{x}\|_1 + \|A\mathbf{x} - b\|^2$ is neither smooth nor strongly convex.

6

Fixed point iteration and Douglas-Rachford splitting

In this chapter, we consider some more sophisticated splitting algorithms to solve a more chanllenging ℓ^1-minimization problem (4.2):

$$\min_{\mathbf{x}} \|\mathbf{x}\|_1 + \iota_{\{\mathbf{x}:A\mathbf{x}=b\}}.$$

6.1 Nonexpansive operators

Definition 6.1. *An operator* $T : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ *is called*

- *contractive if* $\|T(\mathbf{x}) - T(\mathbf{y})\| < \|\mathbf{x} - \mathbf{y}\|, \quad \forall \mathbf{x}, \mathbf{y}.$
- *nonexpansive if* $\|T(\mathbf{x}) - T(\mathbf{y})\| \leq \|\mathbf{x} - \mathbf{y}\|, \quad \forall \mathbf{x}, \mathbf{y}.$
- *firmly nonexpansive if* $\|T(\mathbf{x}) - T(\mathbf{y})\|^2 \leq \langle T(\mathbf{x}) - T(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle, \quad \forall \mathbf{x}, \mathbf{y}.$

With the Cauchy-Schwartz inequality

$$\langle T(\mathbf{x}) - T(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle \leq \|T(\mathbf{x}) - T(\mathbf{y})\|\|\mathbf{x} - \mathbf{y}\|,$$

we can see that a firmly nonexpansive operator must be nonexpansive.

Theorem 6.1. *For an operator* $T : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ *and the identity operator* $I : \mathbb{R}^n \longrightarrow \mathbb{R}^n$*, the following are equivalent:*

1. T *is firmly nonexpansive.*
2. $I - T$ *is firmly nonexpansive.*
3. $2T - I$ *is nonexpansive.*
4. $\|T(\mathbf{x}) - T(\mathbf{y})\|^2 + \|(I - T)(\mathbf{x}) - (I - T)(\mathbf{y})\|^2 \leq \|\mathbf{x} - \mathbf{y}\|^2$

Proof. For 1 and 2:

$$\|T(\mathbf{x}) - T(\mathbf{y})\|^2 \le \langle T(\mathbf{x}) - T(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle$$

$$\|\mathbf{x} - T(\mathbf{x}) - [\mathbf{y} - T(\mathbf{y})]\|^2 = \|\mathbf{x} - \mathbf{y}\|^2 + \|T(\mathbf{x}) - T(\mathbf{y})\|^2 - \langle T(\mathbf{x}) - T(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle$$

$$\le \|\mathbf{x} - \mathbf{y}\|^2 - \langle T(\mathbf{x}) - T(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle = \langle [\mathbf{x} - T(\mathbf{x})] - [\mathbf{y} - T(\mathbf{y})], \mathbf{x} - \mathbf{y}\rangle.$$

For 1 and 3: let $R = 2T - I$, then

$$\|R(\mathbf{x}) - R(\mathbf{y})\|^2 = \|2[T(\mathbf{x}) - T(\mathbf{y})] - (\mathbf{x} - \mathbf{y})\|^2$$

$$= 4\|T(\mathbf{x}) - T(\mathbf{y})\|^2 + \|\mathbf{x} - \mathbf{y}\|^2 - 4\langle T(\mathbf{x}) - T(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle \le \|\mathbf{x} - \mathbf{y}\|^2$$

$$\Leftrightarrow \|T(\mathbf{x}) - T(\mathbf{y})\|^2 \le \langle T(\mathbf{x}) - T(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle$$

For 1 and 4: it can be similarly shown. □

Example 6.1. *Let $T(\mathbf{x}) = \eta \nabla f(\mathbf{x}), \quad 0 < \eta \le \frac{1}{L}$ where ∇f is Lipschitz continuous with Lipschitz constant L:*

1. *T is nonexpansive.*
2. *By Theorem 2.2, if f is convex, then T is firmly nonexpansive.*
3. *If f is convex, the gradient descent operator operator $I - T$ is also firmly nonexpansive.*

6.2 The third convergence rate of proximal point method

One important example of firmly nonexpansive and contractive operators is the proximal operator:

Theorem 6.2. *For a proper closed convex function $f : \mathbb{R}^n \longrightarrow (-\infty, +\infty]$, its proximal operator Prox_f^η is firmly nonexpansive:*

$$\|\mathrm{Prox}_f^\eta(\mathbf{x}) - \mathrm{Prox}_f^\eta(\mathbf{y})\|^2 \le \langle \mathrm{Prox}_f^\eta(\mathbf{x}) - \mathrm{Prox}_f^\eta(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle.$$

If f is also μ-strongly convex, then

$$(1 + \mu\eta)\|\mathrm{Prox}_f^\eta(\mathbf{x}) - \mathrm{Prox}_f^\eta(\mathbf{y})\|^2 \le \langle \mathrm{Prox}_f^\eta(\mathbf{x}) - \mathrm{Prox}_f^\eta(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle.$$

Remark 6.1. *With the Cauchy-Schwartz inequality $\langle T(\mathbf{x}) - T(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle \le \|T(\mathbf{x}) - T(\mathbf{y})\|\|\mathbf{x} - \mathbf{y}\|$, the proximal operator is contractive for a strongly convex f.*

Proof. Let

$$\mathbf{u} = \mathrm{Prox}_f^{\eta}(\mathbf{x}) = (I + \eta\partial f)^{-1}(\mathbf{x}) \Rightarrow \frac{\mathbf{x}-\mathbf{u}}{\eta} \in \partial f(\mathbf{u}),$$

$$\mathbf{v} = \mathrm{Prox}_f^{\eta}(\mathbf{y}) = (I + \eta\partial f)^{-1}(\mathbf{y}) \Rightarrow \frac{\mathbf{y}-\mathbf{v}}{\eta} \in \partial f(\mathbf{v}).$$

For a μ-strongly convex function ($\mu = 0$ for a convex function), by Lemma 4.1, we have

$$\langle \partial f(\mathbf{u}) - \partial f(\mathbf{v}), \mathbf{u} - \mathbf{v}\rangle \geq \mu\|\mathbf{u}-\mathbf{v}\|^2$$

thus

$$\langle \frac{\mathbf{x}-\mathbf{u}}{\eta} - \frac{\mathbf{y}-\mathbf{v}}{\eta}, \mathbf{u}-\mathbf{v}\rangle \geq \mu\|\mathbf{u}-\mathbf{v}\|^2,$$

$$\langle \mathbf{x}-\mathbf{y}, \mathbf{u}-\mathbf{v}\rangle \geq (1+\eta\mu)\|\mathbf{u}-\mathbf{v}\|^2.$$

□

We now reconsider the proximal point method for minimizing $f(\mathbf{x})$:

$$\mathbf{x}_{k+1} = (I + \eta\partial f)^{-1}(\mathbf{x}_k) = \mathrm{Prox}_f^{\eta}(\mathbf{x}_k).$$

Theorem 6.3. *(Linear convergence rate under strong convexity.) Assume $f : \mathbb{R}^n \longrightarrow \mathbb{R} \cup \{+\infty\}$ is a closed (or equivalently lower semicontinous) proper μ-strongly convex function, then it has a unique global minimizers $\mathbf{x}_*$. The iterates of the proximal point method satisfies*

1. $\|\mathbf{x}_k - \mathbf{x}_*\| \leq \left(\frac{1}{1+\eta\mu}\right)^k \|\mathbf{x}_0 - \mathbf{x}_*\|, \quad \forall \eta > 0.$
2. $f(\mathbf{x}_k) - f(\mathbf{x}_*) \leq \frac{1+\eta\mu}{2\eta}\left(\frac{1}{1+\eta\mu}\right)^{2k} \|\mathbf{x}_0 - \mathbf{x}_*\|^2, \quad \forall \eta > 0.$

Remark 6.2. *This is a better rate than the linear rate in Section 5.4 since*

$$\left(\frac{1}{1+\eta\mu}\right)^2 < \frac{1}{1+2\eta\mu} \Leftrightarrow 1 + 2\eta\mu < 1 + 2\eta\mu + \eta^2\mu^2.$$

Proof. By Theorem 6.2, the proximal operator is contractive

$$\|\mathrm{Prox}_f^{\eta}(\mathbf{x}) - \mathrm{Prox}_f^{\eta}(\mathbf{y})\| \leq \frac{1}{1+\mu\eta}\|\mathbf{x}-\mathbf{y}\|.$$

Notice that we have $\mathbf{x}_* = \mathrm{Prox}_f^{\eta}(\mathbf{x}_*)$. Plugging in $\mathbf{x} = \mathbf{x}_k$ and $\mathbf{y} = \mathbf{x}_*$, we get $\|\mathbf{x}_{k+1} - \mathbf{x}_*\| \leq \left(\frac{1}{1+\eta\mu}\right)\|\mathbf{x}_k - \mathbf{x}_*\|$ thus $\|\mathbf{x}_k - \mathbf{x}_*\| \leq \left(\frac{1}{1+\eta\mu}\right)^k \|\mathbf{x}_0 - \mathbf{x}_*\|$.

Let $g(\mathbf{x}) \equiv 0$ in the Prox-Grad Inequality (Theorem 5.6), then we get the following inequality for the proximal operator:

$$f(\mathbf{x}_*) - f(\mathbf{x}_{k+1}) \geq \frac{1}{2\eta}\|\mathbf{x}_* - \mathbf{x}_{k+1}\|^2 - \frac{1}{2\eta}\|\mathbf{x}_* - \mathbf{x}_k\|^2, \forall \eta > 0.$$

Thus

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

□

6.3 Fixed point iteration

We now consider iterative schemes in the form $\mathbf{x}_{k+1} = T(\mathbf{x}_k)$. There are many fixed point theorems and we list a few.

Theorem 6.4 (Banach Fixed Point Theorem)**.** *Let $T : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ be a contractive operator/mapping, then T has a unique fixed point $T(\mathbf{x}_*) = \mathbf{x}_*$.*

In the Banach Fixed Point Theorem (1922), the Euclidean space $\mathbb{R}^n$ can be replaced by a non-empty complete metric space.

Theorem 6.5 (Brouwer Fixed Point Theorem)**.** *Let S^n be the unit ball in $\mathbb{R}^n$. If $T : S^n \longrightarrow S^n$ is continuous, then T has at least one fixed point $T(\mathbf{x}_*) = \mathbf{x}_*$.*

In the Brouwer Fixed Point Theorem (1911), the unit ball can be replaced by a nonempty compact convex set.

Theorem 6.6 (Browder-Göhde-Kirk Fixed Point Theorem)**.** *Let X be a uniformly convex Banach space. Let Y be a non-empty, bounded, closed and convex subset of X. If $T : Y \longrightarrow Y$ is an nonexpansive operator, then it has a fixed point.*

A few quick examples:

1. If $T : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ is nonexpansive, then T may not have a fixed point. Counter example, $T(x, y) = \begin{pmatrix} \cos\theta & -\sin\theta \\ 0 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} 0 \\ 1 \end{pmatrix}$.
2. If $T : S^n \longrightarrow S^n$ is nonexpansive, then T has at least one fixed point but $\mathbf{x}_{k+1} = T(\mathbf{x}_k)$ may not converge. Counter example: $T(\mathbf{x}) = -\mathbf{x}$.

For the counter example above, even though the iterative scheme for the operator $\mathbf{x}_{k+1} = T(\mathbf{x}_k) = -\mathbf{x}_k$ does not converge, the relaxation scheme

$$\mathbf{x}_{k+1} = \theta \mathbf{x}_k + (1 - \theta) T(\mathbf{x}_k)$$

always converges. Such a fact still holds for a general setup:

Theorem 6.7. *Assume $T : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ is nonexpansive and T has at least one fixed point $\mathbf{x}_*$. Then the iteration scheme*

$$\mathbf{x}_{k+1} = S_\theta(\mathbf{x}_k) := \theta \mathbf{x}_k + (1 - \theta) T(\mathbf{x}_k), \quad \theta \in (0, 1)$$

satisfies:

1. *$\mathbf{x}_k$ converges to $\mathbf{y}_*$, one of fixed points of T.*
2. *$\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 \leq \frac{1}{k+1}(\frac{1}{\theta} - 1)\|\mathbf{x}_0 - \mathbf{y}_*\|^2$.*

Proof. Step I:

$$\begin{aligned}\|\mathbf{x}_{k+1}-\mathbf{x}_*\|^2 &= \|\theta[\mathbf{x}_k-\mathbf{x}_*]+(1-\theta)[T(\mathbf{x}_k)-\mathbf{x}_*]\|^2\\ &= \theta\|\mathbf{x}_k-\mathbf{x}_*\|^2+(1-\theta)\|T(\mathbf{x}_k)-\mathbf{x}_*\|^2-\theta(1-\theta)\|T(\mathbf{x}_k)-\mathbf{x}_k\|^2\\ &\le \theta\|\mathbf{x}_k-\mathbf{x}_*\|^2+(1-\theta)\|\mathbf{x}_k-\mathbf{x}_*\|^2-\theta(1-\theta)\|T(\mathbf{x}_k)-\mathbf{x}_k\|^2\\ &= \|\mathbf{x}_k-\mathbf{x}_*\|^2-\frac{\theta}{1-\theta}\|S_\theta(\mathbf{x}_k)-\mathbf{x}_k\|^2\end{aligned}$$

Step II: we first get $\|\mathbf{x}_{k+1}-\mathbf{x}_*\|^2 \le \|\mathbf{x}_k-\mathbf{x}_*\|^2$. Sum it up, we get

$$(n+1)\|\mathbf{x}_{n+1}-\mathbf{x}_n\|^2 \le \sum_{k=0}^{n}\|\mathbf{x}_{n+1}-\mathbf{x}_n\|^2 \le \frac{1-\theta}{\theta}[\|\mathbf{x}_0-\mathbf{x}_*\|^2-\|\mathbf{x}_{n+1}-\mathbf{x}_*\|^2],$$

where the first inequality is implied by

$$\|\mathbf{x}_{k+1}-\mathbf{x}_k\|^2 \le \theta\|\mathbf{x}_k-\mathbf{x}_{k-1}\|+(1-\theta)\|T(\mathbf{x}_k)-T(\mathbf{x}_{k-1})\| \le \|\mathbf{x}_k-\mathbf{x}_{k-1}\|.$$

Step III: $\|\mathbf{x}_{k+1}-\mathbf{x}_*\|^2 \le \|\mathbf{x}_k-\mathbf{x}_*\|^2$ implies that $\{\mathbf{x}_k\}$ is a bounded sequence thus it has a convergent subsequence $\mathbf{x}_{k_j} \to \mathbf{y}_*$. $\|\mathbf{x}_{k+1}-\mathbf{x}_k\|^2 \le \frac{1}{k+1}(\frac{1}{\theta}-1)\|\mathbf{x}_0-\mathbf{x}_*\|^2$ implies

$$\|\mathbf{x}_{k+1}-\mathbf{x}_k\|^2 \to 0 \Rightarrow \|S_\theta(\mathbf{x}_k)-\mathbf{x}_k\| \to 0 \Rightarrow (1-\theta)\|T(\mathbf{x}_k)-\mathbf{x}_k\| \to 0$$

thus $\|T(\mathbf{x}_{k_j})-\mathbf{x}_{k_j}\| \to 0$.

Since $\|T(\mathbf{x})-\mathbf{x}-[T(\mathbf{y})-\mathbf{y}]\| \le 2\|\mathbf{x}-\mathbf{y}\|$, $T-I$ is continuous, thus $\|T(\mathbf{y}_*)-\mathbf{y}_*\|=0$, which implies $\mathbf{y}_*$ is a fixed point.

Finally, $\|\mathbf{x}_{k+1}-\mathbf{x}_*\|^2 \le \|\mathbf{x}_k-\mathbf{x}_*\|^2$ for any fixed point $\mathbf{x}_*$ forces the whole sequence converging to $\mathbf{y}_*$.

□

6.4 Douglas-Rachford splitting

Now we consider a composite optimization problem

$$\min_{\mathbf{x}} f(\mathbf{x})+g(\mathbf{x}),$$

where both $f(\mathbf{x})$ and $g(\mathbf{x})$ are convex but not differentiable. We assume that the proximal operators for $f(\mathbf{x})$ and $g(\mathbf{x})$ are available. Consider the basis pursuit problem (4.2) as an example:

$$\min_{\mathbf{x}} \|\mathbf{x}\|_1+\iota_{\{\mathbf{x}:A\mathbf{x}=b\}},$$

we have $f(\mathbf{x})=\|\mathbf{x}\|_1$ and $g(\mathbf{x})=\iota_{\{\mathbf{x}:A\mathbf{x}=b\}}$. Assume $A\mathbb{R}^{m\times n}$ with $m<n$ has linearly independent rows so that AA^T is invertiable.

For the ℓ^1 function, its proximal operator is called Shrinkage operator:

$$\mathrm{Prox}_f^\gamma(\mathbf{x}) = \mathbf{v}, \quad \mathbf{v}_i = \begin{cases} x_i - \gamma, & x_i > 1, \\ x_i + \gamma, & x_i < -1, \\ 0, & x_i \in [-1, 1]. \end{cases}$$

For the indicator function, the proximal operator is the Euclidean projection:

$$\mathrm{Prox}_g^\gamma(\mathbf{x}) = \mathbf{x} + A^T(AA^T)^{-1}(b - A\mathbf{x}),$$

where $A^T(AA^T)^{-1}$ is also called pseudoinverse of A, see Appendix A.4.

So for the problem (4.2), we have proximal operators for f and g, but not for $f + g$. The divide and concur approach is to do a splitting by using Prox_f and Prox_g. And the most robust splitting method is called *Douglas-Rachford* splitting introduced by Lions and Mercier in 1979:

1. The same splitting was used for solving the heat equation by Peaceman and Rachford in 1955 and Douglas and Rachford in 1956. Such an approach is under the name alternating-direction implicit (ADI) methods for solving PDEs. It is for solving equations like $0 = A\mathbf{x} + B\mathbf{x}$
2. Lions and Mercier in 1979 extended it for solving inclusion equations like $0 \in \partial f(\mathbf{x}) + \partial g(\mathbf{x})$.
3. It is exactly equivalent to the very popular ADMM (Alternating Direction Method of Multipliers) method and some special version of split Bregman method, which are widely used for problems in nonlinear mechanics and image processing.

Definition 6.2. *For minimizing $f(\mathbf{x})+g(\mathbf{x})$, the simplest Douglas-Rachford splitting is given as*

$$\mathbf{y}_{k+1} = \frac{\mathbb{I} + \mathrm{R}_f^\eta \mathrm{R}_g^\eta}{2}(\mathbf{y}_k), \quad \forall \eta > 0,$$

where $\mathbb{I}$ is the identity operator, $\mathrm{R}_f^\eta = 2\,\mathrm{Prox}_f^\eta - \mathbb{I}$ *and* $\mathrm{R}_g^\eta = 2\,\mathrm{Prox}_g^\eta - \mathbb{I}$

The sequence $\mathbf{y}_k$ does NOT converge to $\mathbf{x}_*$ and the variable $\mathbf{y}$ is only an auxiliary variable. Instead, $\mathbf{x}_k = \mathrm{Prox}_g^\eta(\mathbf{y}_k)$ will converge to $\mathbf{x}_*$. So a more explicit expression of the simplest Douglas-Rachford splitting is

$$\text{Douglas-Rachford}: \begin{cases} \mathbf{y}_{k+1} & = \frac{\mathbb{I}+\mathrm{R}_f^\eta \mathrm{R}_g^\eta}{2}(\mathbf{y}_k) = \mathbf{y}_k - \mathbf{x}_k + \mathrm{Prox}_f^\eta(2\mathbf{x}_k - \mathbf{y}_k) \\ \mathbf{x}_k & = \mathrm{Prox}_g^\eta(\mathbf{y}_k) \end{cases}$$

The Douglas-Rachford splitting is robust in the sense that the convergence is true for any convex functions f and g with any step size $\eta > 0$:

Theorem 6.8. *If $f : \mathbb{R} \longrightarrow (-\infty, +\infty]$ and $g : \mathbb{R} \longrightarrow (-\infty, +\infty]$ are proper closed convex functions, assume $f(\mathbf{x})+g(\mathbf{x})$ has at least one minimizer, then the general Douglas-Rachford splitting converges to a fixed point $\mathbf{y}_*$:*

$$\mathbf{y}_{k+1} = [\theta \mathbb{I} + (1-\theta)\,\mathrm{R}_f^\eta\,\mathrm{R}_g^\eta](\mathbf{y}_k), \quad \theta \in (0,1),$$

and $\mathbf{x}_k = \mathrm{Prox}_g^\eta(\mathbf{y}_k)$ converges to $\mathbf{x}_*$, one of the minimizers of $f(\mathbf{x}) + g(\mathbf{y})$ with

1. *$\|\mathbf{y}_{k+1} - \mathbf{y}_*\| \le \|\mathbf{y}_k - \mathbf{y}_*\|$. But $\|\mathbf{x}_{k+1} - \mathbf{x}_*\| \le \|\mathbf{x}_k - \mathbf{x}_*\|$ is not true.*
2. $$\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 \le \|\mathbf{y}_{k+1} - \mathbf{y}_k\|^2 \le \frac{1}{k}(\frac{1}{\theta} - 1)\|\mathbf{y}_0 - \mathbf{y}_*\|^2.$$

Remark 6.3. *If we take $\theta = 0$, we get the Peaceman-Rachford splitting:*

$$\textit{Peaceman-Rachford}: \begin{cases} \mathbf{y}_{k+1} & = \mathrm{R}_f^\eta\,\mathrm{R}_g^\eta(\mathbf{y}_k) \\ \mathbf{x}_k & = \mathrm{Prox}_g^\eta(\mathbf{y}_k) \end{cases},$$

which however does not converge unless at least one of the two functions is strongly convex. For example, if f and g are indicator functions of two lines passing the origin in $\mathbb{R}^2$, then the double reflection iteration will diverge.

Proof. Step I: existence of minimizers implies that $T = \theta\mathbb{I} + (1-\theta)\,\mathrm{R}_f^\eta\,\mathrm{R}_g^\eta$ has fixed points:

$$\begin{aligned} T(\mathbf{y}_*) &= \mathbf{y}_* \\ \Leftrightarrow \mathrm{R}_f^\eta\,\mathrm{R}_g^\eta(\mathbf{y}_*) &= \mathbf{y}_* \\ \Leftrightarrow 2\,\mathrm{Prox}_f^\eta\,\mathrm{R}_g^\eta(\mathbf{y}_*) - \mathrm{R}_g^\eta(\mathbf{y}_*) &= \mathbf{y}_* \\ \Leftrightarrow \mathrm{Prox}_f^\eta\,\mathrm{R}_g^\eta(\mathbf{y}_*) &= \mathrm{Prox}_g^\eta(\mathbf{y}_*) \quad (\mathbf{z} = \mathrm{Prox}_g^\eta(\mathbf{y}_*)) \\ \Leftrightarrow \mathrm{Prox}_f^\eta(2\mathbf{z} - \mathbf{y}_*) &= \mathbf{z} \quad ((\mathbb{I} + \eta\partial g)\mathbf{z} = \mathbf{y}_*) \\ &\Leftrightarrow \mathbf{z} = (\mathbb{I} + \eta\partial f)^{-1}[2\mathbf{z} - (\mathbb{I} + \eta\partial g)\mathbf{z}] \\ &\Leftrightarrow 0 \in \partial f(\mathbf{z}) + \partial g(\mathbf{z}). \end{aligned}$$

Step II: by Theorem 6.2, Prox is firmly nonexpansive, thus $\mathrm{R} = 2\,\mathrm{Prox} - \mathbb{I}$ is nonexpansive (Theorem 6.1). So $T = \theta\mathbb{I} + (1-\theta)\,\mathrm{R}_f^\eta\,\mathrm{R}_g^\eta$ is also nonexpansive since it is a convex combinatin of $\mathbb{I}$ and $\mathrm{R}_f^\eta\,\mathrm{R}_g^\eta$.

Step III: by Theorem 6.7, $\mathbf{y}_k$ converges to one fixed point $\mathbf{y}_*$. Thus $\mathbf{x}_k = \mathrm{Prox}_g^\eta(\mathbf{y}_k)$ converges to one minimizer $\mathbf{x}_*$, due to the fact we have shown in Step I:

$$T(\mathbf{y}_*) = \mathbf{y}_* \Leftrightarrow 0 \in \partial f(\mathbf{z}) + \partial g(\mathbf{z}), \quad \mathbf{z} = \mathrm{Prox}_g^\eta(\mathbf{y}_*).$$

Theorem 6.7 also implies

$$\|\mathbf{y}_k - \mathbf{y}_*\|^2 \le \frac{1}{k}(\frac{1}{\theta} - 1)\|\mathbf{y}_0 - \mathbf{y}_*\|^2.$$

□

6.5 Examples and comparison

Sometimes the general or relaxed Douglas-Rachford splitting is also written in the form of

$$\mathbf{y}_{k+1} = [(1-\lambda)\mathbb{I} + \lambda\frac{\mathbb{I} + \mathrm{R}_f^\eta \mathrm{R}_g^\eta}{2}](\mathbf{y}_k), \quad \lambda \in (0,2). \tag{6.1}$$

See Figure 6.1 for a numerical example.

6.6 Convergence under strong convexity

Lemma 6.1. *If $f : \mathbb{R}^n \longrightarrow (-\infty, +\infty]$ is a proper closed μ-strongly convex function, its proximal and reflection operators satisfy*

$$\| \mathrm{R}_f^\eta(\mathbf{x}) - \mathrm{R}_f^\eta(\mathbf{y})\|^2 + 4\mu\eta\| \mathrm{Prox}_f^\eta(\mathbf{x}) - \mathrm{Prox}_f^\eta(\mathbf{y})\|^2 \le \|\mathbf{x}-\mathbf{y}\|^2.$$

Remark 6.4. *With only convexity $\mu = 0$, the reflection is only nonexpansive $\| \mathrm{R}_f^\eta(\mathbf{x}) - \mathrm{R}_f^\eta(\mathbf{y})\|^2 \le \|\mathbf{x}-\mathbf{y}\|^2$.*

Proof. By Theorem 6.2, we have

$$\begin{aligned}
&\| \mathrm{R}_f^\eta(\mathbf{x}) - \mathrm{R}_f^\eta(\mathbf{y})\|^2 = \|2\,\mathrm{Prox}_f^\eta(\mathbf{x}) - \mathbf{x} - (2\,\mathrm{Prox}_f^\eta(\mathbf{y}) - \mathbf{y})\|^2\\
=&4\mu\eta\| \mathrm{Prox}_f^\eta(\mathbf{x}) - \mathrm{Prox}_f^\eta(\mathbf{y})\|^2 + \|\mathbf{x}-\mathbf{y}\|^2 - 4\langle \mathrm{Prox}_f^\eta(\mathbf{x}) - \mathrm{Prox}_f^\eta(\mathbf{y}), \mathbf{x}-\mathbf{y}\rangle\\
\le& -4\mu\eta\| \mathrm{R}_f^\eta(\mathbf{x}) - \mathrm{R}_f^\eta(\mathbf{y})\|^2 + \|\mathbf{x}-\mathbf{y}\|^2\\
\Rightarrow&\| \mathrm{R}_f^\eta(\mathbf{x}) - \mathrm{R}_f^\eta(\mathbf{y})\|^2 + 4\mu\eta\| \mathrm{Prox}_f^\eta(\mathbf{x}) - \mathrm{Prox}_f^\eta(\mathbf{y})\|^2 \le \|\mathbf{x}-\mathbf{y}\|^2.
\end{aligned}$$

□

Theorem 6.9 (Convergence of Peaceman-Rachford under strong convexity.)**.** *If $f : \mathbb{R}^n \longrightarrow (-\infty, +\infty]$ is a proper closed convex function and $g : \mathbb{R}^n \longrightarrow (-\infty, +\infty]$ is a proper closed μ-strongly convex function, then $f(\mathbf{x}) + g(\mathbf{x})$ has a unique minimizer $\mathbf{x}_*$. The iteration*

$$\begin{cases} \mathbf{y}_{k+1} &= \mathrm{R}_f^\eta \mathrm{R}_g^\eta(\mathbf{y}_k) \\ \mathbf{x}_k &= \mathrm{Prox}_g^\eta(\mathbf{y}_k) \end{cases}, \quad \forall \eta > 0$$

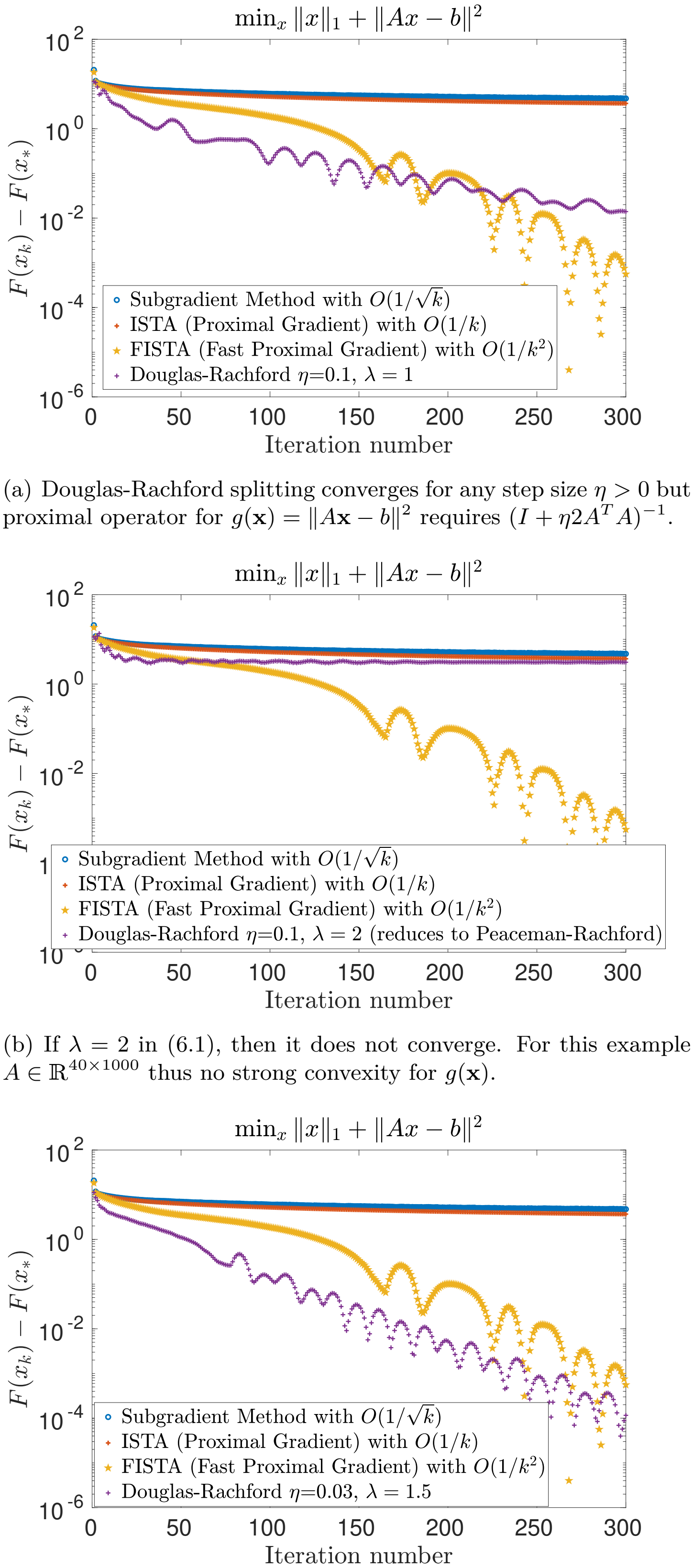
converges and

$$\min_{0\le k\le n} \|\mathbf{x}_k - \mathbf{x}_*\|^2 \le \frac{1}{n+1}\frac{1}{4\mu\eta}\|\mathbf{y}_0 - \mathbf{y}_*\|^2.$$

Remark 6.5. *Notice that we can switch f and g in the Peaceman-Rachford splitting*

$$\begin{cases} \mathbf{z}_{k+1} &= \mathrm{R}_g^\eta \mathrm{R}_f^\eta(\mathbf{z}_k) \\ \mathbf{x}_k &= \mathrm{Prox}_f^\eta(\mathbf{z}_k) \end{cases}, \quad \forall \eta > 0$$

and the same results apply, because it is the same iteration as above if we set $\mathbf{z}_0 = \mathbf{y}_0$.



(a) Douglas-Rachford splitting converges for any step size $\eta > 0$ but proximal operator for $g(\mathbf{x}) = \|A\mathbf{x} - b\|^2$ requires $(I + \eta 2A^T A)^{-1}$.

(b) If $\lambda = 2$ in (6.1), then it does not converge. For this example $A \in \mathbb{R}^{40\times1000}$ thus no strong convexity for $g(\mathbf{x})$.

(c) If tuning parameters, Douglas-Rachford splitting can be faster than FISTA or restarted FISTA for certain accuracy threshold.

Figure 6.1: A LASSO problem (4.3) with $A \in \mathbb{R}^{40\times1000}$. The cost function $F(\mathbf{x}) = \|\mathbf{x}\|_1 + \|A\mathbf{x} - b\|^2$ is neither smooth nor strongly convex.

Proof.

$$
\begin{aligned}
&\| \mathrm{R}_f^\eta \mathrm{R}_g^\eta(\mathbf{y}_k) - \mathrm{R}_f^\eta \mathrm{R}_g^\eta(\mathbf{y}_*)\|^2 + 4\mu\eta\| \mathrm{Prox}_g^\eta(\mathbf{y}_k) - \mathrm{Prox}_g^\eta(\mathbf{y}_*)\|^2 \\
\leq& \| \mathrm{R}_g^\eta(\mathbf{y}_k) - \mathrm{R}_g^\eta(\mathbf{y}_*)\|^2 + 4\mu\eta\| \mathrm{Prox}_g^\eta(\mathbf{y}_k) - \mathrm{Prox}_g^\eta(\mathbf{y}_*)\|^2 \\
\leq& \|\mathbf{y}_k - \mathbf{y}_*\|^2,
\end{aligned}
$$

thus

$$\|\mathbf{y}_{k+1} - \mathbf{y}_*\|^2 + 4\mu\eta\|\mathbf{x}_k - \mathbf{x}_*\|^2 \leq \|\mathbf{y}_k - \mathbf{y}_*\|^2$$

$$\sum_{k=0}^{n} 4\mu\eta\|\mathbf{x}_k - \mathbf{x}_*\|^2 \leq \|\mathbf{y}_0 - \mathbf{y}_*\|^2 - \|\mathbf{y}_{n+1} - \mathbf{y}_*\|^2 \leq \|\mathbf{y}_0 - \mathbf{y}_*\|^2.$$

□

Lemma 6.2. *If $f : \mathbb{R}^n \longrightarrow (-\infty, +\infty)$ is a μ-strongly convex function and ∇f is Lipschitz continuous, its reflection operator is a contraction:*

$$\| \mathrm{R}_f^\eta(\mathbf{x}) - \mathrm{R}_f^\eta(\mathbf{y})\|^2 \leq \left(1 - \frac{4\mu\eta}{(1+\mu L)^2}\right) \|\mathbf{x} - \mathbf{y}\|^2.$$

Proof.

$$
\begin{aligned}
&\|[\mathbb{I} + \eta\nabla f](\mathbf{x}) - [\mathbb{I} + \eta\nabla f](\mathbf{y})\|^2 \\
=& \|\mathbf{x} - \mathbf{y}\|^2 + \eta^2\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|^2 + 2\eta\langle \mathbf{x} - \mathbf{y}, \nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\rangle \\
\leq& \|\mathbf{x} - \mathbf{y}\|^2 + \eta^2 L^2\|\mathbf{x} - \mathbf{y}\|^2 + 2\eta\|\mathbf{x} - \mathbf{y}\|\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \\
\leq& (1 + \eta^2 L^2 + 2\eta L)\|\mathbf{x} - \mathbf{y}\|^2
\end{aligned}
$$

thus

$$\| \mathrm{Prox}_f^\eta(\mathbf{u}) - \mathrm{Prox}_f^\eta(\mathbf{v})\|^2 \geq \frac{1}{[1+\eta L]^2}\|\mathbf{x} - \mathbf{y}\|^2.$$

Plug it into Lemma 6.1, we get the desired result. □

Theorem 6.10. *(Linear rate of general Douglas-Rachford splitting.) Assume $f : \mathbb{R}^n \longrightarrow (-\infty, +\infty]$ is a proper closed convex function, $g : \mathbb{R}^n \longrightarrow (-\infty, +\infty)$ is a μ-strongly convex function and ∇g is Lipschitz continuous with Lipschitz constant L. Then the general Douglas-Rachford*

$$\begin{cases} \mathbf{y}_{k+1} &= \theta\mathbb{I} + (1-\theta)\,\mathrm{R}_f^\eta \mathrm{R}_g^\eta(\mathbf{y}_k) \\ \mathbf{x}_k &= \mathrm{Prox}_g^\eta(\mathbf{y}_k) \end{cases}, \quad \theta \in (0,1], \quad \forall\eta > 0,$$

satisfies

1. $\|\mathbf{y}_{k+1} - \mathbf{y}_*\| \leq \|\mathbf{y}_k - \mathbf{y}_*\|$. *But* $\|\mathbf{x}_{k+1} - \mathbf{x}_*\| \leq \|\mathbf{x}_k - \mathbf{x}_*\|$ *is not true.*

2.

$$\|\mathbf{x}_{k+1}-\mathbf{x}_*\|^2 \leq \|\mathbf{y}_{k+1}-\mathbf{y}_*\|^2 \leq [\theta+(1-\theta)c]^{2k}\|\mathbf{y}_0-\mathbf{y}_*\|^2,$$

where

$$c=\sqrt{1-\frac{4\mu\eta}{(1+\eta L)^2}} \geq \frac{1-\eta L}{1+\eta L}.$$

Remark 6.6. *The best provable rate w.r.t. θ would be $\theta = 1$ (Peaceman-Rachford). On the other hand, such a provable rate is usually much slower than the actual convergence rate, thus the fastest parameter may not be $\theta = 1$ in practice.*

Proof. 1. Lemma 6.2 implies $\|\mathrm{R}_g^\eta(\mathbf{x})-\mathrm{R}_g^\eta(\mathbf{y})\| \leq c\|\mathbf{x}-\mathbf{y}\|$.

2. R_f^η is nonexpansive, so

$$\|\mathrm{R}_f^\eta\mathrm{R}_g^\eta(\mathbf{x})-\mathrm{R}_f^\eta\mathrm{R}_g^\eta(\mathbf{y})\| \leq c\|\mathbf{x}-\mathbf{y}\|,$$

3. Let $T=\theta\mathbb{I}+(1-\theta)\,\mathrm{R}_f^\eta\mathrm{R}_g^\eta$, then

$$\|T(\mathbf{x})-T(\mathbf{y})\| \leq [\theta+(1-\theta)c]\|\mathbf{x}-\mathbf{y}\|$$

thus

$$\|\mathbf{x}_k-\mathbf{x}_*\| \leq \|\mathbf{y}_k-\mathbf{y}_*\| \leq [\theta+(1-\theta)c]^k\|\mathbf{y}_0-\mathbf{y}_*\|.$$

□

6.7 Maximal monotone operators

We summarize the main results in this chapter by considering set valued operators:

1. An operator T is a set-valued mapping or multi-valued function on $\mathbb{R}^n$ if $T(\mathbf{x}) \subset \mathbb{R}^n$, and its domain is defined as

$$\mathrm{dom}(T)=\{\mathbf{x}\in\mathbb{R}^n : T(\mathbf{x})\neq\emptyset\}.$$

2. An operator T is called *monotone* if

$$\langle T(\mathbf{x})-T(\mathbf{y}),\mathbf{x}-\mathbf{y}\rangle, \quad \forall \mathbf{x},\mathbf{y}\in\mathrm{dom}(T).$$

3. The graph of an operator is defined as

$$\mathit{Graph}(T)=\{(\mathbf{x},\mathbf{y})\in\mathbb{R}^n\times\mathbb{R}^n : \mathbf{y}\in T(\mathbf{x})\}.$$

4. A monotone operator T is called *maximal monotone* if there is no other monotone operator S such that Graph(T) is a proper subset of Graph(S).

5. The inverse operator is defined as $T^{-1}(\mathbf{x}) = \{\mathbf{y} : \mathbf{x} \in T(\mathbf{y})\}$.

6. The *resolvent* of an operator T is defined as $J_T = (\mathbb{I} + T)^{-1}$.

7. The *reflection* of an operator T is defined as $R_T = 2J_T - \mathbb{I} = 2(\mathbb{I} + T)^{-1} - \mathbb{I}$.

8. If T is nonexpansive, then $(1-\theta)\mathbb{I}+\theta T$ is called *θ-averaged* operator for $\theta \in (0, 1)$. By Theorem 6.7, the fixed point iteration of any *θ-averaged* operator always converges.

Lemma 6.3. *If A is maximal monotone, R_A is a nonexpansive (single-valued) operator with* $\mathrm{dom}(R_A) = \mathbb{R}^n$ *, and J_A is a (1/2)-averaged with* $\mathrm{dom}(J_A) = \mathbb{R}^n$

Examples:

1. If $f(\mathbf{x})$ is a proper convex function, then ∂f is a monotone operator.

2. If $f(\mathbf{x})$ is a proper closed convex function, then ∂f is a maximal monotone operator, thus Prox_f^η is (1/2)-averaged and R_f^η is nonexpansive.

3. If $f(\mathbf{x})$ and $g(\mathbf{x})$ are proper closed convex functions, then

$$R_f^\eta R_g^\eta \text{ is nonexpansive} \Leftrightarrow \frac{\mathbb{I} + R_f^\eta R_g^\eta}{2} \text{ is firmly nonexpansive}$$

$$R_f^\eta R_g^\eta \text{ is nonexpansive} \Rightarrow \frac{(1-\theta)\mathbb{I} + \theta R_f^\eta R_g^\eta}{2} \text{ converges for any } \theta \in (0,1), \eta > 0.$$

We list the algorithms as follows:

$$\text{General Douglas-Rachford} : \begin{cases} \mathbf{y}_{k+1} & = [(1-\lambda)\mathbb{I} + \lambda \frac{\mathbb{I}+\mathrm{R}_f^\eta \mathrm{R}_g^\eta}{2}](\mathbf{y}_k), \quad \lambda \in (0,2) \\ & = \mathbf{y}_k - \lambda \mathbf{x}_k + \lambda \, \mathrm{Prox}_f^\eta(2\mathbf{x}_k - \mathbf{y}_k) \\ \mathbf{x}_k & = \mathrm{Prox}_g^\eta(\mathbf{y}_k) \end{cases}$$

$$\text{Douglas-Rachford} : \begin{cases} \mathbf{y}_{k+1} & = \frac{\mathbb{I}+\mathrm{R}_f^\eta \mathrm{R}_g^\eta}{2}(\mathbf{y}_k) = \mathbf{y}_k - \mathbf{x}_k + \mathrm{Prox}_f^\eta(2\mathbf{x}_k - \mathbf{y}_k) \\ \mathbf{x}_k & = \mathrm{Prox}_g^\eta(\mathbf{y}_k) \end{cases}$$

$$\text{Peaceman-Rachford} : \begin{cases} \mathbf{y}_{k+1} & = \mathrm{R}_f^\eta \mathrm{R}_g^\eta(\mathbf{y}_k) \\ \mathbf{x}_k & = \mathrm{Prox}_g^\eta(\mathbf{y}_k) \end{cases}.$$

6.8 Davis-Yin splitting

For a composite problem of two functions $\min_{\mathbf{x}} f(\mathbf{x}) + g(\mathbf{x})$, the Douglas-Rachfording splitting is given as

$$\begin{cases} \mathbf{y}_{k+1} & = \mathbf{y}_k - \mathbf{x}_k + \mathrm{Prox}_f^\eta(2\mathbf{x}_k - \mathbf{y}_k) \\ \mathbf{x}_k & = \mathrm{Prox}_g^\eta(\mathbf{y}_k) \end{cases} .$$

To extend it to three functions

$$\min_{\mathbf{x}} f(\mathbf{x}) + g(\mathbf{x}) + h(\mathbf{x}),$$

the Davis-Yin splitting [5] is given as

$$\begin{cases} \mathbf{x}_{k+\frac{1}{2}} & = \mathrm{Prox}_g^\eta(\mathbf{z}_k) \\ \mathbf{x}_{k+1} & = \mathrm{Prox}_f^\eta(2\mathbf{x}_{k+\frac{1}{2}} - \mathbf{z}_k - \eta\nabla h(\mathbf{x}_{k+\frac{1}{2}})) \\ \mathbf{z}_{k+1} & = \mathbf{z}_k + \mathbf{x}_{k+1} - \mathbf{x}_{k+\frac{1}{2}} \end{cases} ,$$

which can be proven convergent for convex functions f, g, h with ∇h being Lipschitz continuous with Lipschitz L, and $\eta < \frac{2}{L}$.

Remark 6.7. *When $h(\mathbf{x}) \equiv 0$, Davis-Yin splitting reduces to Douglas-Rachford splitting. When $g(\mathbf{x}) \equiv 0$, Davis-Yin splitting reduces to the forward-backward splitting.*

Remark 6.8. *The Davis-Yin splitting is equivalent to a 3-block ADMM method (for a dual problem), which however is slightly different from the popular version of 3-block ADMM method.*

7

Fenchel duality and primal dual methods

7.1 Convex conjugate

The *convex conjugate* is also called Legendre Transform, Fenchel Transform and Fenchel dual:

Definition 7.1. *For an extended function $f : \mathbb{R} \longrightarrow [-\infty, +\infty]$, its convex conjugate is defined as*

$$f^*(\mathbf{x}) = \max_{\mathbf{y}} \langle \mathbf{x}, \mathbf{y} \rangle - f(\mathbf{y}).$$

Theorem 7.1. *For any proper function $f : \mathbb{R} \longrightarrow (-\infty, +\infty]$, $f^*(\mathbf{x})$ is a closed convex function on its domain $\operatorname{dom}(f^*) = \{\mathbf{x} : f^*(\mathbf{x}) < +\infty\}$ even if $f(\mathbf{x})$ is not convex.*

Theorem 7.2. *For any proper convex function $f : \mathbb{R} \longrightarrow (-\infty, +\infty]$, $f^*(\mathbf{x})$ is a proper closed convex function.*

Example 7.1. *By solving the critical point equation, we can find that*

1. *For $f(x) = e^x$,*

$$f^*(x) = \begin{cases} x \log x - x, & x > 0 \\ 0, & x = 0 \\ +\infty, & x < 0 \end{cases}.$$

2. *For $f(x) = ax + b$,*

$$f^*(x) = \begin{cases} -b, & x = a \\ +\infty, & x \neq a \end{cases}.$$

Theorem 7.3. *The convex conjugate satisfies:*

- *For any* $f : \mathbb{R} \longrightarrow (-\infty, +\infty]$, $f(\mathbf{x}) + f^*(\mathbf{y}) \geq \langle \mathbf{x}, \mathbf{y} \rangle, \quad \forall \mathbf{x}, \mathbf{y}$.
- *For any* $f : \mathbb{R} \longrightarrow [-\infty, +\infty]$, $f(\mathbf{x}) \geq f^{**}(\mathbf{x}), \quad \forall \mathbf{x}$.
- *For any proper closed convex function* f, $f(\mathbf{x}) = f^{**}(\mathbf{x}), \quad \forall \mathbf{x}$.

Example 7.2. *If* $f(\mathbf{x}) = \|\mathbf{x}\|$ *for some norm, then the convex conjugate is the indicator function of the dual norm unit ball*

$$f^*(\mathbf{x}) = \begin{cases} 0, & \|\mathbf{x}\|_* \leq 1 \\ +\infty, & \|\mathbf{x}\|_* > 1 \end{cases}.$$

1. *For the vector 2-norm* $\|\mathbf{x}\|_2$, $\|\mathbf{x}\|_* = \|\mathbf{x}\|_2$.
2. *For the vector 1-norm* $\|\mathbf{x}\|_1$, $\|\mathbf{x}\|_* = \|\mathbf{x}\|_\infty = \max_i |x_i|$.
3. *For the vector* ∞*-norm* $\|\mathbf{x}\|_\infty$, $\|\mathbf{x}\|_* = \|\mathbf{x}\|_1 = \sum_i |x_i|$.
4. *For the vector* p*-norm* $\|\mathbf{x}\|_p$, $\|\mathbf{x}\|_* = \|\mathbf{x}\|_q$, *where*

$$\|\mathbf{x}\|_p = \left[\sum_i |x_i|^p\right]^{\frac{1}{p}}, \quad \frac{1}{p} + \frac{1}{q} = 1.$$

7.1.1 Fenchel Duality Theorem

Theorem 7.4. *Let* f *and* g *be two proper convex functions. Assume the intersection of relative interior of* $\operatorname{dom}(f)$ *and relative interior of* $\operatorname{dom}(g)$ *is not empty, then*

$$\min_{\mathbf{x}}[f(\mathbf{x}) + g(\mathbf{x})] = -\min_{\mathbf{y}}[f^*(\mathbf{y}) + g^*(-\mathbf{y})]$$

Theorem 7.5. *For a proper closed convex function* f, *the following are equivalent:*

1. $f(\mathbf{x}) + f^*(\mathbf{y}) = \langle \mathbf{x}, \mathbf{y} \rangle$.
2. $\mathbf{y} \in \partial f(\mathbf{x})$.
3. $\mathbf{x} \in \partial f^*(\mathbf{y})$.

Therefore, for two proper closed convex functions, we have the following primal dual relation:

$$\mathbf{y}^* = \operatorname*{argmin}_{\mathbf{y}} f^*(\mathbf{y}) - \langle \mathbf{x}^*, \mathbf{y} \rangle \Leftrightarrow \mathbf{x}^* \in \partial f^*(\mathbf{y}^*)$$

$$\mathbf{x}^* = \operatorname*{argmin}_{\mathbf{x}} \langle \mathbf{x}, \mathbf{y}^* \rangle + g(\mathbf{x}) \Leftrightarrow -\mathbf{y}^* \in \partial g(\mathbf{x}^*).$$

7.1.2 Strong convexity and smoothness

Theorem 7.6 (Conjugate Correspondence.). *Strong convexity is equivalent to smoothness of the conjugate function:*

1. *If f is proper closed μ-strongly convex, then ∇f^* is Lipschitz continuous with $L = \frac{1}{\mu}$.*

2. *If $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ has Lipschitz continuous gradient, then f^* is μ-strongly convex with $\mu = \frac{1}{L}$.*

Example 7.3. *Consider the problem $\min_{\mathbf{x}} f(\mathbf{x})+g(\mathbf{x})$, where $f(\mathbf{x}) = \|\mathbf{x}\|_1 + \frac{1}{2\alpha}\|\mathbf{x}\|^2, g(\mathbf{x}) = \iota_{\{\mathbf{x}:A\mathbf{x}=b\}}(\mathbf{x})$. Then the strong convexity of f implies ∇f^* is Lipschitz-continuous with $L = \alpha$. For finding ∇f^*, there is no need to evaluate f^*:*

$$\mathbf{x} = \nabla f^*(\mathbf{y}) \Leftrightarrow \mathbf{y} \in \partial f(\mathbf{x}) = \partial\|\mathbf{x}\|_1 + \frac{1}{\alpha}\mathbf{x} \Leftrightarrow \mathbf{x} = (\mathbb{I} + \alpha\partial\| \cdot \|_1)^{-1}(\alpha\mathbf{y}).$$

7.1.3 Moreau-Decomposition

Theorem 7.7 (Moreau-Decomposition). *For any proper closed convex function $f : \mathbb{R} \longrightarrow (-\infty, +\infty]$, for any $\mathbf{x}$ and $\eta > 0$,*

$$\mathrm{Prox}_f^{\eta}(\mathbf{x}) + \eta\,\mathrm{Prox}_{f^*}^{\frac{1}{\eta}}(\frac{\mathbf{x}}{\eta}) = \mathbf{x}.$$

Thus by Moreau-Decomposition, we have Prox_{f^*} whenver we have Prox_f.

Example 7.4. *Now consider solving the basis pursuit problem (4.2). It is proven in [18] that the minimizer (4.4) for large enough α minimizes (4.2). So assume we use a large enough α, we consider (4.4) written in the following form*

$$\min_{\mathbf{x}} \|\mathbf{x}\|_1 + \iota_{\{\mathbf{x}:A\mathbf{x}=b\}}(\mathbf{x}) + \frac{1}{2\alpha}\|\mathbf{x}\|^2$$

Let $f(\mathbf{x}) = \|\mathbf{x}\|_1$ and $g(\mathbf{x}) = \iota_{\{\mathbf{x}:A\mathbf{x}=b\}}(\mathbf{x}) + \frac{1}{2\alpha}\|\mathbf{x}\|^2$, then we have both proximal operators. Among all methods introduced so far, we can use only Douglas-Rachford splitting to solve it. But if we consider the Fenchel's dual formulation:

$$\min_{\mathbf{x}}[f(\mathbf{x}) + g(\mathbf{x})] = -\min_{\mathbf{y}}[f^*(\mathbf{y}) + g^*(-\mathbf{y})],$$

then for solving $\min_{\mathbf{y}}[f^*(\mathbf{y}) + g^*(-\mathbf{y})]$, we may use Fast Proximal Gradient method since ∇g^* is Lipschitz continuous due to the strong convexity of $g(\mathbf{x})$. On the other hand, this may not be a good choice since the Lipschtiz constant is exactly α.

7.2 How many different Douglas-Rachford splittings?

Now consider solving

$$\min_{\mathbf{x}}[f(\mathbf{x})+g(\mathbf{x})]=-\min_{\mathbf{y}}[f^*(\mathbf{y})+g^*(-\mathbf{y})],$$

with $f(\mathbf{x})=\|\mathbf{x}\|_1$ and $g(\mathbf{x})=\iota_{\{\mathbf{x}:A\mathbf{x}=b\}}(\mathbf{x})$. To apply the Douglas-Rachford splitting, it seems that there are at least four choices to do fixed point iteration $\mathbf{y}_{k+1}=T(\mathbf{y}_k)$:

1. $T=\frac{1}{2}[\mathbb{I}+\mathrm{R}_{f(\mathbf{x})}\,\mathrm{R}_{g(\mathbf{x})}]$.
2. $T=\frac{1}{2}[\mathbb{I}+\mathrm{R}_{g(\mathbf{x})}\,\mathrm{R}_{f(\mathbf{x})}]$.
3. $T=\frac{1}{2}[\mathbb{I}+\mathrm{R}_{f^*(\mathbf{y})}\,\mathrm{R}_{g^*(-\mathbf{y})}]$.
4. $T=\frac{1}{2}[\mathbb{I}+\mathrm{R}_{g^*(-\mathbf{y})}\,\mathrm{R}_{f^*(\mathbf{y})}]$.

For the first two choices, for the Peaceman-Rachford splitting, it can be easily proven that they are the same if using special intial guess:

Theorem 7.8. *The sequence produced by*

$$\begin{cases}\mathbf{z}_{k+1} &= \mathrm{R}_g^\eta\,\mathrm{R}_f^\eta(\mathbf{z}_k)\\ \mathbf{x}_k &= \mathrm{R}_f^\eta(\mathbf{z}_k)\end{cases},\quad \mathbf{z}_0=\mathrm{R}_g^\eta(\mathbf{y}_0),$$

is the same as the sequence produced by

$$\begin{cases}\mathbf{y}_{k+1} &= \mathrm{R}_f^\eta\,\mathrm{R}_g^\eta(\mathbf{y}_k)\\ \mathbf{x}_k &= \mathrm{R}_g^\eta(\mathbf{y}_k)\end{cases},\quad \forall\mathbf{y}_0.$$

Remark 7.1. *For general Douglas-Rachford splitting, though the same result cannot be shown, in practice the difference in numerical performance between two different versions caused by switching f and g is marginal and minimal.*

Now the question is, does it make a difference if using Douglas-Rachford splitting on the Fenchel's dual problem? It turns out that they is still no difference.

For solving $\min_{\mathbf{x}}[F(\mathbf{x})+G(\mathbf{x})]$, with step size $\eta>0$, it can be written as

$$\text{General Douglas-Rachford}:\mathbf{v}_{k+1}=[(1-\lambda)\mathbb{I}+\lambda\frac{\mathbb{I}+\mathrm{R}_F^\eta\,\mathrm{R}_G^\eta}{2}](\mathbf{v}_k),\quad \lambda\in(0,2).$$

For the primal problem $\min_{\mathbf{x}}[f(\mathbf{x})+g(\mathbf{x})]$, we take $G(\mathbf{x})=f(\mathbf{x})$ and $F(\mathbf{x})=g(\mathbf{x})$, then

$$\text{DR on (P)}:\begin{cases}\mathbf{v}_{k+1} & =[(1-\lambda)\mathbb{I}+\lambda\frac{\mathbb{I}+\mathrm{R}_g^\eta\,\mathrm{R}_f^\eta}{2}](\mathbf{v}_k), \quad \lambda\in(0,2)\\ & =\mathbf{v}_k-\lambda\mathbf{x}_k+\lambda\operatorname{Prox}_f^\eta(2\mathbf{x}_k-\mathbf{v}_k)\\ \mathbf{x}_k & =\operatorname{Prox}_f^\eta(\mathbf{v}_k)\end{cases}. \tag{7.1}$$

For the dual problem $\min_{\mathbf{y}}[f^*(\mathbf{y})+g^*(-\mathbf{y})]$, we take $F(\mathbf{y})=g^*(-\mathbf{y})$ and $G(\mathbf{y})=f^*(\mathbf{y})$, then

$$\operatorname{Prox}_F^\eta(\mathbf{u})=-\operatorname{Prox}_{g^*}^\eta(-\mathbf{u}).$$

Using step size $\tau>0$, we have

$$\text{DR on (D)}:\begin{cases}\mathbf{u}_{k+1} & =[(1-\lambda)\mathbb{I}+\lambda\frac{\mathbb{I}+\mathrm{R}_F^\tau\,\mathrm{R}_G^\tau}{2}](\mathbf{u}_k), \quad \lambda\in(0,2)\\ & =\mathbf{u}_k-\lambda\mathbf{y}_k-\lambda\operatorname{Prox}_{g^*}^\tau(-2\mathbf{y}_k+\mathbf{u}_k)\\ \mathbf{y}_k & =\operatorname{Prox}_{f^*}^\tau(\mathbf{u}_k)\end{cases}. \tag{7.2}$$

Theorem 7.9. *The general Douglas-Rachford splitting on the primal problem* (7.1) *is exactly the same as general Douglas-Rachford splitting on the dual problem* (7.2) *if* $\eta=\frac{1}{\tau}$. *In particular,* $\mathbf{x}_k\to\mathbf{x}_*,\mathbf{y}_k\to\mathbf{y}_*$ *and*

$$\mathbf{u}_k=\frac{\mathbf{v}_k}{\eta},\quad \mathbf{y}_k=\frac{\mathbf{v}_k-\mathbf{x}_k}{\eta}.$$

Problem 7.1. *Prove the two theorems in this section. If using* (7.2), *how to recover the physcial variable* $\mathbf{x}$ *from its iterate* $\mathbf{u}_k$ *or* $\mathbf{y}_k$*?*

7.3 Primal Dual Hybrid Gradient (PDHG) method

For a given composite problem in Theorem 7.4, we have the following three equivalent formulation:

$$\begin{aligned}\text{Primal Problem (P)}: &\quad \min_{\mathbf{x}}[f(\mathbf{x})+g(\mathbf{x})]\\ \text{Dual Problem (D)}: &-\min_{\mathbf{y}}[f^*(\mathbf{y})+g^*(-\mathbf{y})]\\ \text{Primal Dual (PD)}: &\min_{\mathbf{x}}\max_{\mathbf{y}}[\langle\mathbf{x},\mathbf{y}\rangle-f^*(\mathbf{y})+g(\mathbf{x})]\\ \text{Primal Dual relation}: &\mathbf{x}_*\in\partial f^*(\mathbf{y}_*),\quad \mathbf{y}_*\in-\partial g^*(\mathbf{x}_*).\end{aligned}$$

In (PD), the cost function is

$$L(\mathbf{x},\mathbf{y})=\langle\mathbf{x},\mathbf{y}\rangle-f^*(\mathbf{y})+g(\mathbf{x}),$$

and for finding the saddle point $\min_{\mathbf{x}}\max_{\mathbf{y}}L(\mathbf{x},\mathbf{y})$, a simple method is to use implicit gradient descent/ascent:

$$\begin{aligned}\frac{\mathbf{x}_{k+1}-\mathbf{x}_k}{\eta}&=-\frac{\partial L(\mathbf{x}_{k+1},\mathbf{y}_k)}{\partial\mathbf{x}}=-\mathbf{y}_k-\partial g(\mathbf{x}_{k+1})\\ \frac{\mathbf{y}_{k+1}-\mathbf{y}_k}{\eta}&=\frac{\partial L(\mathbf{x}_{k+1},\mathbf{y}_{k+1})}{\partial\mathbf{y}}=\mathbf{x}_{k+1}-\partial f^*(\mathbf{y}_{k+1})\end{aligned}$$

which gives the Arrow-Hurwitz method (1958):

$$\text{Arrow-Hurwitz}: \quad \begin{cases} \mathbf{x}_{k+1} & = \mathrm{Prox}_g^{\eta}[\mathbf{x}_k - \eta \mathbf{y}_k] \\ \mathbf{y}_{k+1} & = \mathrm{Prox}_{f^*}^{\tau}[\mathbf{y}_k + \tau \mathbf{x}_{k+1}] \end{cases}, \quad \eta > 0, \tau > 0.$$

For the Arrow-Hurwitz method to converge, the step sizes must be small enough. A better method is the Primal Dual Hybrid Gradient (PDHG) method introduced around 2010:

$$\text{PDHG}: \quad \begin{cases} \mathbf{x}_{k+1} & = \mathrm{Prox}_g^{\eta}[\mathbf{x}_k - \eta \mathbf{y}_k] \\ \mathbf{y}_{k+1} & = \mathrm{Prox}_{f^*}^{\tau}[\mathbf{y}_k + \tau(2\mathbf{x}_{k+1} - \mathbf{x}_k)] \end{cases}, \quad \eta > 0, \tau > 0, \eta\tau \le 1.$$

Theorem 7.10. *The PDHG method with $\tau = \frac{1}{\eta}$ is equivalent to the Douglas-Rachford splitting $\frac{\mathbb{I} + \mathrm{R}_f \mathrm{R}_g}{2}$. Thus the PDHG method with $\tau = \frac{1}{\eta}$ converges for any $\eta > 0$ if f and g are two convex functions satisying assumptions in the Fenchel's duality Theorem.*

Proof. Define $\mathbf{v}_k = \mathbf{x}_k - \eta \mathbf{y}_k$, then the PDHG method above with $\tau = \frac{1}{\eta}$ becomes

$$\begin{cases} \mathbf{x}_{k+1} & = \mathrm{Prox}_g^{\eta}[\mathbf{v}_k] \\ \mathbf{v}_{k+1} & = \mathbf{v}_k - \mathbf{x}_{k+1} + \mathrm{Prox}_f^{\eta}[(2\mathbf{x}_{k+1} - \mathbf{v}_k)] \end{cases}, \quad \eta > 0.$$

□

7.4 A simple version of ADMM

The Alternating Direction Method of Multipliers (ADMM) introduced in 1970s is a widely use propular method. We first consider its simplest version. For solving $\min_{\mathbf{x}} f(\mathbf{x}) + g(\mathbf{x})$, we rewrite it as

$$\min f(\mathbf{w}) + g(\mathbf{z}), \quad \mathbf{w} = \mathbf{z}.$$

For such a constrained minimization, the *Lagrangian* is defined as

$$L(\mathbf{w}, \mathbf{z}, \mathbf{y}) = f(\mathbf{x}) + g(\mathbf{z}) - \langle \mathbf{y}, \mathbf{w} - \mathbf{z} \rangle,$$

where $\mathbf{y}$ is the Lagrangian multiplier. For finding a saddle point to the Lagrangian, the Augmented Lagrangian is given as

$$\mathcal{L}(\mathbf{w}, \mathbf{z}, \mathbf{y}) = f(\mathbf{x}) + g(\mathbf{z}) - \langle \mathbf{y}, \mathbf{w} - \mathbf{z} \rangle + \frac{\tau}{2}\|\mathbf{w} - \mathbf{z}\|^2.$$

The ADMM method with step sizes $\tau > 0$ and $\sigma > 0$ is given as

$$\begin{aligned} \mathbf{z}_{k+1} &= \operatorname*{argmin}_{\mathbf{z}} \mathcal{L}(\mathbf{w}_k, \mathbf{z}, \mathbf{y}_k) \\ \mathbf{w}_{k+1} &= \operatorname*{argmin}_{\mathbf{w}} \mathcal{L}(\mathbf{w}, \mathbf{z}_{k+1}, \mathbf{y}_k) \\ \mathbf{y}_{k+1} &= \mathbf{y}_k + \sigma \frac{\partial \mathcal{L}}{\partial y}(\mathbf{w}_{k+1}, \mathbf{z}_{k+1}, \mathbf{y}_k) \end{aligned}$$

which is equivalent to

$$\text{(ADMM)}: \quad \begin{aligned} \mathbf{z}_{k+1} &= \operatorname*{argmin}_{\mathbf{z}} g(\mathbf{z}) - \langle \mathbf{y}_k, \mathbf{w}_k - \mathbf{z} \rangle + \frac{\tau}{2}\|\mathbf{w}_k - \mathbf{z}\|^2 \\ \mathbf{w}_{k+1} &= \operatorname*{argmin}_{\mathbf{w}} f(\mathbf{w}) - \langle \mathbf{y}_k, \mathbf{w} - \mathbf{z}_{k+1} \rangle + \frac{\tau}{2}\|\mathbf{w} - \mathbf{z}_{k+1}\|^2 \\ \mathbf{y}_{k+1} &= \mathbf{y}_k + \sigma(\mathbf{w}_{k+1} - \mathbf{z}_{k+1}) \end{aligned}$$

Theorem 7.11. *The ADMM method with $\sigma = \tau = \eta$ is equivalent to the Douglas-Rachford splitting $\frac{\mathbb{I}+\mathrm{R}_F^\eta \mathrm{R}_G^\eta}{2}$ on the dual problem with $F = g^*(-\mathbf{y})$ and $G = f^*(\mathbf{y})$. Thus the ADMM method converges for any two convex functions if using step size $\sigma = \tau > 0$.*

Proof. For the DR splitting $\mathbf{v}_{k+1} = \frac{\mathbb{I}+\mathrm{R}_F^\eta \mathrm{R}_G^\eta}{2}(\mathbf{v}_k), \quad \mathbf{y}_k = \mathrm{Prox}_G(\mathbf{v}_k)$, define $\frac{\mathbf{v}_{k+1}-\mathbf{y}_k}{\eta} = \mathbf{z}_{k+1}$ and $\frac{\mathbf{v}_k-\mathbf{y}_k}{\eta} = \mathbf{w}_k$, then it can be verified. □

Problem 7.2. *Finish the proof above.*

Problem 7.3. *Start with the general DR splitting on the dual problem $\mathbf{v}_{k+1} = [(1-\lambda)\mathbb{I} + \lambda\frac{\mathbb{I}+\mathrm{R}_F^\eta \mathrm{R}_G^\eta}{2}(\mathbf{v}_k)], \quad \mathbf{y}_k = \mathrm{Prox}_G(\mathbf{v}_k)$ to derive a general ADMM method with a relaxation parameter $\lambda \in (0, 2)$.*

7.5 Split Bregman method

The *Bregman distance* for a convex function f is defined as

$$D_f(\mathbf{x}, \mathbf{y}) = f(\mathbf{x}) - f(\mathbf{y}) - \langle \partial f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle.$$

The Bregman distance was used for the original Bregman iteration for minimization. The split Bregman method by Goldstein and Osher in 2009 is also a very popular method. For simplicity, we first consider it for solving

$$\min f(\mathbf{w}) + g(\mathbf{z}), \quad \mathbf{w} = \mathbf{z}.$$

Consider an unconstrained problem

$$\min_{\mathbf{w},\mathbf{z}} f(\mathbf{w}) + g(\mathbf{z}) + \frac{\tau}{2}\|\mathbf{w} - \mathbf{z}\|^2,$$

for which the Bregman iteration is given as

$$\text{(Split Bregman)}: \quad \begin{aligned} \mathbf{z}_{k+1} &= \operatorname*{argmin}_{\mathbf{z}} g(\mathbf{z}) + \frac{\tau}{2}\|\mathbf{w}_k - \mathbf{z} - \mathbf{u}_k\|^2 \\ \mathbf{w}_{k+1} &= \operatorname*{argmin}_{\mathbf{w}} f(\mathbf{w}) + \frac{\tau}{2}\|\mathbf{w} - \mathbf{z}_{k+1} - \mathbf{u}_k\|^2 \\ \mathbf{u}_{k+1} &= \mathbf{u}_k + (\mathbf{w}_{k+1} - \mathbf{z}_{k+1}) \end{aligned}$$

Plug in $\mathbf{u} = \frac{1}{\tau}\mathbf{y}$, then it can be written as

$$\text{(Split Bregman)}: \quad \begin{aligned} \mathbf{z}_{k+1} &= \operatorname*{argmin}_{\mathbf{z}} g(\mathbf{z}) - \langle \mathbf{y}_k, \mathbf{w}_k - \mathbf{z} \rangle + \frac{\tau}{2}\|\mathbf{w}_k - \mathbf{z}\|^2 \\ \mathbf{w}_{k+1} &= \operatorname*{argmin}_{\mathbf{w}} f(\mathbf{w}) - \langle \mathbf{y}_k, \mathbf{w} - \mathbf{z}_{k+1} \rangle + \frac{\tau}{2}\|\mathbf{w} - \mathbf{z}_{k+1}\|^2 \\ \mathbf{y}_{k+1} &= \mathbf{y}_k + \tau(\mathbf{w}_{k+1} - \mathbf{z}_{k+1}) \end{aligned}$$

So this version of split Bregman method is equivalent to ADMM with $\sigma = \tau$, thus also equivalent to the DR spliting on the dual.

7.6 Equivalence of popular algorithms

Now consider popular algorithms for solving

$$\begin{aligned} \text{Primal Problem (P)}: &\quad \min_{\mathbf{x}} [f(\mathbf{x}) + g(\mathbf{x})] \\ \text{Dual Problem (D)}: &- \min_{\mathbf{y}} [f^*(\mathbf{y}) + g^*(-\mathbf{y})] \\ \text{Primal Dual (PD)}: &\min_{\mathbf{x}} \max_{\mathbf{y}} [\langle \mathbf{x}, \mathbf{y} \rangle - f^*(\mathbf{y}) + g(\mathbf{x})], \end{aligned}$$

the following are exactly equivalent:

1. PDHG on (PD) with step size choice $\tau = \frac{1}{\eta}$.
2. Simple Douglas-Rachford splitting on (P) with step size $\eta > 0$.
3. ADMM on (D) with step size $\tau = \sigma = \eta$.
4. Split Bregman on (D) with parameter $\tau = \eta$.
5. Simple Douglas-Rachford splitting on (D) with step size $\frac{1}{\eta} > 0$.
6. ADMM on (P) with step size $\tau = \sigma = \frac{1}{\eta}$.
7. Split Bregman on (P) with parameter $\tau = \frac{1}{\eta}$.

8

Splitting methods for TV norm minimization and similar problems

In this chapter, we consider solving a problem in the form

$$\min_{\mathbf{x}} f(K\mathbf{x}) + g(\mathbf{x}),$$

where K is a matrix or a linear transformation, and we have the proximal operator for $f(\mathbf{y})$ but not for $F(\mathbf{x}) = f(K\mathbf{x})$. One example is the ROF model for TV-norm denoising in Section 4.1.2. We may consider the TV-norm denoising for a one-dimensional signal as an example:

$$\min_{\mathbf{x}\in\mathbb{R}^n} \|D\mathbf{x}\|_1 + \frac{\alpha}{2}\|\mathbf{x} - d\|^2, \tag{8.1}$$

where $d \in \mathbb{R}^n$ is a given 1D noisy signal and D is the finite difference matrix for approximating first order derivatives:

$$D = \begin{pmatrix} -1 & 1 & & & & \\ & -1 & 1 & & & \\ & & -1 & 1 & & \\ & & & \ddots & \ddots & \\ & & & & -1 & 1 \\ & & & & & 0 \end{pmatrix}_{n\times n}.$$

8.1 Lagrangian and the dual problem

For a linear operator K, let K^* denote its adjoint operator. For a real matrix D, its adjoint D^* is simply the transpose matrix D^T. We also consider a more general problem in the following form

$$\min_{\mathbf{x},\mathbf{y}} f(\mathbf{x}) + g(\mathbf{y}), \quad A\mathbf{x} + B\mathbf{y} = C,$$

where A and B are matrices. Notice that $\mathbf{x}$ and $\mathbf{y}$ may not have the same size. For example, for $\min_{\mathbf{x}}[f(K\mathbf{x}) + g(\mathbf{x})]$, we can rewrite it as

$$\min_{\mathbf{v},\mathbf{x}}[f(\mathbf{v}) + g(\mathbf{x})], \quad \mathbf{v} - K\mathbf{x} = 0.$$

For a constrained problem, the Lagrangian is given as

$$L(\mathbf{x}, \mathbf{y}, \mathbf{z}) = f(\mathbf{x}) + g(\mathbf{y}) + \langle \mathbf{z}, A\mathbf{x} + B\mathbf{y} - C\rangle.$$

Under some technical assumptions (see [12, 14]) for the matrices A and B and convex functions f and g for ensuring regularity and total duality, we have

$$\begin{aligned}
&\min_{\mathbf{x},\mathbf{y}} \max_{\mathbf{z}}[f(\mathbf{x}) + g(\mathbf{y}) + \langle \mathbf{z}, A\mathbf{x} + B\mathbf{y} - C\rangle] \\
=&\max_{\mathbf{z}}\{\min_{\mathbf{x}}[f(\mathbf{x}) + \langle \mathbf{z}, A\mathbf{x}\rangle + \min_{\mathbf{y}}[g(\mathbf{y}) + \langle \mathbf{z}, B\mathbf{y}\rangle] + \langle \mathbf{z}, -C\rangle\} \\
=&\max_{\mathbf{z}}\{-\max_{\mathbf{x}}[\langle \mathbf{x}, -A^T\mathbf{z}\rangle - f(\mathbf{x})] - \max_{\mathbf{y}}[\langle \mathbf{y}, -B^T\mathbf{z}\rangle - g(\mathbf{y})] + \langle \mathbf{z}, -C\rangle\} \\
=&\max_{\mathbf{z}}[-f^*(-A^T\mathbf{z}) - g^*(-B^T\mathbf{z}) + \langle \mathbf{z}, -C\rangle] \\
=&-\min_{\mathbf{z}}[f^*(-A^T\mathbf{z}) + g^*(-B^T\mathbf{z}) + \langle \mathbf{z}, C\rangle],
\end{aligned}$$

which implies the following problems are equivalent:

$$\begin{aligned}
\text{Primal Problem (P)}: &\quad \min_{\mathbf{x}}[f(K\mathbf{x}) + g(\mathbf{x})] \\
\text{Dual Problem (D)}: &- \min_{\mathbf{y}}[f^*(\mathbf{y}) + g^*(-K^*\mathbf{y})] \\
\text{Primal Dual (PD)}: &\min_{\mathbf{x}} \max_{\mathbf{y}}[\langle K\mathbf{x}, \mathbf{y}\rangle - f^*(\mathbf{y}) + g(-K^*\mathbf{x})].
\end{aligned}$$

8.2 The dual proximal gradient method

For the example (8.1), we have $f(\mathbf{x}) = \|\mathbf{x}\|_1$ and $g(\mathbf{x}) = \frac{\alpha}{2}\|\mathbf{x} - d\|^2$, and their conjugate functions are computable:

$$f^*(\mathbf{y}) = \iota\{\mathbf{y} : \|\mathbf{y}\|_\infty \le 1\}, \quad g^*(\mathbf{y}) = \frac{1}{2\alpha}\|\mathbf{y} + \alpha d\|^2 - \frac{d^2}{2\alpha}.$$

So we have

$$\begin{aligned}
\text{Primal Problem (P)}: &\min_{\mathbf{x}} \|D\mathbf{x}\|_1 + \frac{\alpha}{2}\|\mathbf{x} - d\|^2, \\
\text{Dual Problem (D)}: &- \min_{\mathbf{y}}[f^*(\mathbf{y}) + \frac{1}{2\alpha}\|D^T\mathbf{y} - \alpha d\|^2 - \frac{d^2}{2\alpha}].
\end{aligned}$$

For the primal problem (P), we can apply the subgradient method which is however as slow as not converging in practice. We cannot use the proximal gradient method for (P) since the proximal operator to $\|D\mathbf{x}\|_1$ is not available. But it is quite straightforward to use the (fast) proximal gradient method on (D), which is also called the *dual proximal gradient method.*

8.3 The PDHG method

For the Lagrangian $L(\mathbf{x},\mathbf{y}) = \langle \mathbf{y}, K\mathbf{x}\rangle - f^*(\mathbf{y}) + g(\mathbf{x})$, to find the saddle point $\min_{\mathbf{x}} \max_{\mathbf{y}} L(\mathbf{x},\mathbf{y})$, the Arrow-Hurwitz method (1958) is to use implicit gradient descent/ascent:

$$\begin{aligned}\frac{\mathbf{x}_{k+1}-\mathbf{x}_k}{\eta} &= -\frac{\partial L(\mathbf{x}_{k+1},\mathbf{y}_k)}{\partial \mathbf{x}} = -K^*\mathbf{y}_k - \partial g(\mathbf{x}_{k+1})\\ \frac{\mathbf{y}_{k+1}-\mathbf{y}_k}{\eta} &= \frac{\partial L(\mathbf{x}_{k+1},\mathbf{y}_{k+1})}{\partial \mathbf{y}} = K\mathbf{x}_{k+1} - \partial f^*(\mathbf{y}_{k+1})\end{aligned}$$

which gives :

$$\text{Arrow-Hurwitz :} \quad \begin{cases}\mathbf{x}_{k+1} &= \text{Prox}_g^{\eta}[\mathbf{x}_k - \eta K^*\mathbf{y}_k]\\ \mathbf{y}_{k+1} &= \text{Prox}_{f^*}^{\tau}[\mathbf{y}_k + \tau K\mathbf{x}_{k+1}]\end{cases}.$$

The convergence of the Arrow-Hurwitz method can be proven if g is strongly convex, and the step sizes must be small enough. A better method is the Primal Dual Hybrid Gradient (PDHG) method:

$$\text{PDHG :} \quad \begin{cases}\mathbf{x}_{k+1} &= \text{Prox}_g^{\eta}[\mathbf{x}_k - \eta K^*\mathbf{y}_k]\\ \mathbf{y}_{k+1} &= \text{Prox}_{f^*}^{\tau}[\mathbf{y}_k + \tau K(2\mathbf{x}_{k+1} - \mathbf{x}_k)]\end{cases}, \quad \eta > 0, \tau > 0, \tau\eta < \frac{1}{\|K\|^2}.$$

Theorem 8.1. *($\mathcal{O}(\frac{1}{k})$ convergence rate.) Let $\tilde{\mathbf{x}}_k = \frac{1}{k}\sum_{i=1}^k \mathbf{x}_i$, the PDHG method with $\tau\eta < \frac{1}{\|K\|^2}$ for convex functions f and g satisfies*

$$L(\tilde{\mathbf{x}}_k, \mathbf{y}) - L(\mathbf{x}, \tilde{\mathbf{y}}_k) \le \frac{1}{k}[\frac{1}{\tau}\|\mathbf{x} - \mathbf{x}_0\|^2 + \frac{1}{\sigma}\|\mathbf{y} - \mathbf{y}_0\|^2].$$

Recall that the PDHG method is equivalent to Douglas-Rachford splitting solving the primal problem, if $K = I$. For a general linear operator K, this equivalence is no longer true. However, the PDHG method for a general linear operator K is still equivalent to Douglas-Rachford splitting solving a different problem [9]:

$$\min_{\mathbf{u},\mathbf{v}} f(K\mathbf{u} + C\mathbf{v}) + g(\mathbf{u}) + \iota_{\mathbf{v}=0},$$

where

$$C = (\gamma^{-2}\mathbb{I} - KK^T)^{\frac{1}{2}}, \quad \gamma\|K\| \le 1.$$

8.4 The accelerated PDHG method

The accelerated PDHG method introduced by Chambolle and Pock around 2010 is a very popular and easy-to-implement method. The PDHG can be equivalently written as

$$\text{PDHG}: \begin{cases} \mathbf{y}_{k+1} & = \text{Prox}^{\tau}_{f^*}[\mathbf{y}_k + \tau K\bar{\mathbf{x}}_k] \\ \mathbf{x}_{k+1} & = \text{Prox}^{\eta}_{g}[\mathbf{x}_k - \eta K^*\mathbf{y}_{k+1}]] \\ \bar{\mathbf{x}}_k & = \mathbf{x}_{k+1} + \theta(\mathbf{x}_{k+1} - \mathbf{x}_k), \quad \theta = 1 \end{cases}.$$

The accelerated PDHG method by Chambolle and Pock is for a μ-strongly convex function g:

$$\text{fast PDHG}: \begin{cases} & \tau_0\eta_0 \le \frac{1}{\|K\|^2}, \quad \bar{\mathbf{x}}_0 = \mathbf{x}_0, \\ \mathbf{y}_{k+1} & = \text{Prox}^{\tau}_{f^*}[\mathbf{y}_k + \tau K\bar{\mathbf{x}}_k] \\ \mathbf{x}_{k+1} & = \text{Prox}^{\eta}_{g}[\mathbf{x}_k - \eta K^*\mathbf{y}_{k+1}]] \\ \theta_k & = \frac{1}{\sqrt{1+2\mu\eta_k}}, \quad \eta_{k+1} = \theta_k\eta_k, \tau_{k+1} = \frac{\tau_k}{\theta_k} \\ \bar{\mathbf{x}}_k & = \mathbf{x}_{k+1} + \theta_k(\mathbf{x}_{k+1} - \mathbf{x}_k) \end{cases}.$$

Theorem 8.2. *($\mathcal{O}(\frac{1}{k^2})$ convergence rate.) Let $\tilde{\mathbf{x}}_k = \frac{1}{k}\sum_{i=1}^{k}\mathbf{x}_i$, the fast PDHG method with $\tau\eta < \frac{1}{\|K\|^2}$ for a convex function f and a strongly convex function g satisfies*

$$0 \le \sup_{\mathbf{y}\in B_2} L(\tilde{\mathbf{x}}_k, \mathbf{y}) - \inf_{\mathbf{x}\in B_1} L(\mathbf{x}, \tilde{\mathbf{y}}_k) \le \sup_{\mathbf{y}\in B_2}\sup_{\mathbf{y}\in B_2} \frac{1}{k^2}[\frac{1}{\tau}\|\mathbf{x}-\mathbf{x}_0\|^2 + \frac{1}{\sigma}\|\mathbf{y}-\mathbf{y}_0\|^2],$$

where (B_1, B_2) is a closed bounded set containing a saddle point. If there is only one saddle point, then $(\mathbf{x}_k, \mathbf{y}_k) \to (\mathbf{x}_*, \mathbf{y}_*)$.

8.5 ADMM

For a more general problem

$$\min_{\mathbf{x},\mathbf{y}} f(\mathbf{x}) + g(\mathbf{y}), \quad A\mathbf{x} + B\mathbf{y} = C,$$

recall that the Lagrangian is given as

$$L(\mathbf{x}, \mathbf{y}, \mathbf{z}) = f(\mathbf{x}) + g(\mathbf{y}) + \langle \mathbf{z}, A\mathbf{x} + B\mathbf{y} - C\rangle.$$

The augmented Lagrangian with a parameter $\sigma > 0$ is given as

$$\mathcal{L}_\sigma(\mathbf{x}, \mathbf{y}, \mathbf{z}) = f(\mathbf{x}) + g(\mathbf{y}) + \langle \mathbf{z}, A\mathbf{x} + B\mathbf{y} - C\rangle + \frac{\sigma}{2}\|A\mathbf{x} + B\mathbf{y} - C\|^2.$$

The ADMM method with step sizes $\tau > 0$ and $\sigma > 0$ is given as

$$\begin{aligned} \mathbf{x}_{k+1} &= \operatorname*{argmin}_{\mathbf{x}} \mathcal{L}_\sigma(\mathbf{x}, \mathbf{y}_k, \mathbf{z}_k) \\ \mathbf{y}_{k+1} &= \operatorname*{argmin}_{\mathbf{y}} \mathcal{L}_\sigma(\mathbf{x}_{k+1}, \mathbf{y}, \mathbf{z}_k) \\ \mathbf{z}_{k+1} &= \mathbf{z}_k + \tau\frac{\partial \mathcal{L}_\sigma}{\partial z}(\mathbf{x}_{k+1}, \mathbf{y}_{k+1}, \mathbf{z}_k) \end{aligned}$$

which is equivalent to

$$
\text{(ADMM)}: \quad
\begin{aligned}
\mathbf{x}_{k+1} &= \operatorname*{argmin}_{\mathbf{x}} f(\mathbf{x}) + \langle \mathbf{z}_k, A\mathbf{x} + B\mathbf{y}_k - C\rangle + \frac{\sigma}{2}\|A\mathbf{x} + B\mathbf{y}_k - C\|^2 \\
\mathbf{y}_{k+1} &= \operatorname*{argmin}_{\mathbf{y}} g(\mathbf{y}) + \langle \mathbf{z}_k, A\mathbf{x}_{k+1} + B\mathbf{y} - C\rangle + \frac{\sigma}{2}\|A\mathbf{x}_{k+1} + B\mathbf{y} - C\|^2 \\
\mathbf{z}_{k+1} &= \mathbf{z}_k + \tau(A\mathbf{x}_{k+1} + B\mathbf{y}_{k+1} - C).
\end{aligned}
$$

From the Lagrangian, we can derive the dual problem:

$$
\begin{aligned}
\text{(P)}: &\quad \min_{\mathbf{x},\mathbf{y}} f(\mathbf{x}) + g(\mathbf{y}), \quad A\mathbf{x} + B\mathbf{y} = C \\
\text{(D)}: &\quad -\min_{\mathbf{z}}[f^*(-A^T\mathbf{z}) + g^*(-B^T\mathbf{z}) + \langle \mathbf{z}, C\rangle],
\end{aligned}
$$

Theorem 8.3. *Assume some technical conditions for matrices A, B and convex functions f, g so that $(P) \Leftrightarrow (D)$ and $F(\mathbf{z}) = f^*(-A^T\mathbf{z})$ and $G(\mathbf{z}) = g^*(-B^T\mathbf{z}) + \langle \mathbf{z}, C\rangle$ are well defined, then the ADMM method with $\sigma = \tau = \eta$ is equivalent to the Douglas-Rachford splitting $\frac{\mathbb{I}+\mathrm{R}_F^\eta \mathrm{R}_G^\eta}{2}$ on the dual problem. Thus the ADMM method converges for any two convex functions if using step size $\sigma = \tau > 0$.*

Problem 8.1. *Start from the general Douglas-Rachford splitting $(1-\lambda)\mathbb{I} + \lambda\frac{\mathbb{I}+\mathrm{R}_F^\eta \mathrm{R}_G^\eta}{2}$ to derive a general ADMM method with an additional relaxation parameter $\lambda \in (0, 2)$. What assumptions do we need so that the limiting Peaceman-Rachford splitting for $\lambda = 2$ will converge?*

8.6 Implementation of TV norm minimization

The TV norm minimization for image denoising has been proposed since early 1990s [13, 15].

8.6.1 Continuum ROF image denoising model

The discussion of a continuum setup only serves as an intuitional guide for us to derive the discrete analog later. Consider a rectangular domain $\Omega = [0, 1] \times [0, 1]$, and a function $u(x, y) \in H^1(\Omega)$, which represents an image with infinite resolution. Then its total variation is defined as

$$
\|u\|_{TV} = \iint_\Omega |\nabla u| dxdy,
$$

where $\nabla u = (u_x, u_y)$ and $|\nabla u| = \sqrt{|u_x|^2 + |u_y|^2}$. With L^2-norm as

$$
\|u\|_{L^2} = \sqrt{\iint_\Omega |u|^2 dxdy},
$$

for a given $a(x, y)$, the ROF (Rudin, Osher, and Fatemi, 1992) model [13] is to minimize (over u in a proper function space)

$$\|u\|_{TV} + \frac{1}{2}\lambda\|u - a\|_{L^2}^2,$$

where λ is a fixed parameter.

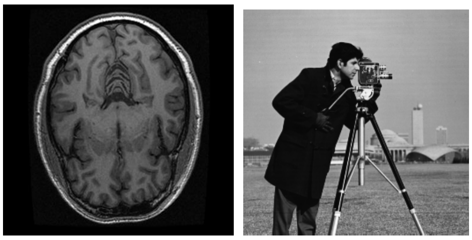


Figure 8.1: Periodic or zero bounary conditions are suitable for MRI images, but not for a generic image.

The function space that the minimizer should belong to, is a subspace of $H^1(\Omega)$ with suitable boundary conditions. For instance, periodic or homogeneous Dirichlet boundary conditions make sense for MRI images, but not for a generic image. For convenience, for a generic image, we just consider homogeneous Neumann boundary conditions, which will naturally emerge in the discrete setup as will be seen in the following subsections. See Figure 8.1.

To this end, we define

$$\mathcal{H} = \{u \in H^1(\Omega) : \nabla u \cdot \mathbf{n}|_{\partial\Omega} = 0\},$$

where $\mathbf{n}$ is the unit normal vector of the boundary $\partial\Omega$.

The gradient operator ∇ is a linear mappping, and we use an abstract name for it $\mathcal{K} = \nabla$:

$$\begin{aligned} \mathcal{K} = \nabla : \mathcal{H} &\longrightarrow \mathcal{V} = (L^2(\Omega), L^2(\Omega)) \\ u &\longmapsto \nabla u = (u_x, u_y) \end{aligned}$$

To understand the adjoint operator of $\mathcal{K} = \nabla$, we need the $H(div)$-space:

$$H(div) = \{\mathbf{q} = (q^1, q^2) \in (L^2(\Omega), L^2(\Omega)) : \nabla \cdot (q^1, q^2) \in L^2(\Omega)\} \subset \mathcal{V}.$$

Remark 8.1. *Elements in $H(div)$ are not necessarily in $H^1(\Omega)$. For instance, let $f(x) = \begin{cases} 1, & x \geq 0 \\ 0, & x < 0 \end{cases}$, then $\mathbf{q}(x, y) := (0, f(x))$ is in $H(div)$ but $f(x) \notin H^1(\Omega)$.*

The divergence operator $\nabla\cdot$ is a linear mapping from $H(div)$ to $L^2(\Omega)$. If we assume suitable boundary conditions for smooth $\mathbf{q}$ so that boundary terms in the integration-by-parts vanish, then $\mathcal{K}^* = -\nabla\cdot$ is the adjoint operator of $\mathcal{K} = \nabla$ since

$$\langle \mathcal{K}u, \mathbf{q}\rangle := \iint_\Omega \nabla u\cdot\mathbf{q}dxdy = -\iint_\Omega u\nabla\cdot\mathbf{q}dxdy = \langle u, -\nabla\cdot\mathbf{q}\rangle, \forall \mathbf{q} \in (C_0^1(\Omega), C_0^1(\Omega)).$$

8.6.2 Discrete ROF model

Consider an image of size $n \times n$, corresponding to domain $[0,1] \times [0,1]$ and a uniform grid $x_i, y_j = (j-1)h$, $j = 1, \cdots, n$ with $h = \frac{1}{n-1}$. Notice that an image does not have any necessary association of a domain of size $[0,1] \times [0,1]$, and this assumption of domain $[0,1] \times [0,1]$ should not affect the final implementation.

Recall that D is the finite difference matrix approximating first order derivative, then we have

$$D^T = \begin{pmatrix} -1 & & & & & \\ 1 & -1 & & & & \\ & 1 & -1 & & & \\ & & \ddots & \ddots & & \\ & & & 1 & -1 & \\ & & & & 1 & 0 \end{pmatrix}_{n\times n}, D^T D = \begin{pmatrix} 1 & -1 & & & & \\ -1 & 2 & -1 & & & \\ & -1 & 2 & -1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & -1 & 2 & -1 \\ & & & & -1 & 1 \end{pmatrix}.$$

Now for a bivariate function $u(x,y)$, let U be a 2D array with $U(j,i) = u(x_i, y_j)$, then

$$U \approx u \Rightarrow \frac{1}{h}UD^T \approx u_x, \quad \frac{1}{h}DU \approx u_y.$$

Remark 8.2. *We may also choose the notation $U(i,j) = u(x_i, y_j)$, then $\frac{1}{h}DU \approx u_x$.*

For function $u(x,y)$ and $a(x,y)$, we have

$$||u||_{TV} \approx \sum_i \sum_j h^2 \sqrt{|u_x(x_i,y_j)|^2 + |u_y(x_i,y_j)|^2}$$

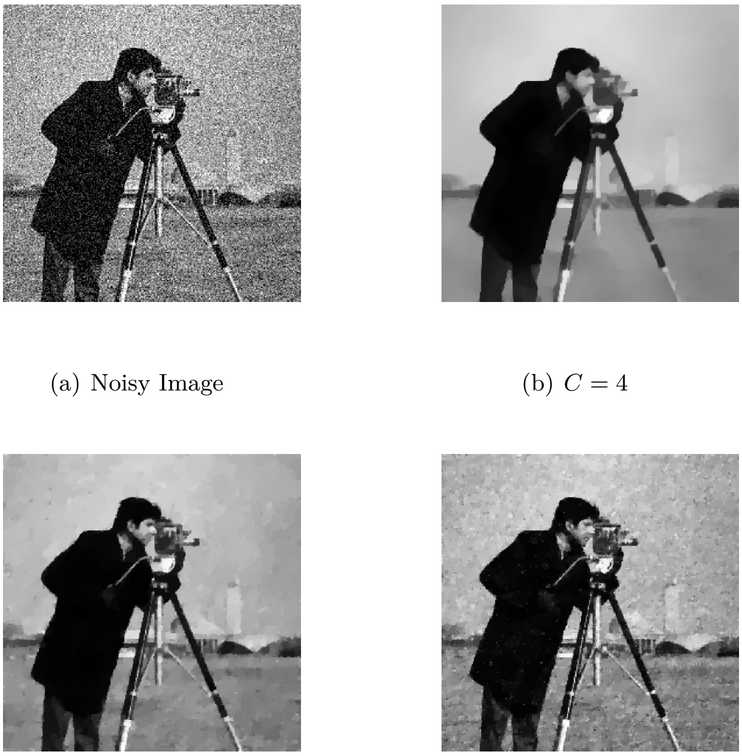
$$||u-a||_{L^2}^2 \approx \sum_i \sum_j h^2 |u(x_i,y_j) - a(x_i,y_j)|^2.$$

Introduce $U_x := \frac{1}{h}UD^T$ and $U_y := \frac{1}{h}DU$. Let A be a 2D array with $A(j,i) = a(x_i, y_j)$, then the discrete ROF model is

$$\min_{U\in\mathbb{R}^{n\times n}} \sum_i \sum_j \left(h^2 \sqrt{U_x(j,i)^2 + U_y(j,i)^2} + h^2 \frac{1}{2}\lambda |U(j,i) - A(j,i)|^2 \right)$$

$$= \min_{U\in\mathbb{R}^{n\times n}} \sum_i \sum_j \left(h\sqrt{[UD^T](j,i)^2 + [DU](j,i)^2} + h^2 \frac{1}{2}\lambda |U(j,i) - A(j,i)|^2 \right).$$

Notice that its minimizer does not depend on the choice of h if $\lambda = \frac{C}{h}$ for some constant C. An image has no physical grid spacing anyway, so h should be arbitrary, for which we should take $\lambda = \frac{C}{h}$. In practice, $C = 10$ usually produces a reasonable result. See Figure 8.2 and Figure 8.3.



(a) Noisy Image

(b) $C = 4$

(c) $C = 8$

(d) $C = 12$

Figure 8.2: ROF solutions using isotropic TV-norm with different $\lambda = \frac{C}{h}$.

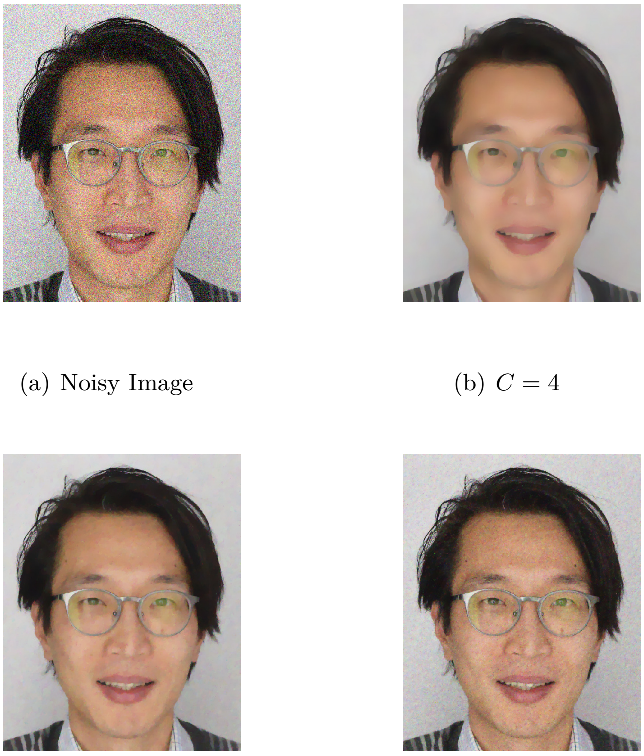
8.6.3 Primal, dual and primal-dual forms

Using all notation above, the discrete ROF model can be written as

$$\min_{U\in\mathbb{R}^{n\times n}} f(\mathcal{K}U) + g(U), \tag{8.2a}$$

where $\mathcal{K} = \nabla_h : \mathbb{R}^{n\times n} \longrightarrow \mathbb{R}^{2(n\times n)}$ is a linear mapping

$$\mathcal{K}(U) = \nabla_h U = \frac{1}{h}(UD^T, DU), \tag{8.2b}$$



(a) Noisy Image

(b) $C = 4$

(c) $C = 8$

(d) $C = 12$

Figure 8.3: ROF solutions using isotropic TV-norm with different $\lambda = \frac{C}{h}$.

and

$$f(P,Q) = \sum_{i,j} h^2 \sqrt{P^2(i,j) + Q^2(i,j)}, \quad g(U) = \lambda \sum_{i,j} h^2 |U(i,j) - a(i,j)|^2. \tag{8.2c}$$

It is straightforward to verify that the adjoint operator of $\mathcal{K}$ is given by

$$\begin{aligned} \mathcal{K}^* = -\nabla_h \cdot : \mathbb{R}^{2(n\times n)} &\longrightarrow \mathbb{R}^{n\times n} \\ (P,Q) &\longmapsto \frac{1}{h}(PD + D^T Q) \end{aligned}$$

The convex minimization (8.2a) is called *primal* form. To solve (8.2a), equivalently we can solve its *dual* form

$$-\min_{\mathbf{P}\in\mathbb{R}^{2(n\times n)}} f^*(\mathbf{P}) + g^*(-\mathcal{K}^*\mathbf{P}). \tag{8.3}$$

Both (8.2a) and (8.3) are also equivalent to the *primal-dual* form:

$$\min_{U\in\mathbb{R}^{n\times n}} \max_{\mathbf{P}\in\mathbb{R}^{2(n\times n)}} \langle \mathcal{K}U, \mathbf{P}\rangle - f^*(\mathbf{P}) + g(U), \tag{8.4}$$

Recall that the minimizer U^* to (8.2a) and the minimizer $\mathbf{P}^*$ to (8.3) are related via the optimality condition in the primal-dual form in the previous chapter. To recover the physical image U from $\mathbf{P}$, we need the relation obtained from the Legendre transform of $g(U)$:

$$0 \in \mathcal{K}^*P + \partial g(U),$$

which gives

$$0 = \mathcal{K}^*P + \lambda(U - A) \Rightarrow U = A - \frac{1}{\lambda}\mathcal{K}^*P.$$

8.6.4 ADMM on the primal problem

Both alternating direction method of multipliers (ADMM) (Glowenski and Marrocco 75) and Douglas-Rachford splitting (Lions and Mercier 79) are popular and successful splitting convex minimization algorithms, and we have seen that they are equivalent in the following sense:

ADMM on primal ⇔ Douglas-Rachford on dual,

ADMM on dual ⇔ Douglas-Rachford on primal.

By plugging in the linear constraint $A\mathbf{x}+B\mathbf{y} = C$ as $P-\mathcal{K}U = 0$, ADMM with $\tau = \sigma$ applied on $f(\mathbf{P}) + g(U)$ in the primal form (8.2a) becomes

$$\text{(ADMM)}: \quad \begin{aligned} \mathbf{P}_{k+1} &= \text{argmin}_{\mathbf{P}} f(\mathbf{P}) + \langle \mathbf{Q}_k, \mathbf{P} - \mathcal{K}U_k\rangle + \frac{\sigma}{2}\|\mathbf{P} - \mathcal{K}U_k\|^2 \\ U_{k+1} &= \text{argmin}_U g(U) + \langle \mathbf{Q}_k, \mathbf{P}_{k+1} - \mathcal{K}U\rangle + \frac{\sigma}{2}\|\mathbf{P}_{k+1} - \mathcal{K}U\|^2 \\ \mathbf{Q}_{k+1} &= \mathbf{Q}_k + \sigma(\mathbf{P}_{k+1} - \mathcal{K}U_{k+1}). \end{aligned}$$

For $f(\mathbf{P})$ where $\mathbf{P} = (P, Q)$, its proximal operator for $(\mathbf{V}) = (U, V)$ is defined as

$$\begin{aligned} &\text{Prox}_f^\eta(\mathbf{V}) = \underset{\mathbf{P}}{\text{argmin}}\, f(P,Q) + \frac{1}{2\eta}\|\mathbf{P} - \mathbf{V}\|^2 \\ &= \underset{\mathbf{P}}{\text{argmin}} \sum_{i,j} h^2\sqrt{P(i,j)^2 + Q^2(i,j)} + \frac{1}{2\eta}h^2(|P(i,j) - U(i,j)|^2 + |Q(i,j) - V(i,j)|^2) \\ &= \underset{\mathbf{P}}{\text{argmin}} \sum_{i,j} h^2[\sqrt{P(i,j)^2 + Q^2(i,j)} + \frac{1}{2\eta}(|P(i,j) - U(i,j)|^2 + |Q(i,j) - V(i,j)|^2)]. \end{aligned}$$

Example 8.1. *Consider $f(x,y)=\sqrt{x^2+y^2}$, its subdifferential can be computed as*

$$\partial f(x,y)=\begin{cases}\nabla f=(\frac{x}{\sqrt{x^2+y^2}},\frac{y}{\sqrt{x^2+y^2}}) & , \quad |x|+|y|>0\\ ([-\frac{\sqrt{2}}{2},\frac{\sqrt{2}}{2}],[-\frac{\sqrt{2}}{2},\frac{\sqrt{2}}{2}]) & , \quad |x|+|y|=0\end{cases}.$$

With the subdifferential, we can compute the conjugate function as an indicator of the unit ball:

$$f^*(x,y)=\iota_{\{(x,y):x^2+y^2\leq 1\}}(x,y).$$

Problem 8.2. *Derive the formula for* $\mathrm{Prox}_f^{\eta}(\mathbf{V})$ *using Moreau's formula* $\mathrm{Prox}_f^{\eta}(x)+\eta\,\mathrm{Prox}_{f^*}^{1/\eta}(x\eta)=x$.

To implement the second line in ADMM, by ignoring constants, we consider

$$U_{k+1}=\mathrm{argmin}_U g(U)-\langle\mathbf{Q}_k,\mathcal{K}U\rangle+\frac{\tau}{2}\|\mathcal{K}U-\mathbf{P}_{k+1}\|^2.$$

Notice that $g(U)$ is a simple quadratic function, thus the minimizer is obtained by finding critical point, for which we need to take derivative of $\|\mathcal{K}U-\mathbf{P}_{k+1}\|^2$ w.r.t. U:

$$\frac{\partial}{\partial U}\langle\mathcal{K}U-\mathbf{P}_{k+1},\mathcal{K}U-\mathbf{P}_{k+1}\rangle=h^2(2\mathcal{K}^*\mathcal{K}U-2\mathcal{K}^*\mathbf{P}_{k+1}).$$

So the second line can be equivalently written as

$$\lambda(U_{k+1}-A)-\mathcal{K}^*\mathbf{Q}_k+\tau\mathcal{K}^*\mathcal{K}U_{k+1}-\tau\mathcal{K}^*\mathbf{P}_{k+1}=0$$

which is

$$(\lambda I+\tau\mathcal{K}^*\mathcal{K})U_{k+1}=-\lambda A+\mathcal{K}^*\mathbf{Q}_k+\tau\mathcal{K}^*\mathbf{P}_{k+1}.$$

Notice that $\mathcal{K}^*\mathcal{K}=-\Delta_h$ is precisely the discrete Laplacian with purely Neumann boundary conditions, and $\lambda I-\tau\Delta_h$ can be inverted efficiently in a simple way, see Appendix B.3.

8.6.5 Douglas-Rachford splitting on the dual problem

Using notation in this section, for the TV-norm denoising problem of a 2D image $B\in\mathbb{R}^{n\times n}$, the primal problem is equivalently written as

$$\min_{U\in\mathbb{R}^{n\times n}}\|\mathcal{K}U\|_1+\frac{\lambda}{h}\|U-B\|_F^2,$$

where $\|\cdot\|_F$ is the Frobenius norm for a matrix $\|U-B\|_F=\sqrt{\sum_{i,j}|U(i,j)-a(i,j)|^2}$ and the 1-norm for a pair of matrices $\mathbf{V}=(P,Q)$ is

$$F(\mathbf{V})=\|(P,Q)\|_1=\sum_{i,j}\sqrt{P(i,j)^2+Q(i,j)^2}.$$

The convex conjugate of $F(\mathbf{V})$ is

$$F^*(\mathbf{V}) = \sum_{i,j} \iota_{\{P(i,j)^2+Q(i,j)^2\leq 1\}}.$$

Up to a constant shift, the dual problem can be written as

$$-\min F^*(\mathbf{V}) + \frac{h}{2\lambda}\|\mathcal{K}^*\mathbf{V} - \frac{\lambda}{h}B\|_F^2.$$

Problem 8.3. *Derive the dual problem.*

The proximal operator of F^* can be easily computed as the projection to the unit ball for each entry (i,j).

Now consider the proximal operator of the function

$$G^*(\mathbf{V}) = \frac{h}{2\lambda}\|\mathcal{K}^*\mathbf{V} - \frac{\lambda}{h}B\|_F^2,$$

which is written as

$$\text{Prox}_{G^*}^{\eta}(\mathbf{W}) = \operatorname*{argmin}_{\mathbf{V}} \frac{h}{2\lambda}\|\mathcal{K}^*\mathbf{V} - \frac{\lambda}{h}B\|_F^2 + \frac{1}{2\eta}\|\mathbf{V} - \mathbf{W}\|_F^2.$$

Let $\mathbf{V} = \text{argmin}$, then the critical point equation gives

$$\frac{h}{\lambda}\mathcal{K}(\mathcal{K}^*\mathbf{V} - \frac{\lambda}{h}B) + \frac{1}{\eta}(\mathbf{V} - \mathbf{W}) = 0$$

$$\Rightarrow (\frac{1}{\eta}\mathbb{I} + \frac{h}{\lambda}\mathcal{K}\mathcal{K}^*)\mathbf{V} = \mathcal{K}B + \frac{1}{\eta}\mathbf{W}.$$

We need to solve $\mathbf{V}$ in an equation in the form

$$\mathcal{K}\mathcal{K}^*\mathbf{V} + \beta\mathbf{V} = \mathbf{F}$$

where $\beta = \frac{\lambda}{\eta h}$ and $\mathbf{F} = \eta\mathcal{K}B + \mathbf{W}$ is some known vector field. At first glance, this corresponds to an equation

$$\nabla(-\nabla\cdot\vec{p}) + \beta\vec{p} = \vec{f},$$

which is a harder equation to solve due to the mixed second order derivatives, compared to the Poisson equation.

However, to solve this seemingly difficult equation, we can just compute $(-\Delta_h + \beta\mathbb{I})^{-1}$, which can be computed similarly for computing $(-\Delta_h)^{-1}$ (see Appendix B.3), mainly due to a simple linear algebra fact:

Lemma 8.1. *For a linear operator $\mathcal{K} : \mathbb{R}^{n\times n} \longrightarrow \mathbb{R}^{2(n\times n)}$, assume $\mathcal{K}^*\mathcal{K}$ has an inverse or a right pseudo-inverse $(\mathcal{K}^*\mathcal{K})^{-1}$, then the solution to the equation $\mathcal{K}\mathcal{K}^*\mathbf{V} + \beta\mathbf{V} = \mathbf{F}$ can be written as*

$$\mathbf{V} = \frac{1}{\beta}[\mathbf{F} - \mathcal{K}(\beta\mathbb{I} + \mathcal{K}^*\mathcal{K})^{-1}\mathcal{K}^*\mathbf{F}].$$

Proof. The kernal of $\mathcal{K}^*$ is orthogonal to the range of $\mathcal{K}$ (column space of a matrix K is orthogonal to the left null space of K), thus

$$\mathbb{R}^{2(n\times n)} = \text{Kernel}(\mathcal{K}^*) \oplus \text{Range}(\mathcal{K}),$$

which implies a very useful fact (corresponding to Helmholtz decomposition for suitable vector fields):

$$\mathbf{V} = \mathcal{K}W + \mathbf{G} = \nabla_h W + \mathbf{G}, \quad \text{where } \mathbf{G} \in \text{Kernel}(\mathcal{K}^*),\ \text{i.e., } \nabla_h \cdot \mathbf{G} = 0.$$

Apply $\mathcal{K}^*$ to both sides of the equation, we can first solve for W as follows

$$\mathcal{K}^*\mathcal{K}\mathcal{K}^*\mathbf{V} + \beta\mathcal{K}^T\mathbf{V} = K^*\mathbf{F} \Rightarrow \mathcal{K}^*\mathbf{V} = (\beta\mathbb{I} + \mathcal{K}^*\mathcal{K})^{-1}\mathcal{K}^*\mathbf{F},$$

$$\mathbf{V} = \mathcal{K}W+\mathbf{G} \Rightarrow \mathcal{K}^*\mathbf{V} = \mathcal{K}^*(\mathcal{K}W+\mathbf{G}) \Rightarrow W = (\mathcal{K}^*\mathcal{K})^{-1}\mathcal{K}^*\mathbf{V} = (\mathcal{K}^*\mathcal{K})^{-1}(\beta\mathbb{I}+\mathcal{K}^*\mathcal{K})^{-1}\mathcal{K}^*\mathbf{F}.$$

Then we can solve $\mathbf{G}$ by

$$\mathcal{K}\mathcal{K}^*(\mathcal{K}W + \mathbf{G}) + \beta(\mathcal{K}W + \mathbf{G}) = \mathbf{F} \Rightarrow \mathbf{G} = \frac{1}{\beta}[\mathbf{F} - \mathcal{K}\mathcal{K}^*\mathcal{K}W] - \mathcal{K}W.$$

Finally we get

$$\mathbf{V} = \mathcal{K}W + \mathbf{G} = \frac{1}{\beta}[\mathbf{F} - \mathcal{K}\mathcal{K}^*\mathcal{K}W] = \frac{1}{\beta}[\mathbf{F} - \mathcal{K}(\beta\mathbb{I} + \mathcal{K}^*\mathcal{K})^{-1}\mathcal{K}^*\mathbf{F}].$$

□

Remark 8.3. *It is not a surprise that the seemingly more difficult equation* $\nabla(-\nabla\cdot\vec{p})+\beta\vec{p}=\vec{f}$ *can actually be solved by computing* $(-\Delta_h+\beta\mathbb{I})^{-1}$*, since the Douglas-Rachford splitting on the dual problem is equivalent to ADMM on the primal problem, which involves solving* $(-\Delta_h + \beta\mathbb{I})^{-1}$*.*

8.7 Comparisons and concluding remarks

So for the TV norm minimization problem in this chapter, we have considered three algorithms:

1. The (fast) proximal gradient method on the dual problem (D).
2. The (fast) PDHG method by Chambolle and Pock.
3. The ADMM method on (P) (Douglas-Rachford splitting on (D)).

In terms of implementation, the first two methods do not involve inverting big matrices, e.g., solving a Poisson equation, which is however needed in ADMM. On the other hand, ADMM may converge faster than the first two methods in terms of iteration numbers, though each iteration of ADMM is more expensive. Moreover, in practice, one may use an inaccurate approximation to the solution of the Poisson equation in ADMM and its performance is often quite satisfying. For instance, for solving the Poisson equation, a few iterations of conjugate gradient method can be used as a (very inaccurate) approximation. See [19] for examples and justifications.

Part III

Randomized algorithms

Part IV

Riemannian optimization

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 $Ax = \sum\limits_{j=1}^{n} a_j A v_j = \sum\limits_{j=1}^{n} a_j \lambda_j v_j = V\Lambda a$. 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 $x \in \mathbb{C}^n$,*

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

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 $A = U_1\Sigma_1 V_1$ (***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 Pseudoinverse

Let the compact SVD of $A \in \mathbb{C}^{m\times n}$ be

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

with

$$\Sigma_1 = \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_k \end{bmatrix}, \quad \sigma_i > 0.$$

The pseudoinverse $A^\dagger \in \mathbb{C}^{n\times m}$ is defined as $A^\dagger = V_1\Sigma_1^{-1}U_1^*$. Special cases:

1. A has linearly independent columns, then $A^\dagger = (A^*A)^{-1}A^*$ and $A^\dagger A = \mathbb{I}_{n\times n}$. In this case, $A^\dagger$ is also called left inverse of A.
2. A has linearly independent rows, then $A^\dagger = A^*(AA^*)^{-1}$ and $AA^\dagger = \mathbb{I}_{m\times m}$. In this case, $A^\dagger$ is also called right inverse of A.

A.5 Vector norms

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

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

A.6 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.7 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

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

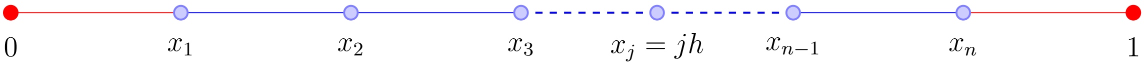


Figure B.1: An illustration of the discretized domain.

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:

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

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

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

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

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

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

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

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

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

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

B.3 Efficient inversion of discrete Laplacian

See Section 2.8 in MA/CS 615 notes.

Appendix C

Basic Theorems in Analysis

The following results are standard in many real analysis books, e.g. [7].

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] Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on information theory*, 52(2):489–509, 2006.

[2] Heinz H Bauschke and Patrick L Combettes. *Convex analysis and monotone operator theory in Hilbert spaces.* Springer, 2013.

[3] Amir Beck. *First-order methods in optimization.* SIAM, 2017.

[4] Scott Shaobing Chen, David L. Donoho, Michael, and A. Saunders. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20:33–61, 1998.

[5] Damek Davis and Wotao Yin. A three-operator splitting scheme and its optimization applications. *Set-valued and variational analysis*, 25:829–858, 2017.

[6] 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.

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

[8] 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.

[9] Daniel O'Connor and Lieven Vandenberghe. On the equivalence of the primal-dual hybrid gradient method and Douglas–Rachford splitting. *Mathematical Programming*, 179(1):85–108, 2020.

[10] C. Planiden and X. Wang. Strongly convex functions, moreau envelopes, and the generic nature of convex functions with strong minimizers. *SIAM Journal on Optimization*, 26(2):1341–1364, 2016.

[11] Ralph Rockafellar. Characterization of the subdifferentials of convex functions. *Pacific Journal of Mathematics*, 17(3):497–510, 1966.

[12] Ralph Tyrell Rockafellar. Convex analysis. 2015.

[13] Leonid I Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: nonlinear phenomena*, 60(1-4):259–268, 1992.

[14] Ernest K Ryu and Wotao Yin. *Large-scale convex optimization: algorithms & analyses via monotone operators.* Cambridge University Press, 2022.

[15] Ken Sauer and Charles Bouman. Bayesian estimation of transmission tomograms using segmentation based optimization. *IEEE Transactions on Nuclear Science*, 39(4):1144–1152, 1992.

[16] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288, 1996.

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

[18] Wotao Yin. Analysis and Generalizations of the Linearized Bregman Method. *SIAM J. Img. Sci.*, 3(4):856–877, October 2010.

[19] Wotao Yin and Stanley Osher. Error forgetting of Bregman iteration. *Journal of Scientific Computing*, 54:684–695, 2013.